One Practical Algorithm for Both Stochastic and Adversarial Bandits

Full Version Including Appendices

Yevgeny Seldin YEVGENY.SELDIN@GMAIL.COM
Queensland University of Technology, Brisbane, Australia

Aleksandrs Slivkins SLIVKINS@MICROSOFT.COM
Microsoft Research, New York NY, USA

Abstract

We present an algorithm for multiarmed bandits that achieves almost optimal performance in both stochastic and adversarial regimes without prior knowledge about the nature of the environment. Our algorithm is based on augmentation of the EXP3 algorithm with a new control lever in the form of exploration parameters that are tailored individually for each arm. The algorithm simultaneously applies the "old" control lever, the learning rate, to control the regret in the adversarial regime and the new control lever to detect and exploit gaps between the arm losses. This secures problem-dependent "logarithmic" regret when gaps are present without compromising on the worst-case performance guarantee in the adversarial regime. We show that the algorithm can exploit both the usual expected gaps between the arm losses in the stochastic regime and deterministic gaps between the arm losses in the adversarial regime. The algorithm retains "logarithmic" regret guarantee in the stochastic regime even when some observations are contaminated by an adversary, as long as on average the contamination does not reduce the gap by more than a half. Our results for the stochastic regime are supported by experimental validation.

1. Introduction

Stochastic multiarmed bandits (Thompson, 1933; Robbins, 1952; Lai & Robbins, 1985; Auer et al., 2002a) and adversarial multiarmed bandits (Auer et al., 1995; 2002b) have co-existed in parallel for almost two decades by now, in the sense that no algorithm for stochastic multiarmed bandits is applicable to adversarial multiarmed bandits and algorithms for adversarial bandits are unable to exploit the simpler regime of stochastic bandits. The recent attempt of Bubeck & Slivkins (2012) to bring them together did not make it in the full sense of unification, since the algorithm of Bubeck and Slivkins relies on the knowledge of time horizon and makes a one-time irreversible switch between stochastic and adversarial operation modes if the beginning of the game is estimated to exhibit adversarial behavior.

We present an algorithm that treats both stochastic and adversarial multiarmed bandit problems without distinguishing between them. Our algorithm "just runs", as most other bandit algorithms, without knowledge of time horizon and without making any hard statements about the nature of the environment. We show that if the environment happens to be adversarial the performance of the algorithm is just a factor of 2 worse than the performance of the `EXP3` algorithm (with the best constants, as described in Bubeck & Cesa-Bianchi (2012)) and if the environment happens to be stochastic the performance of our algorithm is comparable to the performance of `UCB1` of Auer et al. (2002a). Thus, we cover the full range and achieve almost optimal performance at the extreme points.

Furthermore, we show that the new algorithm can exploit both the usual expected gaps between the arm losses in the stochastic regime and deterministic gaps between the arm losses in the adversarial regime. We also show that the algorithm retains "logarithmic" regret guarantee in the stochastic regime even when some observations are adversarially contaminated, as long as on average the contamination does not reduce the gap by more than a half. To the best of our knowledge, no other algorithm has been yet shown to be able to exploit gaps in the adversarial or adversarially contaminated stochastic regimes. The contaminated stochastic regime is a very practical model, since in many real-life situations we are dealing with stochastic environments with occasional disturbances.

Since the introduction of Thompson's sampling (Thompson, 1933) (which was analyzed only after 80 years (Kaufmann et al., 2012; Agrawal & Goyal, 2013)) a variety of al-

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

gorithms were invented for the stochastic multiarmed bandit problem. The most powerful for today are KL-UCB (Cappé et al., 2013), EwS (Maillard, 2011), and the aforementioned Thompson's sampling. It is easy to show that any deterministic algorithm can potentially suffer linear regret in the adversarial regime (see the supplementary material for a proof). Although nothing is known about the performance of randomized algorithms for stochastic bandits in the adversarial regimes, empirically they are extremely sensitive to deviations from the stochastic assumption.

