JMLR: Workshop and Conference Proceedings vol (2012) 1–23

The best of both worlds: stochastic and adversarial bandits

Sébastien Bubeck SBUBECK@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ, USA

Aleksandrs Slivkins SLIVKINS@MICROSOFT.COM
Microsoft Research
Mountain View, USA

Abstract

We present a new bandit algorithm, SAO (Stochastic and Adversarial Optimal) whose regret is (essentially) optimal both for adversarial rewards and for stochastic rewards. Specifically, SAO combines the $O(\sqrt{n})$ worst-case regret of Exp3 (Auer et al., 2002b) and the (poly)logarithmic regret of UCB1 (Auer et al., 2002a) for stochastic rewards. Adversarial rewards and stochastic rewards are the two main settings in the literature on multi-armed bandits (MAB). Prior work on MAB treats them separately, and does not attempt to jointly optimize for both. This result falls into the general agenda to design algorithms that combine the optimal worst-case performance with improved guarantees for "nice" problem instances.

1. Introduction

Multi-armed bandits (henceforth, MAB) is a simple model for sequential decision making under uncertainty that captures the crucial tradeoff between *exploration* (acquiring new information) and *exploitation* (optimizing based on the information that is currently available). Introduced in early 1950-ies (Robbins, 1952), it has been studied intensively since then in Operations Research, Electrical Engineering, Economics, and Computer Science.

The "basic" MAB framework can be formulated as a game between the player (i.e., the algorithm) and the adversary (i.e., the environment). The player selects actions ("arms") sequentially from a fixed, finite set of possible options, and receives rewards that correspond to the selected actions. For simplicity, it is customary to assume that the rewards are bounded in $[0, 1]$. In the adversarial model one makes no other restrictions on the sequence of rewards, while in the stochastic model we assume that the rewards of a given arm is an i.i.d sequence of random variables. The performance criterion is the so-called regret, which compares the rewards received by the player to the rewards accumulated by a hypothetical benchmark algorithm. A typical, standard benchmark is the best single arm. See Figure 1 for a precise description of this framework.

Adversarial rewards and stochastic rewards are the two main reward models in the MAB literature. Both are now very well understood, in particular thanks to the seminal papers (Lai and Robbins, 1985; Auer et al., 2002a,b). In particular, the `Exp3` algorithm from (Auer et al., 2002b) attains a regret growing as $O(\sqrt{n})$ in the adversarial model, where n is the number of rounds, and `UCB1` algorithm from (Auer et al., 2002a) attains $O(\log n)$ in the stochastic model. Both results

Known parameters: K arms; n rounds ($n \geq K \geq 2$).
Unknown parameters (stochastic model):
K probability distributions $\nu_1, \ldots, \nu_K$ on $[0,1]$ with resp. means $\mu_1, \ldots, \mu_K$.

For each round $t = 1, 2, \ldots, n$;

(1) algorithm chooses $I_t \in \{1, \ldots, K\}$, possibly using external randomization;

(2) adversary simultaneously selects rewards $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0,1]^K$.
- in the stochastic model, each reward $g_{i,t} \sim \nu_i$ is drawn independently.

(3) the forecaster receives (and observes) the reward $g_{I_t,t}$.
He does not observe the rewards from the other arms.

Goal: Minimize the regret, defined in the adversarial model by:

$$R_n = \max_{i \in \{1,\ldots,K\}} \sum_{t=1}^{n} g_{i,t} - \sum_{t=1}^{n} g_{I_t,t},$$

and in the stochastic model by:

$$\overline{R}_n = \sum_{t=1}^{n} \left(\max_{i \in \{1,\ldots,K\}} \mu_i - \mu_{I_t} \right).$$

Figure 1: The MAB framework: adversarial rewards and stochastic rewards.

are essentially optimal. It is worth noting that `UCB1` and `Exp3` have influenced, and to some extent inspired, a number of follow-up papers on richer MAB settings.

However, it is easy to see that `UCB1` incurs a trivial $\Omega(n)$ regret in the adversarial model, whereas `Exp3` has $\Omega(\sqrt{n})$ regret even in the stochastic model.[1] This raises a natural question that we aim to resolve in this paper: *can we achieve the best of both worlds?* Is there a bandit algorithm which matches the performance of `Exp3` in the adversarial model, and attains the performance of `UCB1` if the rewards are in fact stochastic? A more specific (and slightly milder) formulation is as follows:

> Is there a bandit algorithm that has $\tilde{O}(\sqrt{n})$ regret in the adversarial model and $\mathrm{polylog}(n)$ regret in the stochastic model?

We are not aware of any prior work on this question. Intuitively, we introduce a new tradeoff: a bandit algorithm has to balance between *attacking* the weak adversary (stochastic rewards) and *defending* itself from a more devious adversary that targets algorithm's weaknesses, such as being too aggressive if the reward sequence is seemingly stochastic. In particular, while the basic exploration-exploitation tradeoff induces $O(\log n)$ regret in the stochastic model, and $O(\sqrt{n})$ regret in the adversarial model, it is not clear a priori what are the optimal regret guarantees for this new *attack-defense* tradeoff.

1. This is clearly true for the original version of `Exp3` with a mixing parameter. However, this mixing is unnecessary against oblivious adversaries (Stoltz, 2005). The regret of the resulting algorithm in the stochastic model is unknown.

We answer the above question affirmatively, with a new algorithm called SAO (Stochastic and Adversarial Optimal). To formulate our result, we need to introduce some notation. In the stochastic model, let μ_i be the expected single-round reward from arm i. A crucial parameter is the *minimal gap*: $\Delta = \min_{i:\, \mu_i<\mu^*} \mu^* - \mu_i$, where $\mu^* = \max_i \mu_i$. With this notation, `UCB1` attains regret $O(\frac{K}{\Delta} \log n)$ in the stochastic model, where K is the number of arms. We are looking for the following: regret $\mathbb{E}[R_n] = \tilde{O}(\sqrt{Kn})$ in the adversarial model and regret $\mathbb{E}[\overline{R}_n] = \tilde{O}(\frac{K}{\Delta})$ in the stochastic model, where $\tilde{O}(\cdot)$ hides $\mathrm{polylog}(n)$ factors. Our main result is as follows.

Theorem 1 *There exists an algorithm SAO for the MAB problem such that:*

(a) in the adversarial model, SAO achieves regret $\mathbb{E}[R_n] \leq O(\sqrt{nK}\, \log^{3/2}(n)\, \log K)$.

(b) in the stochastic model, SAO achieves regret $\mathbb{E}[\overline{R}_n] \leq O(\frac{K}{\Delta}\, \log^2(n) \log K)$.

Moreover, with very little extra work we can obtain the corresponding high-probability versions (see Theorem 15 for a precise statement).

It is easier, and more instructive, to explain the main ideas on the special case of two arms and oblivious adversary.[2] This special case (with a simplified algorithm) is presented in Section 3. Due to the page limit, the general case is fleshed out in the Appendix.

Discussion. The question raised in this paper touches upon an important theme in Machine Learning, and more generally in the design of algorithms with partially known inputs: how to achieve a good worst-case performance *and* also take advantage of "nice" problem instances. In the context of MAB it is natural to focus on the distinction between stochastic and adversarial rewards, especially given the prominence of the two models in the MAB literature. Then our "best-of-both-worlds" question is the first-order specific question that one needs to resolve. Also, we provide the first analysis of the same MAB algorithm under both adversarial and stochastic rewards.

Once the "best-of-both-worlds" question is settled, several follow-up questions emerge. Most immediately, it is not clear whether the polylog factors can be improved to match the optimal guarantees for each respective model; a lower bound would indicate that the "attack-defence" tradeoff is fundamentally different from the familiar explore-exploit tradeoffs. A natural direction for further work is rewards that are adversarial on a few short time intervals, but stochastic most of the time. Moreover, it is desirable to adapt not only to the binary distinction between the stochastic and adversarial rewards, but also to some form of continuous tradeoff between the two reward models.

Finally, we acknowledge that our solution is no more (and no less) than a theoretical proof of concept. More work, theoretical and experimental, and perhaps new ideas or even new algorithms, are needed for a practical solution. In particular, a practical algorithm should probably go beyond what we accomplish in this paper, along the lines of the two possible extensions mentioned above.

Related work. The general theme of combining worst-case and optimistic performance bounds have received considerable attention in prior work on online learning. A natural incarnation of this theme in the context of MAB concerns proving upper bounds on regret that can be written in terms of some complexity measure of the rewards, and match the optimal worst-case bounds. To this end, a version of `Exp3` achieves regret $\tilde{O}(\sqrt{KG_n^*})$, where $G_n^* \leq n$ is the maximal cumulative reward of a single arm, and the corresponding high probability result was recently proved in Audibert and Bubeck (2010). In Hazan and Kale (2009), the authors obtain regret $\tilde{O}(\sqrt{KV_n^*})$, where $V_n^* \leq n$

2. An oblivious adversary fixes the rewards $g_{i,t}$ for all round t without observing the algorithm's choices.

is the maximal "temporal variation" of the rewards.[3] Similar results have been obtained for the full-feeback ("experts") version in Cesa-Bianchi et al. (2007); Abernethy et al. (2008). Also, the regret bound for UCB1 depends on the gap Δ, and matches the optimal worst-case bound for the stochastic model (up to logarithmic factors). Moreover, adaptivity to "nice" problem instances is a crucial theme in the work on bandits in metric spaces (Kleinberg et al., 2008; Bubeck et al., 2011; Slivkins, 2011), an MAB setting in which some information on similarity between arms is a priori available to an algorithm.

The distinction between $\mathrm{polylog}(n)$ and $\Omega(\sqrt{n})$ regret has been crucial in other MAB settings: bandits with linear rewards (Dani et al., 2008), bandits in metric spaces (Kleinberg and Slivkins, 2010), and an extension of MAB to auctions (Babaioff et al., 2009; Devanur and Kakade, 2009; Babaioff et al., 2010). Interestingly, here we have four different MAB settings (including the one in this paper) in which this distinction occurs for four different reasons, with no apparent connections.

A proper survey of the literature on multi-armed bandits is beyond the scope of this paper; a reader is encouraged to refer to Cesa-Bianchi and Lugosi (2006) for background. An important high-level distinction is between Bayesian and non-Bayesian MAB formulations. Both have a rich literature; this paper focuses on the latter. The "basic" MAB version defined in this paper has been extended in various papers to include additional information and/or assumptions about rewards.

Most relevant to this paper are algorithms UCB1 (Auer et al., 2002a) and Exp3 (Auer et al., 2002b). UCB1 has a slightly more refined regret bound than the one that we cited earlier: $\overline{R}_n = O(\sum_{i:\,\mu_i<\mu^*} \frac{\log n}{\mu^*-\mu_i})$ with high probability. A matching lower bound (up to the considerations of the variance and constant factors) is proved in Lai and Robbins (1985). Several recent papers (Auer and Ortner, 2010; Honda and Takemura, 2010; Audibert et al., 2009; Audibert and Bubeck, 2010; Maillard and Munos, 2011; Garivier and Cappé, 2011; Perchet and Rigollet, 2011) improve over UCB1, obtaining algorithms with regret bounds that are even closer to the lower bound.

The regret bound for Exp3 is $\mathbb{E}[R_n] = O(\sqrt{nK\log K})$, and a version of Exp3 achieves this with high probability (Auer et al., 2002b). There is a nearly matching lower bound of $\Omega(\sqrt{Kn})$. Recently Audibert and Bubeck (2010) have shaved off the $\log K$ factor, achieving an algorithm with regret $O(\sqrt{Kn})$ in the adversarial model against an oblivious adversary.

High-level ideas. For clarity, let us consider the simplified algorithm for the special case of two arms and oblivious adversary. The algorithm starts with the assumption that the stochastic model is true, and then proceeds in three phases: "exploration", "exploitation", and the "adversarial phase". In the exploration phase, we alternate the two arms until one of them (say, arm 1) appears significantly better than the other. When and if that happens, we move to the exploitation phase where we focus on arm 1, but re-sample arm 2 with small probability. After each round we check several consistency conditions which should hold with high probability if the rewards are stochastic. When and if one of these conditions fails, we declare that we are not in the case of stochastic rewards, and switch to running a bandit algorithm for the adversarial model (a version of Exp3).

Here we have an incarnation of the "attack-defense" tradeoff mentioned earlier in this section: the consistency conditions should be (a) strong enough to justify using the stochastic model as an operating assumption while the conditions hold, and (b) weak enough so that we can check them

3. The result in Hazan and Kale (2009) does not shed light on the question in the present paper, because the "temporal variation" concerns actual rewards rather than expected rewards. In particular, temporal variation is minimal when the actual reward of each arm is constant over time, and (essentially) maximal in the stochastic model with 0-1 rewards.

despite the low sampling probability of arm 2. The fact that (a) and (b) are not mutually exclusive is surprising and unexpected.

More precisely, the consistency conditions should be strong enough to insure us from losing too much in the first two phases even if we are in the adversarial model. We use a specific re-sampling schedule for arm 1 which is rare enough so that we do not accumulate much regret if this is indeed a bad arm, and yet sufficient to check the consistency conditions.

To extend to the K-arm case, we "interleave" exploration and exploitation, "deactivating" arms one by one as they turn out to be suboptimal. The sampling probability of a given arm increases while the arm stays active, and then decreases after it is deactivated, with a smooth transition between the two phases. This complicated behavior (and the fact that we handle general adversaries) in turn necessitate a more delicate analysis.

2. Preliminaries

We consider randomized algorithms, in the sense that I_t (the arm chosen at time t) is drawn from a probability distribution p_t on $\{1, \ldots, K\}$. We denote by $p_{i,t}$ the probability that $I_t = i$. For brevity, let $I_{i,t} = \mathbb{1}_{\{I_t = i\}}$. Given such a randomized algorithm, it is a well-known trick to use $\tilde{g}_{i,t} = \frac{g_{i,t}\, I_{i,t}}{p_{i,t}}$ as an unbiased estimate of the reward $g_{i,t}$. Now for arm i and time t we introduce:

- $G_{i,t} = \sum_{s=1}^{t} g_{i,s}$ (*fixed-arm cumulative reward* from arm i up to time t),
- $\widetilde{G}_{i,t} = \sum_{s=1}^{t} \tilde{g}_{i,s}$ (*estimated cumulative reward* from arm i up to time t),
- $\widehat{G}_{i,t} = \sum_{s=1}^{t} g_{i,s}\, I_{i,s}$ (*algorithm's cumulative reward* from arm i up to time t),
- $T_i(t) = \sum_{s=1}^{t} I_{i,s}$ (the *sampling time* of arm i up to time t).
- The corresponding averages: $H_{i,t} = \frac{1}{t}\, G_{i,t}$, $\widetilde{H}_{i,t} = \frac{1}{t}\, \widetilde{G}_{i,t}$, and $\widehat{H}_{i,t} = \widehat{G}_{i,t}/T_i(t)$.

$G_{i,t}$ is the cumulative reward of a "fixed-arm algorithm" that always plays arm i. Recall that our benchmarks are $\max_i G_{i,t}$ for the adversarial model, and $\max_i \mathbb{E}[G_{i,t}]$ for the stochastic model.

Note that $\widetilde{H}_{i,t}$, $\widehat{H}_{i,t}$ (and $\widetilde{G}_{i,t}$, $\widehat{G}_{i,t}$) are observed by an algorithm whereas $H_{i,t}$ (and $G_{i,t}$) is not. Informally, $\widetilde{H}_{i,t}$ and $\widehat{H}_{i,t}$ are estimates for the expected reward μ_i in the stochastic model, and $\widetilde{H}_{i,t}$ is an estimate for the benchmark reward $H_{i,t}$ in the adversarial model.

In the stochastic model we define the *gap* of arm i as $\Delta_i = (\max_{1 \le j \le K} \mu_j) - \mu_i$, and the *minimal gap* $\Delta = \min_{i:\, \Delta_i > 0} \Delta_i$.

Following the literature, we measure algorithm's performance in terms of *regret* R_n and $\overline{R}_n$ as defined in Figure 1. The two notions of regret are somewhat different, in particular the "stochastic regret" $\overline{R}_n$ is *not* exactly equal to the expected "adversarial regret" R_n. However, in the stochastic model they are approximately equal:[4] $\mathbb{E}[\overline{R}_n] \le \mathbb{E}[R_n] \le \mathbb{E}[\overline{R}_n] + \sqrt{\frac{1}{2}\, n \log K}$.

3. The case of $K = 2$ arms

We will derive a (slightly weaker version of) the main result for the special case of $K = 2$ arms and oblivious adversary, using a simplified algorithm. This version contains most of the ideas from the general case, but can be presented in a more lucid fashion.

4. This fact is well-known and easy to prove, e.g. see Proposition 34 in Audibert and Bubeck (2010).

We are looking for the "best-of-both-worlds" feature: $\tilde{O}(\sqrt{n})$ regret in the adversarial model, and $\tilde{O}(\frac{1}{\Delta})$ regret in the stochastic model, where $\Delta = |\mu_1 - \mu_2|$ is the gap. Our goal in this section is to obtain this feature in the simplest way possible. In particular, we will hide the constants under the $O()$ notation, and will not attempt to optimize the polylog(n) factors; also, we will assume oblivious adversary. We will prove the following theorem:

Theorem 2 *Consider a MAB problem with two arms. There exists a algorithm such that:*

(a) against an oblivious adversary, its expected regret is $\mathbb{E}[R_n] \leq O(\sqrt{n}\,\log^2 n)$.

(b) in the stochastic model, its expected regret satisfies $\mathbb{E}[R_n] \leq O(\frac{1}{\Delta}\,\log^3 n)$.

Both regret bounds also hold with probability at least $1 - \frac{1}{n}$.

Note that in the stochastic model, regret trivially cannot be larger than Δn, so part (b) trivially implies regret $\mathbb{E}[R_n] \leq \tilde{O}(\sqrt{n})$.

Our analysis proceeds via high-probability arguments and directly obtains the high-probability guarantees. The high-probability arguments tend to make the analysis cleaner; we suspect it cannot be made much simpler if we only seek bounds on expected regret.

3.1. A simplified SAO (Stochastic and Adversarial Optimal) Algorithm

The algorithm proceeds in three phases: exploration, exploitation, and adversarial phase. In the exploration phase, we alternate the two arms until one of them appears significantly better than the other. In exploitation phase, we focus on the better arm, but re-sample the other arm with small probability. We check several consistency conditions which should hold with high probability if the rewards are stochastic. When and if one of these conditions fails, we declare that we are not in the case of stochastic rewards, and switch to running a bandit algorithm for the adversarial model, namely algorithm `Exp3.P` (Auer et al., 2002b).

The algorithm is parameterized by $C_{\texttt{crn}} = \Theta(\log n)$ which we will chose later in Section 3.2. The formal description of the three phases is as follows.