In the adversarial world the most powerful algorithm for today is INF (Audibert & Bubeck, 2009; Bubeck & Cesa-Bianchi, 2012). Nevertheless, the EXP3 algorithm of Auer et al. (2002b) still retains an important place, mainly due to its simplicity and wide applicability, which covers combinatorial bandits, partial monitoring games, and many other adversarial problems. Since any stochastic problem can be seen as an instance of an adversarial problem, both INF and EXP3 have the worst-case "root-t" regret guarantee in the stochastic regime, but it is not known whether they can do better. Empirically in the stochastic regime EXP3 is inferior to all other known algorithms for this setting, including the simplest UCB1 algorithm.

It is interesting to take a brief look into the development of EXP3. The algorithm was first suggested in Auer et al. (1995) and its parametrization and analysis were improved in Auer et al. (2002b). The EXP3 of Auer et. al. was designed for the multiarmed bandit game with rewards and its playing strategy is based on mixing Gibbs distribution (also known as "exponential weights") with a uniform exploration distribution in proportion to the learning rate. The uniform exploration leaves no hope for achieving "logarithmic" regret in the stochastic regime simultaneously with the "root-t" regret in the adversarial regime, since each arm is played at least $\Omega(\sqrt{t})$ times in t rounds of the game. By changing the learning rate Cesa-Bianchi & Fischer (1998) managed to derive a different parametrization of the algorithm that was shown to achieve "logarithmic" regret in the stochastic regime, but it had no regret guarantees in the adversarial regime. Stoltz (2005) has observed that in the game with losses the "root-t" regret guarantee in the adversarial regime can be achieved without mixing in the uniform distribution (and even lead to better constants).[1] However, mixing in any distribution that element-wise does not exceed the learning rate does not break the worst-case performance of the algorithm in the game with losses. We exploit this emerged freedom in order to derive a modification of the EXP3 algorithm that achieves almost optimal regret in both adversarial and stochastic regimes without prior knowledge about the nature of the environment.

2. Problem Setting

We study the multiarmed bandit (MAB) game with losses. In each round t of the game the algorithm chooses one action A_t among K possible actions, a.k.a. *arms*, and observes the corresponding loss $\ell_t^{A_t}$. The losses of other arms are not observed. There is a large number of loss generation models, four of which are considered below. In this work we restrict ourselves to loss sequences $\{\ell_t^a\}_{t,a}$ that are generated independently of the algorithm's actions. Under this assumption we can assume that the loss sequences are written down before the game starts (but not revealed to the algorithm). We also make a standard assumption that the losses are bounded in the $[0,1]$ interval.

The performance of the algorithm is quantified by *regret*, defined as the difference between the expected loss of the algorithm up to round t and the expected loss of the best arm up to round t:

$$R(t) = \sum_{s=1}^{t} \mathbb{E}\left[\ell_s^{A_s}\right] - \min_a \left\{ \mathbb{E}\left[\sum_{s=1}^{t} \ell_s^a\right] \right\}.$$

The expectation is taken over the possible randomness of the algorithm and loss generation model. The goal of the algorithm is to minimize the regret.

We consider two standard loss generation models, the *adversarial regime* and the *stochastic regime* and two intermediate regimes, the *contaminated stochastic regime* and the *adversarial regime with a gap*.

Adversarial regime. In this regime the loss sequences are generated by an unrestricted adversary (who is oblivious to the algorithm's actions). This is the most general setting and the other three regimes can be seen as special cases of the adversarial regime. An arm $a \in \arg\min_{a'} \left(\sum_{s=1}^t \ell_s^{a'}\right)$ is known as a *best arm in hindsight* for the first t rounds.

Stochastic regime. In this regime the losses ℓ_t^a are sampled independently from an unknown distribution that depends on a, but not on t. We use $\mu(a) = \mathbb{E}\left[\ell_t^a\right]$ to denote the expected loss of arm a. Arm a is called a *best arm* if $\mu(a) = \min_{a'}\{\mu(a')\}$ and *suboptimal* otherwise; let a^* denote some best arm. For each arm a, define the *gap* $\Delta(a) = \mu(a) - \mu(a^*)$. Let $\Delta = \min_{a:\Delta(a)>0}\{\Delta(a)\}$ denote the minimal gap.

Letting $N_t(a)$ be the number of times arm a was played up to (and including) round t, the regret can be rewritten as

$$R(t) = \sum_a \mathbb{E}\left[N_t(a)\right]\Delta(a). \tag{1}$$

Contaminated stochastic regime. In this regime the adversary picks some round-arm pairs (t,a) ("locations") before the game starts and assigns the loss values there in an

[1] Rewards can be transformed into losses by taking $\ell = 1 - r$.

arbitrary way. The remaining losses are generated according to the stochastic regime.

We call a contaminated stochastic regime *moderately contaminated after τ rounds* if for all $t \geq \tau$ the total number of contaminated locations of each suboptimal arm up to time t is at most $t\Delta(a)/4$ and the number of contaminated locations of each best arm is at most $t\Delta/4$. By this definition, for all $t \geq \tau$ on average (over stochasticity of the loss sequences) the adversary can reduce the gap of every arm by at most a half.

Adversarial regime with a gap. An adversarial regime is named by us *an adversarial regime with a gap* if there exists a round τ and an arm a^*_τ that persists to be the best arm in hindsight for all rounds $t \geq \tau$. We name such arm a *consistently best arm after round* τ. If no such arm exists then a^*_τ is undefined. Note that if a^*_τ is defined for some τ then $a^*_{\tau'}$ is defined for all $\tau' > \tau$.

We use $\lambda_t(a) = \sum_{s=1}^t \ell_s^a$ to denote the cumulative loss of arm a. Whenever a^*_τ is defined we define a *deterministic gap* of arm a on round τ as:

$$\Delta(\tau, a) = \min_{t \geq \tau} \left\{ \frac{1}{t} \left(\lambda_t(a) - \lambda_t(a^*_\tau) \right) \right\}.$$

If a^*_τ is undefined, $\Delta(\tau, a)$ is defined as zero.

Notation. We use $\mathbb{1}_{\{E\}}$ to denote the indicator function of event E and $\mathbb{1}_t^a = \mathbb{1}_{\{A_t = a\}}$ to denote the indicator function of the event that arm a was played on round t.

3. Main Results

Our main results include a new algorithm, which we name EXP3++, and its analysis in the four regimes defined in the previous section. The EXP3++ algorithm, provided in Algorithm 1 box, is a generalization of the EXP3 algorithm with losses.

Algorithm 1 Algorithm EXP3++.

```
Remark: See text for definition of $\eta_t$ and $\xi_t(a)$.
$\forall a$: $\tilde{L}_0(a) = 0$.
for $t = 1, 2, ...$ do
    $\beta_t = \frac{1}{2}\sqrt{\frac{\ln K}{tK}}$.
    $\forall a$: $\varepsilon_t(a) = \min\left\{\frac{1}{2K}, \beta_t, \xi_t(a)\right\}$.
    $\forall a$: $\rho_t(a) = e^{-\eta_t \tilde{L}_{t-1}(a)} \big/ \sum_{a'} e^{-\eta_t \tilde{L}_{t-1}(a')}$.
    $\forall a$: $\tilde{\rho}_t(a) = \left(1 - \sum_{a'} \varepsilon_t(a')\right) \rho_t(a) + \varepsilon_t(a)$.
    Draw action $A_t$ according to $\tilde{\rho}_t$ and play it.
    Observe and suffer the loss $\ell_t^{A_t}$.
    $\forall a$ : $\tilde{\ell}_t^a = \frac{\ell_t^{A_t}}{\tilde{\rho}_t(a)} \mathbb{1}_t^a$.
    $\forall a$ : $\tilde{L}_t(a) = \tilde{L}_{t-1}(a) + \tilde{\ell}_t^a$.
end for
```

The EXP3++ algorithm has two control levers: the *learning rate* η_t and the *exploration parameters* $\xi_t(a)$. The EXP3 with losses (as described in Bubeck & Cesa-Bianchi (2012)) is a special case of the EXP3++ with $\eta_t = 2\beta_t$ and $\xi_t(a) = 0$.