(Exploration phase) In each round t, pick an arm at random: $p_{1,t} = p_{2,t} = \frac{1}{2}$. Go to the next phase as soon as $t > \Omega(C^2_{\texttt{crn}})$ and the following condition fails:

$$|\widetilde{H}_{1,t} - \widetilde{H}_{2,t}| < 24\, C_{\texttt{crn}}/\sqrt{t}. \tag{1}$$

Let τ_* be the duration of this phase. Without loss of generality, assume $\widetilde{H}_{1,\tau_*} > \widetilde{H}_{2,\tau_*}$. This means, informally, that arm 1 is selected for exploitation.

(Exploitation phase) In each round $t > \tau_*$, pick arm 2 with probability $p_{2,t} = \frac{\tau_*}{2t}$, and arm 1 with the remaining probability $p_{1,t} = 1 - \frac{\tau_*}{2t}$.

After the round, check the following *consistency conditions*:

$$8\, C_{\texttt{crn}}/\sqrt{\tau_*} \leq \widetilde{H}_{1,t} - \widetilde{H}_{2,t} \leq 40\, C_{\texttt{crn}}/\sqrt{\tau_*} \tag{2}$$

$$\begin{cases} |\widetilde{H}_{1,t} - \widehat{H}_{1,t}| \leq 6\, C_{\texttt{crn}}/\sqrt{t} \\ |\widetilde{H}_{2,t} - \widehat{H}_{2,t}| \leq 6\, C_{\texttt{crn}}/\sqrt{\tau_*}\,. \end{cases} \tag{3}$$

If one of these conditions fails, go to the next phase.

(Adversarial phase) Run algorithm `Exp3.P` from Auer et al. (2002b).

Discussion. The exploration phase is simple: Condition (1) is chosen so that once it fails then (assuming stochastic rewards) the seemingly better arm is indeed the best arm with high probability.

In the exploitation phase, we define the re-sampling schedule for arm 2 and a collection of "consistency conditions". The re-sampling schedule should be sufficiently rare to avoid accumulating much regret if arm 2 is indeed a bad arm. The consistency conditions should be sufficiently strong to justify using the stochastic model as an operating assumption while they hold. Namely, an adversary constrained by these conditions should not be able to inflict too much regret on our algorithm in the first two phases. Yet, the consistency conditions should be weak enough so that they hold with high-probability in the stochastic model, despite the low sampling probability of arm 2.

It is essential that we use both $\widehat{H}_{i,t}$ and $\widetilde{H}_{i,t}$ in the consistency conditions: the interplay of these two estimators allows us to bound regret in the adversarial model. Other than that, the conditions that we use are fairly natural (the surprising part is that they work). Condition (2) checks whether the relation between the two arms is consistent with the outcome of the exploration phase, i.e. whether arm 1 still seems better than arm 2, but not *too much* better. Condition (3) checks whether for each arm i, the estimate $\widetilde{H}_{i,t}$ is close to the average $\widehat{H}_{i,t}$. In the stochastic model, both estimate the expected gain μ_i, so we expect them to be not too far apart. However, our definition of "too far" should be consistent with how often a given arm is sampled.

3.2. Concentration inequalities

The "probabilistic" aspect of the analysis is confined to proving that several properties of estimates and sampling times hold with high probability. The rest of the analysis can proceed *as if* these properties hold with probability 1. In particular, we have made our core argument essentially deterministic, which greatly simplifies presentation.

All high-probability results are obtained using an elementary concentration inequality loosely known as *Chernoff Bounds*. For the sake of simplicity, we use a slightly weaker formulation below (see Appendix A for a proof), which uses just one inequality for all cases.

Theorem 3 (Chernoff Bounds) *Let X_t, $t \in [n]$ be a independent random variables such that $X_t \in [0,1]$ for each t. Let $X = \sum_{t=1}^{n} X_t$ be their sum, and let $\mu = \mathbb{E}[X]$. Then*

$$\Pr\left[\,|X-\mu| > C\,\max(1,\sqrt{\mu})\,\right] < 2\,e^{-C/3}, \quad \text{for any } C > 1. \tag{4}$$

We will often need to apply Chernoff Bounds to sums whose summands depend on some events in the execution of the algorithms and therefore are not mutually independent. However, in all cases these issues are but a minor technical obstacle which can be side-stepped using a slightly more careful setup.[5] In particular, we sometimes find it useful to work in the probability space obtained by conditioning on the outcome of the exploration phase. Specifically, the *post-exploration probability space* is the probability space obtained by conditioning on the following events: that the exploration phase ends, that it has a specific duration τ_*, and that arm 1 is chosen for exploitation.

Throughout the analysis, we will obtain concentration bounds that hold with probability at least $1-2n^{-4}$. We will often take a Union Bound over all rounds t, which will imply success probability at least $1-2n^{-3}$. To simplify presentation, we will allow a slight abuse of notation: we will say

5. However, the independence issues appear prohibitive for $K > 2$ arms or if we consider a non-oblivious adversary. So for the general case we resorted to a more complicated analysis via martingale inequalities.

with high probability (abbreviated *w.h.p.*), which will mean mean with probability at least $1-2n^{-3}$ or at least $1-2n^{-4}$, depending on the context.

To parameterize the algorithm, let us fix some $C_{\texttt{crn}} = 12\,\ln(n)$ such that Theorem 3 with $C = C_{\texttt{crn}}$ ensures success probability at least $1-2\,n^{-4}$.

3.3. Analysis: adversarial model

We need to analyze our algorithm in two different reward models. We start with the adversarial model, so that we can re-use some of the claims proved here to analyze the stochastic model.

Recall that τ_* denotes the duration of the exploration phase (which in general is a random variable). Following the convention from Section 3.1 that whenever the exploration phase ends, the arm chosen for exploitation is arm 1. (Note that we do not assume that arm 1 is the best arm.)

We start the analysis by showing that the re-sampling schedule in the exploitation phase does not result in playing arm 2 too often.

Claim 4 *During the exploitation phase, arm 2 is played at most $O(\tau_* \log n)$ times w.h.p..*

Proof. We will work in the post-exploration probability space $\mathcal{S}$. We need to bound from above the sum $\sum_t I_{2,t}$, where t ranges over the exploitation phase. However, Chernoff Bounds do not immediately apply since the number of summands itself is a random variable. Further, if we condition on a specific duration of exploitation then we break independence between summands. We sidestep this issue by considering an alternative algorithm in which exploitation lasts indefinitely (i.e., without the stopping conditions), and which uses the same randomness as the original algorithm. It suffices to bound from above the number of times that arm 2 is played during the exploitation phase in this alternative algorithm; denote this number by N. Letting J_t be the arm selected in round t of the alternative algorithm, we have that $N = \sum_{t=\tau_*+1}^{n} \mathbb{1}_{\{J_t=i\}}$ is a sum of 0-1 random variables, and in $\mathcal{S}$ these variables are independent. Moreover, in $\mathcal{S}$ it holds that

$$\mathbb{E}[N] = \textstyle\sum_{t=\tau_*+1}^{n} p_{2,t} = \tau_* \sum_{t=\tau_*+1}^{n} \frac{1}{2t} = O(\tau_* \log n).$$

Therefore, the claim follows from Chernoff Bounds. □

Now we connect the estimated cumulative rewards $\widetilde{G}_{i,t}$ with the benchmark $G_{i,t}$. More specifically, we will bound from above several expressions of the form $|\widetilde{H}_{i,t} - H_{i,t}|$. Naturally, the upper bound for arm 1 will be stronger since this arm is played more often during exploitation. To ensure that the bound for arm 2 is strong enough we need to play this arm "sufficiently often" during exploitation. (Whereas Claim 4 ensures that we do not play it "too often".) Here and elsewhere in this analysis, we find it more elegant to express some of the claims in terms of the average cumulative rewards (such as $H_{i,t}$, etc.)

Claim 5 *(a) With high probability, $|\widetilde{H}_{i,\tau_*} - H_{i,\tau_*}| < 2\,C_{\texttt{crn}}/\sqrt{\tau_*}$ for each arm i.*
(b) For any round t in the exploitation phase, with high probability it holds that

$$\begin{cases} |\widetilde{H}_{1,t} - H_{1,t}| < 3\,C_{\texttt{crn}}/\sqrt{t}, \\ |\widetilde{H}_{2,t} - H_{2,t}| < 3\,C_{\texttt{crn}}/\sqrt{\tau_*}. \end{cases} \tag{5}$$

Proof. For part (a), we are interested in the sum $\sum_{t\leq\tau_*}\ g_{i,t}\ I_{i,t}$. As in the proof of Claim 4, Chernoff Bounds do not immediately apply since the number of summands τ_* is a random variable (and conditioning on a particular value of τ_* tampers with independence between summands). So let us consider an alternative algorithm in which the exploration phase proceeds indefinitely, without the stopping condition, and uses the same randomness as the original algorithm.[6] Let J_t be the arm selected in round t of this alternative algorithm, and define $A_{i,t} = \sum_{s=1}^{t} g_{i,t}\,\mathbb{1}_{\{J_t=i\}}$. Then (when run on the same problem instance) both algorithms coincide for any $t \leq \tau_*$, so in particular $\widetilde{G}_{i,t} = 2\,A_{i,t}$. Now, $A_{i,t}$ is the sum of bounded independent random variables with expectation $G_{i,t}/2$. Therefore by Chernoff Bounds w.h.p. it holds that $|A_{i,t} - G_{i,t}/2| < C_{\texttt{crn}}\sqrt{t}$ for each t, which implies the claim.

For part (b), we will analyze the exploitation phase separately. Let us work in the post-exploration probability space $\mathcal{S}$. We will consider the alternative algorithm from the proof of Claim 4 (in which exploitation continues indefinitely). This way we do not need worry that we implicitly condition on the event that a particular round $t > \tau_*$ belongs to the exploitation phase. Clearly, it suffices to prove (5) for this alternative algorithm. To facilitate the notation, define the time interval $\texttt{INT} = \{\tau_*+1,\ldots,t\}$, and denote $G_{i,\texttt{INT}} = \sum_{s\in\texttt{INT}} g_{i,t}$ and $\widetilde{G}_{i,\texttt{INT}} = \sum_{s\in\texttt{INT}} \widetilde{g}_{i,t}$.

To handle arm 1, note that in $\mathcal{S}$, $\widetilde{G}_{i,\texttt{INT}}$ is a sum of independent random variables, with expectation $G_{i,\texttt{INT}}$. Since $p_{1,t} \geq \frac{1}{2}$ for any $t \in \texttt{INT}$, the summands $\widetilde{g}_{1,t}$ are bounded by 2. Therefore by Chernoff Bounds with high probability it holds that

$$|\widetilde{G}_{1,\texttt{INT}} - G_{1,\texttt{INT}}| < 2\,C_{\texttt{crn}}\sqrt{t-\tau_*}.$$

From this and part (a) it follows that w.h.p.

$$|\widetilde{G}_{1,t} - G_{1,t}| < 2\,C_{\texttt{crn}}(\sqrt{\tau_*} + \sqrt{t-\tau_*}) < 3\,C_{\texttt{crn}}\sqrt{t},$$

which implies the claim for arm 1.

Handling arm 2 requires a little more work since the summands $\widetilde{g}_{2,t}$ may be large (since they have a small probability $p_{2,t}$ in the denominator). For each $t \in \texttt{INT}$,

$$\widetilde{G}_{2,\texttt{INT}} = \sum_{s\in\texttt{INT}} \frac{2s}{\tau_*}\ g_{2,s}\ I_{2,s} = \frac{2t}{\tau_*}\sum_{s\in\texttt{INT}} \frac{s}{t}\ g_{2,s}\ I_{2,s} = \frac{2t}{\tau_*}\sum_{s\in\texttt{INT}} X_s,$$

where $X_s = \frac{s}{t}\ g_{2,s}\,I_{2,s} \in [0,1]$. In $\mathcal{S}$, random variables X_s, $s \in \texttt{INT}$ are mutually independent, and the expectation of their sum is

$$\mu \triangleq \mathbb{E}\left[\sum_{s\in\texttt{INT}} X_s\right] = \tfrac{\tau_*}{2t}\ \mathbb{E}\left[\widetilde{G}_{2,\texttt{INT}}\right] = \tfrac{\tau_*}{2t}\ G_{2,\texttt{INT}} \leq \tfrac{\tau_*}{2}\ \tfrac{t-\tau_*}{t}.$$

Noting that $G_{2,\texttt{INT}} \leq t-\tau_*$ and letting $\alpha = \frac{\tau_*}{t}$, we obtain $\mu \leq \frac{\tau_*}{2}(1-\alpha)$. By Chernoff Bounds w.h.p. it holds that

$$|\textstyle\sum_{s\in\texttt{INT}} X_s - \mu| < C_{\texttt{crn}}\sqrt{\tau_*\,(1-\alpha)}.$$

Going back to $\widetilde{G}_{2,\texttt{INT}}$ and $G_{2,\texttt{INT}}$, we obtain:

$$|\widetilde{G}_{2,\texttt{INT}} - G_{2,\texttt{INT}}| < \frac{2t}{\tau_*}\,C_{\texttt{crn}}\sqrt{\tau_*\,(1-\alpha)} < C_{\texttt{crn}}\,\frac{2t}{\sqrt{\tau_*}}\,\sqrt{1-\alpha}.$$

6. Note that this is not the same "alternative algorithm" as the one in the proof of Claim 4.

From part (a), we have that $|\widetilde{G}_{i,\tau_*} - G_{i,\tau_*}| < C_{\texttt{crn}} \frac{2t}{\sqrt{\tau_*}} \alpha$. Therefore,

$$|\widetilde{G}_{2,t} - G_{2,t}| < C_{\texttt{crn}} \frac{2t}{\sqrt{\tau_*}} (\sqrt{\alpha} + \sqrt{1-\alpha}) < C_{\texttt{crn}} \frac{3t}{\sqrt{\tau_*}}. \quad \square$$

Combining Claim 5(b) and Condition (2), we obtain:

Corollary 6 *In the exploitation phase, for any round t (except possibly the very last round in the phase) it holds w.h.p. that $G_{1,t} > G_{2,t}$.*

By Corollary 6, regret accumulated by round t in the exploitation phase is, with high probability, equal to $G_{1,t} - \widehat{G}_{1,t} - \widehat{G}_{2,t}$. The following claim upper-bounds this quantity by $O(\sqrt{t}\,\log^2 n)$. The proof of this claim contains our main regret computation.

Claim 7 *For any round t in the exploitation phase it holds w.h.p. that*

$$\widehat{G}_{1,t} + \widehat{G}_{2,t} - G_{1,t} \geq -O(\sqrt{t}\,\log^2 n).$$

Proof. Throughout this proof, let us assume that the high-probability events in Claim 4 and Claim 5(b) actually hold; we will omit "with high probability" from here on.

Let t be some (but not the last) round in the exploitation phase. First,

$$\widehat{H}_{1,t} - H_{1,t} = \left[\widetilde{H}_{1,t} - H_{1,t}\right] + \left[\widehat{H}_{1,t} - \widetilde{H}_{1,t}\right] \geq -O(C_{\texttt{crn}}/\sqrt{t}). \tag{6}$$

We have upper-bounded the two square brackets in (6) using, respectively, Claim 5(b) and Condition (3). We proved that algorithm's average for arm 1 ($\widehat{H}_{1,t}$) is not too small compared to the corresponding benchmark average $H_{1,t}$, and we used the estimate $\widetilde{H}_{1,t}$ as an intermediary in the proof.

Similarly, using Condition (2), Condition (3), and Claim 5(b) to upper-bound the three square brackets in the next equation, we obtain that

$$\widehat{H}_{2,t} - H_{1,t} = \left[\widetilde{H}_{2,t} - \widetilde{H}_{1,t}\right] + \left[\widehat{H}_{2,t} - \widetilde{H}_{2,t}\right] + \left[\widetilde{H}_{1,t} - H_{1,t}\right] \geq -O(C_{\texttt{crn}}/\sqrt{\tau_*}). \tag{7}$$

Here we have proved that the algorithm did not do too badly playing arm 2, even though this arm was supposed to be suboptimal. Specifically, we establish that algorithm's average for arm 2 ($\widehat{H}_{2,t}$) is not too small compared to the benchmark average *for arm 1* ($H_{1,t}$). Again, the estimates $\widetilde{H}_{1,t}$ and $\widetilde{H}_{2,t}$ served us as intermediaries in the proof.

Finally, let us go from bounds on average rewards to bounds on cumulative rewards (and prove the claim). Combining (6), (7) and Claim 4, we have:

$$\begin{aligned}
\widehat{G}_{1,t} + \widehat{G}_{2,t} - G_{1,t} &= \textstyle\sum_{i=1,2} T_j(t) \left[\widehat{H}_{i,t} - H_{1,t}\right] \\
&\geq -O(C_{\texttt{crn}}) \left[T_1(t)/\sqrt{t} + T_2(t)/\sqrt{\tau_*}\right] \\
&\geq -O(C_{\texttt{crn}})(\sqrt{t} + \sqrt{\tau_*}\log n) \\
&\geq -O(\sqrt{t}\,\log^2 n). \qquad \square
\end{aligned}$$

Now we are ready for the final computations. We will need to consider three cases, depending on which phase the algorithm is in when it halts (i.e., reaches the time horizon).

First, if the exploration phase never ends then by Claim 5(a) w.h.p. it holds that $|\widetilde{H}_{i,n} - H_{i,n}| < 2\,C_{\texttt{crn}}/\sqrt{n}$ for each arm i, and the exit condition (1) never fails. This implies the claimed regret bound $R_n \leq O(\sqrt{n}\log n)$.

From here on let us assume that the exploration phase ends at some $\tau_* < n$. Define regret on the time interval $[a,b]$ as

$$R_{[a,b]} = \max_{i\in\{1,2\}} \sum_{a=1}^{b} g_{i,t} - \sum_{s=a}^{b} g_{I_t,t}.$$

Let t be the last round in the exploitation phase. By Corollary 6 and Claim 7 we have

$$R_{[1,t-1]} = G_{1,t} - \widehat{G}_{1,t} - \widehat{G}_{2,t} \leq O(\sqrt{n}\,\log^2 n).$$

If $t = n$ (i.e., the algorithm halts during exploitation) then we are done.

Third, if the algorithm enters the adversarial phase then we can use the regret bound for `Exp3.P` in Auer et al. (2002b), which states that w.h.p. $R_{[t,n]} \leq O(\sqrt{n})$. Therefore

$$R_n \leq R_{[1,t-1]} + R_{[t,n]} \leq O(\sqrt{n}\,\log^2 n).$$

This completes the proof of Theorem 2(a).

3.4. Analysis: stochastic model

We start with a simple claim that w.h.p. each arm is played sufficiently often during exploration, and arm 1 is played sufficiently often during exploitation. This claim complements Claim 4 from the previous subsection which states that arm 2 is not played too often during exploitation (we will re-use Claim 4 later in the proofs).

Claim 8 *With high probability it holds that:*

(a) during the exploration phase, each arm is played at least $\tau_*/4$ times.

(b) during the exploitation phase, $T_1(t) \geq t/4$ for each time t.

Proof. Both parts follow from Chernoff Bounds. The only subtlety is to ensure that we do not condition the summands (in the sum that we apply the Chernoff Bounds to) on a particular value of τ_* or on the fact that arm 1 is chosen for exploitation.