The crucial innovation in EXP3++ is the introduction of exploration parameters $\xi_t(a)$, which are tuned individually for each arm depending on the past observations. In the sequel we show that tuning only the learning rate η_t suffices to control the regret of EXP3++ in the adversarial regime, irrespective of the choice of the exploration parameters $\xi_t(a)$. Then we show that tuning only the exploration parameters $\xi_t(a)$ suffices to control the regret of EXP3++ in the stochastic regime irrespective of the choice of η_t, as long as $\eta_t \geq \beta_t$. Applying the two control levers simultaneously we obtain an algorithm that achieves the optimal "root-t" regret in the adversarial regime (up to logarithmic factors) and almost optimal "logarithmic" regret in the stochastic regime (though with a suboptimal power in the logarithm). Then show that the new control lever is even more powerful and allows to detect and exploit the gap in even more challenging situations, including moderately contaminated stochastic regime and adversarial regime with a gap.

Adversarial Regime

First, we show tuning η_t is sufficient to control the regret of EXP3++ in the adversarial regime.

Theorem 1. *For $\eta_t = \beta_t$ and any $\xi_t(a) \geq 0$ the regret of* EXP3++ *for any t satisfies:*

$$R(t) \leq 4\sqrt{Kt \ln K}.$$

Note that the regret bound in Theorem 1 is just a factor of 2 worse than the regret of EXP3 with losses (Bubeck & Cesa-Bianchi, 2012).

Stochastic Regime

Now we show that for any $\eta_t \geq \beta_t$ tuning the exploration parameters $\xi_t(a)$ suffices to control the regret of the algorithm in the stochastic regime. By choosing $\eta_t = \beta_t$ we obtain algorithms that have both the optimal "root-t" regret scaling in the adversarial regime *and* "logarithmic" regret scaling in the stochastic regime.

We consider a number of different ways of tuning the exploration parameters $\xi_t(a)$, which lead to different parametrizations of EXP3++. We start with an idealistic assumption that the gap is known, just to give an idea of what is the best result we can hope for.

Theorem 2. *Assume that the gaps $\Delta(a)$ are known. For any choice of $\eta_t \geq \beta_t$ and any $c \geq 18$, the regret of* EXP3++ *with $\xi_t(a) = \frac{c \ln(t\Delta(a)^2)}{t\Delta(a)^2}$ in the stochastic regime* *satisfies:*

$$R(t) \leq \sum_a O\left(\frac{\ln(t)^2}{\Delta(a)}\right) + \sum_a \tilde{O}\left(\frac{K}{\Delta(a)^3}\right).$$

The constants in this theorem are small and are provided explicitly in the analysis. We also show that c can be made almost as small as 2.

Next we show that using the empirical gap as an estimate of the true gap

$$\hat{\Delta}_t(a) = \min\left\{1, \frac{1}{t}\left(\tilde{L}_t(a) - \min_{a'}\left(\tilde{L}_t(a')\right)\right)\right\} \qquad (2)$$

we can also achieve polylogarithmic regret guarantee. We call this algorithm $\texttt{EXP3++}^{\texttt{AVG}}$.

Theorem 3. *Let $c \geq 18$ and $\eta_t \geq \beta_t$. Let t^* be the minimal integer that satisfies $t^* \geq \frac{4c^2 K \ln(t^*)^4}{\ln(K)}$ and let $t^*(a) = \max\left\{t^*, \left\lceil e^{1/\Delta(a)^2}\right\rceil\right\}$. The regret of* EXP3++ *with $\xi_t(a) = \frac{c(\ln t)^2}{t\hat{\Delta}_{t-1}(a)^2}$ (termed $\texttt{EXP3++}^{\texttt{AVG}}$) in the stochastic regime satisfies:*

$$R(t) \leq \sum_a O\left(\frac{\ln(t)^3}{\Delta(a)}\right) + \sum_a \Delta(a) t^*(a).$$

Although the additive constants $t^*(a)$ in this theorem are very large, in the experimental section we show that a minor modification of this algorithm performs comparably to UCB1 in the stochastic regime (and has the adversarial regret guarantee in addition).

In the following theorem we show that if we assume a known time horizon T, then we can eliminate the additive term $e^{1/\Delta(a)^2}$ in the regret bound. The algorithm in Theorem 4 replaces the empirical gap estimate (2) in the definition of $\xi_t(a)$ with a lower confidence bound on the gap and slightly adjusts other terms. We name this algorithm $\texttt{EXP3++}^{\texttt{LCBT}}$.

Theorem 4. *Consider the stochastic regime with a known time horizon T. The $\texttt{EXP3++}^{\texttt{LCBT}}$ algorithm with any $\eta_t \geq \beta_t$ and appropriately defined $\xi_t(a)$ achieves regret $R(T) \leq \frac{O(\log^3 T)}{\Delta^3}$.*

The precise definition of $\texttt{EXP3++}^{\texttt{LCBT}}$ and the proof of Theorem 4 are provided in the supplementary material. It seems that simultaneous elimination of the assumption on the known time horizon and the exponentially large additive term is a very challenging problem and we defer it for future work.

Contaminated Stochastic Regime

Next we show that $\texttt{EXP3++}^{\texttt{AVG}}$ can sustain moderate contamination in the stochastic regime without a significant deterioration in performance.

Theorem 5. *Under the parametrization given in Theorem 3, for $t^*(a) = \max\left\{t^*, \left\lceil e^{4/\Delta(a)^2}\right\rceil\right\}$, where t^* is defined as before, the regret of $\texttt{EXP3++}^{\texttt{AVG}}$ in the stochastic regime that is moderately contaminated after τ rounds satisfies:*

$$R(t) \leq \sum_a O\left(\frac{\ln(t)^3}{\Delta(a)}\right) + \sum_a \max\left\{t^*(a), \tau\right\}.$$

The price that is paid for moderate contamination after τ rounds is the scaling of $\Delta(a)$ by a factor of $1/2$ and the additive factor of τ. (The scaling of Δ affects the definition of t^* and the constant in $O\left(\frac{\ln(t)^3}{\Delta(a)}\right)$.) As before, the regret guarantee of Theorem 5 comes *in addition* to the guarantee of Theorem 1.

Adversarial Regime with a Gap

Finally, we show that $\texttt{EXP3++}^{\texttt{AVG}}$ can also take advantage of deterministic gap in the adversarial regime.

Theorem 6. *Under the parametrization given in Theorem 3, the regret of $\texttt{EXP3++}^{\texttt{AVG}}$ in the adversarial regime satisfies:*

$$R(t) \leq \sum_a \min_\tau \left\{\max\left\{t^*, \tau, e^{1/(\Delta(\tau,a))^2}\right\} + O\left(\frac{\ln(t)^3}{\Delta(\tau,a)}\right)\right\}.$$

We remind the reader that in the absence of consistently best arm $\Delta(\tau, a)$ is defined as zero and the regret bound is vacuous (but the regret bound of Theorem 1 still holds). We also note that $\Delta(\tau, a)$ is a non-decreasing function of τ. Therefore, there is a trade-off: increasing τ increases $\Delta(\tau, a)$, but loses the regret guarantee on the rounds before τ (for simplicity, we assume that we have no guarantees before τ). Theorem 6 allows to pick τ that minimizes this trade-off. An important implication of the theorem is that if the deterministic gap is growing with time the regret guarantee improves too.