For part (a), without loss of generality assume that n fair coins are tossed in advance, so that in the t-th round of exploration we use the t-th coin toss to decide which arm is chosen. Then by Chernoff Bounds for each t w.h.p. it holds that among the first t coin tosses there are at least $t/2 - C_{\texttt{crn}}\sqrt{t/2}$ heads and at least this many tails. We take the Union Bound over all t, so in particular this holds for $t = \tau_*$. Therefore w.h.p. we have:

$$T_i(\tau_*) \geq \tau_*/2 - C_{\texttt{crn}}\sqrt{\tau_*/2}. \tag{8}$$

The claim follows from (8) because we force exploration to last for at least $\Omega(C^2_{\texttt{crn}})$ rounds.

For part (b), let us analyze the exploitation phase separately. We are interested in the sum $\sum_s I_{1,s}$, where s ranges over all rounds in the exploitation phase. We will work in the post-exploration probability space. The indicator variables $I_{1,s}$, for all rounds s during exploitation, are

mutually independent. Therefore Chernoff Bounds apply, and w.h.p. $T_1(t) - T_1(\tau_*) \geq (t-\tau_*)/2 - C_{\texttt{crn}}\sqrt{t-\tau_*}$. Using (8), it follows that $T_1(t) \geq t/2 - C_{\texttt{crn}}(\sqrt{\tau_*} + \sqrt{t-\tau_*}) \geq t/2 - C_{\texttt{crn}}\sqrt{t} \geq t/4$. □

Recall that Claim 5(b) connects algorithm's estimate $\widetilde{H}_{i,t}$ and the benchmark average $H_{i,t}$ (we will re-use this claim later in the proofs). In the stochastic model these two quantities, as well as the algorithm's average $\widehat{H}_{i,t}$, are close to the respective expected reward μ_i. The following lemma makes this connection precise.

Claim 9 *Assume the stochastic model. Then during the exploitation phase for each arm i and each time t the following holds with high probability:*

$$\begin{cases} |H_{i,t} - \mu_i| \leq C_{\texttt{crn}}/\sqrt{t}, \\ |\widehat{H}_{1,t} - \mu_1| \leq 2\, C_{\texttt{crn}}/\sqrt{t}, \\ |\widehat{H}_{2,t} - \mu_2| \leq 2\, C_{\texttt{crn}}/\sqrt{\tau_*}. \end{cases}$$

Proof. All three inequalities follow from Chernoff Bounds. The first inequality follows immediately. To obtain the other two inequalities, we claim that w.h.p. it holds that

$$|\widehat{H}_{i,t} - \mu_i| \leq C_{\texttt{crn}}/\sqrt{T_i(t)}. \tag{9}$$

Indeed, note that without loss of generality T independent samples from the reward distribution of arm i are drawn in advance, and then the reward from the ℓ-th play of arm i is the ℓ-th sample. Then by Chernoff Bounds the bound (9) holds w.h.p. for each $T_i(t) = l$, and then one can take the Union Bound over all l to obtain (9). Claim proved.

Finally, we use (9) and plug in the lower bounds on $T_i(t)$ from Claim 8(ab). □

Now that we have all the groundwork, let us argue that in the stochastic model the consistency condition in the algorithm are satisfied with high probability.

Corollary 10 *Assume the stochastic model. Then in each round t of the exploitation phase, with high probability the following holds:*

$$16\, C_{\texttt{crn}}/\sqrt{\tau_*} \leq \mu_1 - \mu_2 \leq 32\, C_{\texttt{crn}}/\sqrt{\tau_*}. \tag{10}$$

Moreover, conditions (2-3) are satisfied.

Proof. Condition (3) follows simply by combining Claim 5(b) and Claim 9.

To obtain (10), we note that by Claim 5(b) and Claim 9 w.h.p. it holds that

$$|\widetilde{H}_{1,t} - \mu_1| + |\widetilde{H}_{2,t} - \mu_2| \leq 8\, C_{\texttt{crn}}/\sqrt{\tau_*}. \tag{11}$$

Recall that Condition (1) holds at time $t = \tau_* - 1$, and fails at $t = \tau_*$. This in conjunction with (11) (for $t = \tau_*$) implies (10). In turn, (10) with (11) imply Condition (2). □

To complete the proof of Theorem 2(b), assume we are in the stochastic model with gap $\Delta = |\mu_1 - \mu_2|$. In the rest of the argument, we omit "with high probability". If the exploration phase never ends, it is easy to see that $\Delta \leq O(1/\sqrt{n})$, and we are done since trivially $\overline{R}_n \leq \Delta\, n \leq O(\frac{1}{\Delta})$. Else, by Corollary 10 it holds that arm 1 is optimal, $\tau_* = \Theta(C_{\texttt{crn}}/\Delta)^2$ and moreover that the exploitation phase never ends. Now, by Claim 4 in the exploitation phase the suboptimal arm 2 is played at most $O(\tau_* \log n)$ times. Therefore $\overline{R}_n \leq O(\frac{1}{\Delta}\, \log^3 n)$.

References

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the Dark: An Efficient Algorithm for Bandit Linear Optimization. In *21th Conf. on Learning Theory (COLT)*, pages 263–274, 2008.

J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.

J.Y. Audibert and S. Bubeck. Regret Bounds and Minimax Policies under Partial Monitoring. *J. of Machine Learning Research (JMLR)*, 11:2785–2836, 2010. A preliminary version has been published in *COLT 2009*.

P. Auer and R. Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61:55–65, 2010.

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

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

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *10th ACM Conf. on Electronic Commerce (EC)*, pages 79–88, 2009.

Moshe Babaioff, Robert Kleinberg, and Aleksandrs Slivkins. Truthful mechanisms with implicit payment computation. In *11th ACM Conf. on Electronic Commerce (EC)*, pages 43–52, 2010. Best Paper Award.

S. Bubeck. *Bandits Games and Clustering Foundations*. PhD thesis, Université Lille 1, 2010.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvari. Online Optimization in X-Armed Bandits. *J. of Machine Learning Research (JMLR)*, 12:1587–1627, 2011. Preliminary version in *NIPS 2008*.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge Univ. Press, 2006.

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007. Preliminary version in *COLT 2005*.

Varsha Dani, Thomas P. Hayes, and Sham Kakade. Stochastic Linear Optimization under Bandit Feedback. In *21th Conf. on Learning Theory (COLT)*, 2008.

Nikhil Devanur and Sham M. Kakade. The price of truthfulness for pay-per-click auctions. In *10th ACM Conf. on Electronic Commerce (EC)*, pages 99–106, 2009.

D. A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3:100–118, 1975.

Aurélien Garivier and Olivier Cappé. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *24th Conf. on Learning Theory (COLT)*, 2011.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. In *20th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 38–47, 2009.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *23rd annual conference on learning theory*, 2010.

Robert Kleinberg and Aleksandrs Slivkins. Sharp Dichotomies for Regret Minimization in Metric Spaces. In *21st ACM-SIAM Symp. on Discrete Algorithms (SODA)*, 2010.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-Armed Bandits in Metric Spaces. In *40th ACM Symp. on Theory of Computing (STOC)*, pages 681–690, 2008.

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

Odalric-Ambrym Maillard and Rémi Munos. Adaptive Bandits: Towards the best history-dependent strategy. In *24th Conf. on Learning Theory (COLT)*, 2011.

Colin McDiarmid. Concentration. In M. Habib. C. McDiarmid. J. Ramirez and B. Reed, editors, *Probabilistic Methods for Discrete Mathematics*, pages 195–248. Springer-Verlag, Berlin, 1998.

V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *Arxiv preprint arXiv:1110.6084*, 2011.

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

Aleksandrs Slivkins. Contextual Bandits with Similarity Information. In *24th Conf. on Learning Theory (COLT)*, 2011.

G. Stoltz. *Incomplete Information and Internal Regret in Prediction of Individual Sequences*. PhD thesis, Université Paris-Sud, Orsay, France, May 2005.

Appendix A. Concentration inequalities

Recall that the analysis in Section 3 relies on Chernoff Bounds as stated in Theorem 3. Let us derive Theorem 3 from a version of Chernoff Bounds that can be found in the literature.

Theorem 11 (Chernoff Bounds: Theorem 2.3 in McDiarmid (1998)) *Consider n i.i.d. random variables $X_1 \dots X_n$ on $[0,1]$. Let $X = \frac{1}{n}\sum_{t=1}^{n} X_t$ be their average, and let $\mu = \mathbb{E}[X]$. Then for any $\varepsilon > 0$ the following two properties hold:*

(a) $\Pr[X \geq (1+\varepsilon)\mu] < \exp\left(-\frac{\varepsilon^2\mu}{2(1+\varepsilon/3)}\right) < \begin{cases} e^{-\varepsilon^2\mu/3}, & \varepsilon \leq 1 \\ e^{-\varepsilon\mu/3}, & \text{otherwise.} \end{cases}$

(b) $\Pr[X \leq (1-\varepsilon)\mu] < e^{-\varepsilon^2\mu/2}$.

Corollary 12 *In the setting of Theorem 11, for any $\beta > 0$ we have:*

$$\Pr\left[\,|X-\mu| > \beta\,\max(\beta,\sqrt{\mu})\,\right] < 2\,e^{-\beta^2/3}. \tag{12}$$

We obtain Theorem 3 by taking $\beta = \sqrt{C}$, noting that $\beta\,\max(\beta,\sqrt{\mu}) \leq C\max(1,\sqrt{\mu})$ for $C > 1$.

Proof. Fix $\beta > 0$ and consider two cases: $\mu \geq \beta^2$ and $\mu < \beta^2$.

If $\mu \geq \beta^2$ then we can take $\varepsilon = \beta/\sqrt{\mu} \leq 1$ in Theorem 11(ab) and obtain

$$\Pr[|X - \mu| \geq \beta\sqrt{\mu}] = \Pr[|X - \mu| \geq \varepsilon\mu] < 2\, e^{-\varepsilon^2\mu/3} = 2\, e^{-\beta^2/3}.$$

Now assume $\mu < \beta^2$. We can still take $\varepsilon = \beta/\sqrt{\mu}$ in Theorem 11(b) to obtain

$$\Pr[X - \mu \leq -\beta^2] \leq \Pr[X - \mu \leq -\beta\sqrt{\mu}] < e^{-\varepsilon^2\mu/2} = e^{-\beta^2/2}.$$

Then let us take $\varepsilon = \beta^2/\mu > 1$ in Theorem 11(a) to obtain

$$\Pr[X - \mu \geq \beta^2] = \Pr[X - \mu \geq \varepsilon\mu] < e^{-\varepsilon\mu/3} = e^{-\beta^2/3}.$$

It follows that $\Pr[\,|X - \mu| \geq \beta^2\,] < 2\, e^{-\beta^2/3}$, completing the proof. □

Next we present two important concentration inequalities for martingale sequences that we use in the analysis of the general case ($K > 1$ arms).

Theorem 13 (Hoeffding-Azuma's inequality for martingales, Hoeffding (1963)) *Let $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ real random variables such that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}(X_t|\mathcal{F}_{t-1}) = 0$ and $X_t \in [A_t, A_t + c_t]$ where A_t is a random variable $\mathcal{F}_{t-1}$-measurable and c_t is a positive constant. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{P}\Big(\sum_{t=1}^{n} X_t \geq \varepsilon\Big) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{t=1}^{n} c_t^2}\right), \tag{13}$$

or equivalently for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\sum_{t=1}^{n} X_t \leq \sqrt{\frac{\log(\delta^{-1})}{2} \sum_{t=1}^{n} c_t^2}. \tag{14}$$

Theorem 14 (Bernstein's inequality for martingales, Freedman (1975)) *Let $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ real random variables such that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}(X_t|\mathcal{F}_{t-1}) = 0$, $|X_t| \leq b$ for some $b > 0$ and let $V_n = \sum_{t=1}^{n} \mathbb{E}(X_t^2|\mathcal{F}_{t-1})$. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{P}\Big(\sum_{t=1}^{n} X_t \geq \varepsilon \text{ and } V_n \leq V\Big) \leq \exp\left(-\frac{\varepsilon^2}{2V + 2b\varepsilon/3}\right), \tag{15}$$

and for any $\delta > 0$, with probability at least $1 - \delta$, we have either $V_n > V$ or

$$\sum_{t=1}^{n} X_t \leq \sqrt{2V \log(\delta^{-1})} + \frac{b \log(\delta^{-1})}{3}. \tag{16}$$

Appendix B. SAO (Stochastic and Adversarial Optimal algorithm)

In this section we treat the general case: K arms and adaptive adversary. The proposed algorithm, SAO, is described precisely in Algorithm 1 (see page 17). On a high-level, SAO proceeds similarly to the simplified version in Section 3, but there are a few key differences.

First, the exploration and exploitation phases are now interleaved. Indeed, SAO starts with all arms being "active", and then it successively "deactivates" them as they turn out to be suboptimal. Thus, the algorithm evolves from pure exploration (when all arms activated) to pure exploitation (when all arms but the optimal one are deactivated).

Second, in order to make the above evolution smooth we adopt a more complicated (re)sampling schedule that the one we used in Section 3. Namely, the probability of selecting a given arm continuously increases while this arm stays active, and then continuously decreases when it gets deactivated, and the transition between the two phases is also continuous. For the precise equation, see Equation (21) in Algorithm 1.

Third, this more subtle behavior of the (re)sampling probabilities $p_{i,t}$ in turn necessitates more complicated consistency conditions (e.g. see Condition (18) compared to Condition (3)), and a more intricate analysis. The key in the analysis is to obtain the good concentration properties of the different estimators, which we accomplish by exhibiting martingale sequences and resorting to Bernstein's inequality for martingales (Theorem 14).

Recall that the crucial parameter for the stochastic model is the *minimal gap* $\Delta = \min_{i:\,\Delta_i>0} \Delta_i$, where $\Delta_i = (\max_{1\leq j\leq K} \mu_j) - \mu_i$ is the *gap* of arm i. Our main result is formulated as follows:

Theorem 15 *SAO with $\beta = n^4$ satisfies in the stochastic model:*

$$\mathbb{E}\overline{R}_n \leq O\left(\frac{K\log(K)\log^2(n)}{\Delta}\right),$$

and in the adversarial model:

$$\mathbb{E}R_n \leq O\left(\log(K)\log^{3/2}(n)\sqrt{nK}\right).$$

More precisely, for any $\delta \in (0,1)$, with probability at least $1-\delta$, SAO with $\beta = 10Kn^3\delta^{-1}$ satisfies in the stochastic model:

$$\overline{R}_n \leq \frac{260K(1+\log K)\log^2(\beta)}{\Delta},$$

and in the adversarial model:

$$R_n \leq 60(1+\log K)(1+\log n)\sqrt{nK\log(\beta) + 5K^2\log^2(\beta)} + 200K^2\log^2(\beta).$$

We divide the proof into three parts. In Section B.1 we propose several concentration inequalities for the different quantities involved in the algorithm. Then we make a deterministic argument conditional on the event that all these concentration inequalities hold true. First, in Section B.2, we analyze stochastic rewards, and Section B.3 concerns the adversarial rewards.

Algorithm 1 The SAO strategy with parameter $\beta > 1$

$A \leftarrow \{1, \dots, K\}$ ▷ A is the set of active arms
for $i = 1, \dots, K$ **do** ▷ Initialization
  $\tau_i \leftarrow n$ ▷ τ_i is the time when arm i is deactivated
  $p_i \leftarrow 1/K$ ▷ p_i is the probability of selecting arm i
end for
for $t = 1, \dots, n$ **do** ▷ Main loop
  Play I_t at random from p ▷ Selection of the arm to play
  for $i = 1, \dots, K$ **do** ▷ Test of four properties for arm i
    if ▷ Test if arm i should be deactivated

$$i \in A, \text{ and } \max_{j \in A} \tilde{H}_{j,t} - \tilde{H}_{i,t} > 6\sqrt{\frac{4K\log(\beta)}{t} + 5\left(\frac{K\log(\beta)}{t}\right)^2} \tag{17}$$

      then $A \leftarrow A \setminus \{i\}$, $\tau_i \leftarrow t$ and $q_i \leftarrow p_i$ ▷ Deactivation of arm i
    end if ▷ q_i denotes the probability of arm i at the moment when it was de-activated
    if one of the three following properties is satisfied
      then Start Exp3.P with the parameters described in [Theorem 2.4, Bubeck (2010)]
      ▷ Test if stochastic model still valid for arm i
      ▷ First, test if the two estimates of $H_{i,t}$ are consistent; let $t_i^* = \min(\tau_i, t)$.

$$\left|\tilde{H}_{i,t} - \widehat{H}_{i,t}\right| > \sqrt{\frac{2\log(\beta)}{T_i(t)}} + \sqrt{4\left(\frac{Kt_i^*}{t^2} + \frac{t - t_i^*}{q_i \tau_i t}\right)\log(\beta) + 5\left(\frac{K\log(\beta)}{t_i^*}\right)^2}. \tag{18}$$

      ▷ Second, test if the estimated suboptimality of arm i did not increase too much

$$i \notin A, \text{ and } \max_{j \in A} \tilde{H}_{j,t} - \tilde{H}_{i,t} > 10\sqrt{\frac{4K\log(\beta)}{\tau_i - 1} + 5\left(\frac{K\log(\beta)}{\tau_i - 1}\right)^2}. \tag{19}$$

      ▷ Third, test if arm i still seems significantly suboptimal

$$i \notin A, \text{ and } \max_{j \in A} \tilde{H}_{j,t} - \tilde{H}_{i,t} \leq 2\sqrt{\frac{4K\log(\beta)}{\tau_i} + 5\left(\frac{K\log(\beta)}{\tau_i}\right)^2}. \tag{20}$$

    end if
  end for ▷ End of testing
  for $i = 1, \dots, K$ **do** ▷ Update of the probability of selecting arm i

$$p_i \leftarrow \frac{q_i\, \tau_i}{t+1}\, \mathbb{1}_{i \notin A} + \frac{1}{|A|}\left(1 - \sum_{j \notin A} \frac{q_j\, \tau_j}{t+1}\right) \mathbb{1}_{i \in A}. \tag{21}$$

  end for
end for

Let us discuss some notation. Recall that we denote by $p_{i,t}$ the probability that the algorithm selects arm i at time t; this probability is denoted by p_i in the description of the algorithm. As in Algorithm 1, q_i will denote the probability of arm i at the moment when this arm was deactivated. Let A_t denote the set of active arms at the end of time step t. We also introduce τ_0 as the last time step before we start Exp3.P, with a convention that $\tau_0 = n$ if we never start Exp3.P. Moreover note that with this notation, if $\tau_i < \tau_0$ then we have $q_i = p_{i,\tau_i}$. We generalize this notation and set $q_i := p_{i,\,\min(\tau_i,\tau_0)}$. For sake of notation, in the following τ_i denotes the minimum between the time when arm i is deactivated and the last time before we start Exp3.P, that is $\tau_i \leftarrow \min(\tau_i, \tau_0)$.

B.1. Concentration inequalities

First we derive a version of Bernstein's inequality for martingales that suits our needs.