4. Proofs

We prove the theorems from the previous section in the order they were presented.

The Adversarial Regime

The proof of Theorem 1 relies on the following lemma, which is an intermediate step in the analysis of EXP3 by Bubeck (2010) (see also Bubeck & Cesa-Bianchi (2012)).

Lemma 7. *For any K sequences of non-negative numbers $X_1^a, X_2^a, \ldots$ indexed by $a \in \{1, \ldots, K\}$ and any non-increasing positive sequence $\eta_1, \eta_2, \ldots$, for $\rho_t(a) =$*

$\frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} X_s^a\right)}{\sum_{h'} \exp\left(-\eta_t \sum_{s=1}^{t-1} X_s^a\right)}$ *(assuming for $t = 1$ the sum in the exponent is zero) we have:*

$$\sum_{t=1}^{T} \sum_a \rho_t(a) X_t^a - \min_a \left(\sum_{t=1}^{T} X_t^a\right) \leq \frac{1}{2} \sum_{t=1}^{T} \eta_t \sum_a \rho_t(a) \left(X_t^a\right)^2 + \frac{\ln K}{\eta_T}. \qquad (3)$$

More precisely, we are using the following corollary, which follows by allowing X_t^a-s to be random variables and taking expectations of the two sides of (3) and using the fact that $\mathbb{E}\left[\min[\cdot]\right] \leq \min\left[\mathbb{E}\left[\cdot\right]\right]$. We decompose expectations of incremental sums into incremental sums of conditional expectations and use $\mathbb{E}_t\left[\cdot\right]$ to denote expectations conditioned on realization of all random variables up to round t.

Corollary 8. *Let $X_1^a, X_2^a, \dots$ for $a \in \{1, \dots, K\}$ be nonnegative random variables and let η_t and ρ_t as defined in Lemma 7. Then:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t\left[\sum_a \rho_t(a) X_t^a\right]\right] - \min_a \left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t\left[X_t^a\right]\right]\right) \leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{\eta_t}{2} \mathbb{E}_t\left[\sum_a \rho_t(a) \left(X_t^a\right)^2\right]\right] + \frac{\ln K}{\eta_T}.$$

Proof of Theorem 1. We associate X_t^a in (3) with $\tilde{\ell}_t^a$ in the EXP3++ algorithm. We have $\mathbb{E}_t\left[\tilde{\ell}_t^a\right] = \ell_t^a$ and since

$$\rho_t(a) = \frac{1}{1 - \sum_{a'} \varepsilon_t(a')} \left(\tilde{\rho}_t(a) - \varepsilon_t(a)\right) \geq \tilde{\rho}_t(a) - \varepsilon_t(a)$$

and $\ell_t^a \in [0, 1]$ we also have:

$$\mathbb{E}_t\left[\sum_a \rho_t(a) \tilde{\ell}_t^a\right] \geq \mathbb{E}_t\left[\sum_a \left(\tilde{\rho}_t(a) - \varepsilon_t(a)\right) \ell_t^a\right] \geq \mathbb{E}_t\left[\ell_t^{A_t}\right] - \sum_a \varepsilon_t(a).$$

As well, we have:

$$\mathbb{E}_t\left[\sum_a \rho_t(a) \left(\tilde{\ell}_t^a\right)^2\right] = \mathbb{E}_t\left[\sum_a \rho_t(a) \left(\frac{\ell_t^{A_t}}{\tilde{\rho}_t(a)} \mathbb{1}_t^a\right)^2\right] \leq \mathbb{E}_t\left[\sum_a \frac{\rho_t(a)}{\tilde{\rho}_t(a)^2} \mathbb{1}_t^a\right] = \sum_a \frac{\rho_t(a)}{\tilde{\rho}_t(a)} = \sum_a \frac{\rho_t(a)}{\left(1 - \sum_{a'} \varepsilon_t(a')\right) \rho_t(a) + \varepsilon_t(a)} \leq 2K,$$