Lemma 16 *Let $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ real random variables such that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}(X_t|\mathcal{F}_{t-1}) = 0$ and $|X_t| \leq b$ for some $b > 0$. Let $V_n = \sum_{t=1}^n \mathbb{E}(X_t^2|\mathcal{F}_{t-1})$ and $\delta > 0$. Then with probability at least $1-\delta$,*

$$\sum_{t=1}^n X_t \leq \sqrt{4V_n \log(n\delta^{-1}) + 5b^2 \log^2(n\delta^{-1})}.$$

Proof. The proof follows from Theorem 14 along with an union bound on the events $V_n \in [x, x+b]$, $x \in \{0, b^2, 2b^2, \ldots, (n-1)b^2\}$. It also uses $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$. □

Now let us use this martingale inequality to derive the concentration bound for (average) estimated cumulative rewards $\tilde{H}_{i,t}$. Recall that $\tilde{H}_{i,t}$ is an estimator of $H_{i,t}$, so we want to upper-bound the difference $|\tilde{H}_{i,t} - H_{i,t}|$, and in the stochastic model $\tilde{H}_{i,t}$ is an estimator of the true expected reward μ_i, so we want to upper-bound the difference $|\tilde{H}_{i,t} - \mu_i|$.

Lemma 17 *For any arm $i \in \{1, \ldots, K\}$ and any time $t \in \{1, \ldots, n\}$, in the stochastic model we have with probability at least $1-\delta$, if $t \leq \tau_0$,*

$$\left|\tilde{H}_{i,t} - \mu_i\right| \leq \sqrt{4\left(\frac{K\min(\tau_i,t)}{t^2} + \frac{\max(t-\tau_i,0)}{q_i\tau_i t}\right)\log(2t^2\delta^{-1}) + 5\left(\frac{K\log(2t^2\delta^{-1})}{\min(\tau_i,t)}\right)^2}.$$

Moreover in the adversarial model we have with probability at least $1-\delta$, if $t \leq \tau_0$,

$$\left|\tilde{H}_{i,t} - H_{i,t}\right| \leq \sqrt{4\left(\frac{K\min(\tau_i,t)}{t^2} + \frac{\max(t-\tau_i,0)}{q_i\tau_i t}\right)\log(2t^2\delta^{-1}) + 5\left(\frac{K\log(2t^2\delta^{-1})}{\min(\tau_i,t)}\right)^2}.$$

Proof. The proof of the two concentration inequalities is similar, so we restrict our attention to the adversarial model. Let $(\mathcal{F}_s)$ be the filtration associated to the historic of the strategy. We introduce the following sequence of independent random variables: for $1 \leq i \leq K$, $1 \leq s \leq n$ and $p \in [0,1]$, let $Z_s^i(p) \sim \texttt{Bernoulli}(p)$. Then for $t \leq \tau_0$ we have,

$$\tilde{G}_{i,t} = \sum_{s=1}^t g_{i,s}\left(\frac{Z_s^i(p_{i,s})}{p_{i,s}}\mathbb{1}_{s\leq\tau_i} + \frac{s}{q_i\tau_i}Z_s^i\left(\frac{q_i\tau_i}{s}\right)\mathbb{1}_{s>\tau_i}\right).$$

For $T \in \{1, \ldots, n\}$, let

$$X_s^i(T) = \left(\frac{Z_s^i(p_{i,s})}{p_{i,s}} - 1 \right) g_{i,s} \mathbb{1}_{s \leq \tau_i \leq T} + \left(\frac{s}{q_i T} Z_s^i \left(\frac{q_i T}{s} \right) - 1 \right) g_{i,s} \mathbb{1}_{s > T \geq \tau_i}.$$

We have, for $t \leq \tau_0$,

$$\tilde{G}_{i,t} - G_{i,t} = \sum_{s=1}^{t} X_s^i(\tau_i).$$

Now remark that $(X_s^i(T))_{1 \leq s \leq t}$ is a martingale difference sequences such that $|X_s^i(T)| \leq K \max\left(\frac{t}{T}, 1\right)$ (since $p_{i,s} \geq 1/K$ when $s \leq \tau_i$) and

$$\sum_{s=1}^{t} \mathbb{E}\left((X_s^i(T))^2 | \mathcal{F}_{s-1} \right) \leq \sum_{s=1}^{\min(\tau_i, t)} \frac{1}{p_{i,s}} + \frac{t \max(t-T, 0)}{q_i T}.$$

Thus, using Lemma 16, we obtain that with probability at least $1 - \delta$,

$$\sum_{s=1}^{t} X_s^i(T) \leq \sqrt{4 \left(\sum_{s=1}^{\min(\tau_i, t)} \frac{1}{p_{i,s}} + \frac{t \max(t-T, 0)}{q_i T} \right) \log(t\delta^{-1}) + 5K^2 \max\left(\left(\frac{t}{T} \right)^2, 1 \right) \log^2(t\delta^{-1})}.$$

Then, using an union bound over T, we obtain the claimed inequality by taking $T = \tau_i$ (with another union bound to get the two-sided inequality). □

Next, we analyze the (average) cumulative reward $\widehat{H}_{i,t}$ collected by the algorithm. Again, in the stochastic model $\widehat{H}_{i,t}$ can be used as an estimate of the true expected reward μ_i, and it is not hard to see that it is a reasonably sharp estimate.

Lemma 18 *For any arm $i \in \{1, \ldots, K\}$, in the stochastic model we have with probability at least $1 - \delta$, for any time $t \in \{1, \ldots, n\}$, if $T_i(t) \geq 1$,*

$$\left| \widehat{H}_{i,t} - \mu_i \right| \leq \sqrt{\frac{2 \log(2n\delta^{-1})}{T_i(t)}}.$$

Proof. This follows via an union bound over the value of $T_i(t)$ and a standard Hoeffding's inequality for independent random variables, see Theorem 13. □

Next we show that, essentially, $T_i(t) \leq \tilde{O}(q_i \tau_i + \sqrt{q_i \tau_i})$.

Lemma 19 *For any $i \in \{1, \ldots, K\}, t \in \{1, \ldots, n\}$, with probability at least $1 - \delta$, if $t \leq \tau_0$,*

$$T_i(t) \leq q_i \tau_i (1 + \log t) + \sqrt{4 q_i \tau_i (1 + \log t) \log(t\delta^{-1}) + 5 \log^2(t\delta^{-1})}.$$

Proof. Using the notation of the proof of Lemma 17, we have for $t \leq \tau_0$,

$$T_i(t) = \sum_{s=1}^{t} Z_s^i(p_{i,s}) \mathbb{1}_{s \leq \tau_i} + Z_s^i \left(\frac{q_i \tau_i}{s} \right) \mathbb{1}_{s > \tau_i}.$$

Let

$$X_s^i = (Z_s^i(p_{i,s}) - p_{i,s})\mathbb{1}_{s\leq\tau_i} + \left(Z_s^i\left(\frac{q_i\tau_i}{s}\right) - \frac{q_i\tau_i}{s}\right)\mathbb{1}_{s>\tau_i}.$$

Then (X_s^i) is a martingale difference sequence such that $|X_s^i| \leq 1$ and, since $p_{i,s}$ is increasing in s for $s \leq \tau_i$, it follows that

$$\sum_{s=1}^{t} \mathbb{E}((X_s^i)^2|\mathcal{F}_{s-1}) \leq q_i\tau_i + \sum_{s=\tau_i+1}^{t} \frac{q_i\tau_i}{s} \leq q_i\tau_i(1+\log t).$$

Thus using Lemma 16 we obtain that with probability at least $1-\delta$:

$$\sum_{s=1}^{t} X_s^i(T) \leq \sqrt{4q_i\tau_i(1+\log t)\log(t\delta^{-1}) + 5\log^2(t\delta^{-1})}.$$

It implies that

$$\sum_{s=1}^{t} Z_s^i(p_{i,s})\mathbb{1}_{s\leq\tau_i} + Z_s^i\left(\frac{q_i\tau_i}{s}\right)\mathbb{1}_{s>\tau_i} \leq q_i\tau_i(1+\log t) + \sqrt{4q_i\tau_i(1+\log t)\log(t\delta^{-1}) + 5\log^2(t\delta^{-1})},$$

which is the claimed inequality. □

The next lemma restates regret guarantee for `Exp3.P` in terms of our setting. Instead of using the original guarantee from Auer et al. (2002b), we take an improved bound from Bubeck (2010) (namely, Theorem 2.4 in Bubeck (2010)).

Lemma 20 *In the adversarial model, with probability at least* $1-\delta$*, we have*

$$\max_{i\in\{1,\ldots,K\}} \sum_{t=\tau_0+1}^{n} g_{i,t} - \sum_{t=\tau_0+1}^{n} g_{I_t,t} \leq 5.15\sqrt{(n-\tau_0)K\log(K\delta^{-1})}.$$

Let $\beta = 10Kn^3\delta^{-1}$. Putting together the results of Lemma 17, 18, 19 and 20, we obtain that with probability at least $1-\delta$, the following inequalities hold true for any arm $i \in \{1,\ldots,K\}$ and

any time $t \in \{1, \dots, \tau_0\}$:

In the stochastic model,

$$\left|\tilde{H}_{i,t} - \mu_i\right| \leq \sqrt{4\left(\frac{K\min(\tau_i, t)}{t^2} + \frac{\max(t-\tau_i, 0)}{q_i \tau_i t}\right)\log(\beta) + 5\left(\frac{K\log(\beta)}{\min(\tau_i, t)}\right)^2}. \tag{22}$$

In the adversarial model,

$$\left|\tilde{H}_{i,t} - H_{i,t}\right| \leq \sqrt{4\left(\frac{K\min(\tau_i, t)}{t^2} + \frac{\max(t-\tau_i, 0)}{q_i \tau_i t}\right)\log(\beta) + 5\left(\frac{K\log(\beta)}{\min(\tau_i, t)}\right)^2}. \tag{23}$$

In the stochastic model,

$$\left|\widehat{H}_{i,t} - \mu_i\right| \leq \sqrt{\frac{2\log(\beta)}{T_i(t)}}. \tag{24}$$

In both models,

$$T_i(t) \leq q_i\tau_i(1+\log t) + \sqrt{4q_i\tau_i(1+\log t)\log(\beta) + 5\log^2(\beta)}. \tag{25}$$

In the adversarial model,

$$\max_{i\in\{1,\dots,K\}} \sum_{t=\tau_0+1}^{n} g_{i,t} - \sum_{t=\tau_0+1}^{n} g_{I_t,t} \leq 5.15\sqrt{(n-\tau_0)K\log(\beta)}. \tag{26}$$

We will now make a deterministic reasoning on the event that the above inequalities are indeed true.

B.2. Analysis in the stochastic model

First note that by equations (22) and (24), test (18) is never satisfied.

Let $i^* \in \operatorname{argmax}_i \mu_i$. Remark that by equation (22), test (17) is never satisfied for i^*, since if $i, i^* \in A_t$ then

$$\tilde{H}_{i,t} - \tilde{H}_{i^*,t} \leq -\Delta_i + 2\sqrt{\frac{4K\log(\beta)}{t} + 5\left(\frac{K\log(\beta)}{t}\right)^2}.$$

Thus we have $i^* \in A_t, \forall t$. Moreover if $i \notin A_t$, then it means that $\tau_i \leq t$ and test (17) was satisfied at time step τ_i (and not satisfied at time $\tau_i - 1$). Thus, using (22), we see that if $i \notin A_t$ then it implies:

$$\Delta_i + 2\sqrt{\frac{4K\log(\beta)}{\tau_i} + 5\left(\frac{K\log(\beta)}{\tau_i}\right)^2} > 6\sqrt{\frac{4K\log(\beta)}{\tau_i} + 5\left(\frac{K\log(\beta)}{\tau_i}\right)^2},$$

and (since $i^* \in A_t$)

$$\Delta_i - 2\sqrt{\frac{4K\log(\beta)}{\tau_i - 1} + 5\left(\frac{K\log(\beta)}{\tau_i - 1}\right)^2} \leq 6\sqrt{\frac{4K\log(\beta)}{\tau_i - 1} + 5\left(\frac{K\log(\beta)}{\tau_i - 1}\right)^2}. \tag{27}$$

Thus test (19) is never satisfied since:

$$\max_{j\in A_t} \tilde{H}_{j,t} - \tilde{H}_{i,t} \leq \Delta_i + 2\sqrt{\frac{4K\log(\beta)}{\tau_i} + 5\left(\frac{K\log(\beta)}{\tau_i}\right)^2} \leq 10\sqrt{\frac{4K\log(\beta)}{\tau_i - 1} + 5\left(\frac{K\log(\beta)}{\tau_i - 1}\right)^2}.$$

Moreover (20) is also never satisfied, indeed since $i^* \in A_t$ we have:

$$\max_{j \in A_t} \tilde{H}_{j,t} - \tilde{H}_{i,t} \geq \Delta_i - 2\sqrt{\frac{4K\log(\beta)}{\tau_i} + 5\left(\frac{K\log(\beta)}{\tau_i}\right)^2} > 2\sqrt{\frac{4K\log(\beta)}{\tau_i} + 5\left(\frac{K\log(\beta)}{\tau_i}\right)^2}.$$

In conclusion we proved that Exp3 is never started in the stochastic model, that is $\tau_0 = n$. Thus, using (25), we obtain:

$$\begin{aligned} \overline{R}_n &= \sum_{i=1}^K \Delta_i T_i(n) \\ &\leq \sum_{i=1}^K \Delta_i \left(q_i \tau_i (1+\log n) + \sqrt{4 q_i \tau_i (1+\log n)\log(\beta) + 5\log^2(\beta)} \right). \end{aligned}$$

Now remark that for any arm i with $\Delta_i > 0$, one can see that (27) implies:

$$\tau_i \leq 259 \frac{K\log(\beta)}{\Delta_i^2} + 1 \leq 260 \frac{K\log(\beta)}{\Delta_i^2}.$$

Indeed if $\tau_i > 259\frac{K\log(\beta)}{\Delta_i^2} + 1$, then

$$8\sqrt{\frac{4K\log(\beta)}{\tau_i - 1} + 5\left(\frac{K\log(\beta)}{\tau_i - 1}\right)^2} < 8\sqrt{\frac{4\Delta_i^2}{259} + \frac{5\Delta_i^4}{259}} < \Delta_i,$$

which contradicts (27).

The proof is concluded with straightforward computations and by showing that

$$\sum_{i=1}^K q_i \leq 1 + \log K. \tag{28}$$

Denote by $\tau_{(1)} \leq \ldots \leq \tau_{(K)}$ the ordered random variables $\tau_1, \ldots, \tau_K$. Then we clearly have $q_{(i)} \leq \frac{1}{K-i+1}$, which proves (28).

B.3. Analysis in the adversarial model

Let $i^* \in \operatorname{argmax}_{1 \leq i \leq K} G_{i,\tau_0-1}$. First we show that $i^* \in A_{\tau_0-1}$. Let $I^* \in \operatorname{argmax}_{i \in A_{\tau_0-1}} G_{i,\tau_0-1}$ and $i \notin A_{\tau_0-1}$, then we have, by $\tau_i \leq \tau_0 - 1$, (23) and since (20) is not satisfied at time $\tau_0 - 1$:

$$\begin{aligned} &G_{I^*,\tau_0-1} - G_{i,\tau_0-1} \\ &= G_{I^*,\tau_0-1} - \tilde{G}_{I^*,\tau_0-1} + \tilde{G}_{I^*,\tau_0-1} - \tilde{G}_{i,\tau_0-1} + \tilde{G}_{i,\tau_0-1} - G_{i,\tau_0-1} \\ &> -\sqrt{4\left(\frac{K\tau_i}{(\tau_0-1)^2} + \frac{\tau_0 - 1 - \tau_i}{q_i\tau_i(\tau_0-1)}\right)\log(\beta) + 5\left(\frac{K\log(\beta)}{\tau_i}\right)^2} - \sqrt{\frac{4K\log(\beta)}{\tau_0-1} + 5\left(\frac{K\log(\beta)}{\tau_0-1}\right)^2} \\ &\quad + 2\sqrt{\frac{4K\log(\beta)}{\tau_i} + 5\left(\frac{K\log(\beta)}{\tau_i}\right)^2} \\ &\geq -\sqrt{4\left(\frac{K\tau_i}{(\tau_0-1)^2} + \frac{\tau_0 - 1 - \tau_i}{q_i\tau_i(\tau_0-1)}\right)\log(\beta) + 5\left(\frac{K\log(\beta)}{\tau_i}\right)^2} + \sqrt{\frac{4K\log(\beta)}{\tau_i} + 5\left(\frac{K\log(\beta)}{\tau_i}\right)^2}, \end{aligned}$$

where the last inequality follows from $q_i \geq 1/K$ and

$$\frac{\tau_i}{(\tau_0-1)^2} + \frac{\tau_0 - 1 - \tau_i}{\tau_i(\tau_0-1)} \leq \frac{1}{\tau_i}.$$

This proves $i^* \in A_{\tau_0-1}$. Thus we get, using the fact that (18) and (19) are not satisfied at time $\tau_0 - 1$, as well as (23), and the fact that (17) is not satisfied for active arms at time $\tau_0 - 1$,

$$\begin{aligned}
R_{\tau_0-1} &= G_{i^*,\tau_0-1} - \sum_{i=1}^K \widehat{G}_{i,\tau_0-1} \\
&= \sum_{i=1}^K T_i(\tau_0-1)\left(H_{i^*,\tau_0-1} - \widehat{H}_{i,\tau_0-1}\right) \\
&= \sum_{i=1}^K T_i(\tau_0-1)\left(H_{i^*,\tau_0-1} - \tilde{H}_{i^*,\tau_0-1} + \tilde{H}_{i^*,\tau_0-1} - \tilde{H}_{i,\tau_0-1} + \tilde{H}_{i,\tau_0-1} - \widehat{H}_{i,\tau_0-1}\right) \\
&\leq \sum_{i=1}^K T_i(\tau_0-1)\left(12\sqrt{\frac{4K\log(\beta)}{\tau_i - 1} + 5\left(\frac{K\log(\beta)}{\tau_i-1}\right)^2} + \sqrt{\frac{2\log(\beta)}{T_i(\tau_0-1)}}\right).
\end{aligned}$$

Then, using (25) and (26) we get, thanks to $\tau_i \geq 2$,

$$\begin{aligned}
R_n &\leq 1 + 6.6\sqrt{nK\log(\beta)} + 12\sum_{i=1}^K q_i(1+\log n)\sqrt{16K\tau_i\log(\beta) + 20(K\log(\beta))^2} \\
&+ 12\sum_{i=1}^K \sqrt{\left(4q_i\tau_i(1+\log n)\log(\beta) + 5\log^2(\beta)\right)\left(\frac{4K\log(\beta)}{\tau_i-1} + 5\left(\frac{K\log(\beta)}{\tau_i-1}\right)^2\right)} \\
&\leq 60(1+\log K)(1+\log n)\sqrt{nK\log(\beta) + K^2\log^2(\beta)} + 200K^2\log^2(\beta),
\end{aligned}$$

where the last inequality follows from (28) and straightforward computations.