where the last inequality follows by the fact that $\left(1 - \sum_{a'} \varepsilon_t(a')\right) \geq \frac{1}{2}$ by the definition of $\varepsilon_t(a)$. Substitution of the above calculations into Corollary 8 yields:

$$R(t) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{A_t}\right] - \min_a \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^a\right] \leq K \sum_{t=1}^{T} \eta_t + \frac{\ln K}{\eta_T} + \sum_a \varepsilon_t(a) \leq 2K \sum_{t=1}^{T} \eta_t + \frac{\ln K}{\eta_T}.$$

The result of the theorem follows by the choice of η_t. □

The Stochastic Regime

Our proofs are based on the following form of Bernstein's inequality, which is a minor improvement over Cesa-Bianchi & Lugosi (2006, Lemma A.8) based on the ideas from Boucheron et al. (2013, Theorem 2.10).

Theorem 9 (Bernstein's inequality for martingales)**.** *Let $X_1, \dots, X_n$ be a martingale difference sequence with respect to filtration $\mathcal{F} = (\mathcal{F}_i)_{1 \leq i \leq n}$ and let $S_i = \sum_{j=1}^{i} X_j$ be the associated martingale. Assume that there exist positive numbers ν and c, such that $X_j \leq c$ for all j with probability 1 and $\sum_{i=1}^{n} \mathbb{E}\left[(X_i)^2 \middle| \mathcal{F}_{i-1}\right] \leq \nu$ with probability 1. Then for all $b > 0$:*

$$\mathbb{P}\left[S_n > \sqrt{2\nu b} + \frac{cb}{3}\right] \leq e^{-b}.$$

We are also using the following technical lemma, which is proved in the supplementary material.

Lemma 10. *For any $c > 0$: $\sum_{t=0}^{\infty} e^{-c\sqrt{t}} = O\left(\frac{2}{c^2}\right)$.*

The proof of Theorems 2 and 3 is based on the following lemma.

Lemma 11. *Let $\{\underline{\varepsilon}_t(a)\}_{t=1}^{\infty}$ be non-increasing deterministic sequences, such that $\underline{\varepsilon}_t(a) \leq \varepsilon_t(a)$ with probability 1 and $\underline{\varepsilon}_t(a) \leq \underline{\varepsilon}_t(a^*)$ for all t and a. Define $\nu_t(a) = \sum_{s=1}^{t} \frac{1}{\underline{\varepsilon}_s(a)}$ and define the event E_t^a*

$$t\Delta(a) - \left(\tilde{L}_t(a) - \tilde{L}_t(a^*)\right) \leq \sqrt{2\left(\nu_t(a) + \nu_t(a^*)\right) b_t} + \frac{1.25 b_t}{3\underline{\varepsilon}_t(a^*)}. \qquad (E_t^a)$$

Then for any positive sequence $b_1, b_2, \dots$ and any $t^* \geq 2$ the number of times arm a is played by EXP3++ *up to round t is bounded as:*

$$\mathbb{E}\left[N_t(a)\right] \leq (t^* - 1) + \sum_{s=t^*}^{t} e^{-b_s} + \sum_{s=t^*}^{t} \varepsilon_s(a) \mathbb{1}_{\{E_t^a\}} + \sum_{s=t^*}^{t} e^{-\eta_s g_{s-1}(a)},$$

where

$$g_t(a) = t\Delta(a) - \sqrt{2tb_t\left(\frac{1}{\underline{\varepsilon}_t(a)} + \frac{1}{\underline{\varepsilon}_t(a^*)}\right)} - \frac{1.25 b_t}{3\underline{\varepsilon}_t(a^*)}.$$

Proof. Note that elements of the martingale difference sequence $\left\{\Delta(a) - \left(\tilde{\ell}_t^a - \tilde{\ell}_t^{a^*}\right)\right\}_{t=1}^{\infty}$ are upper bounded by $\frac{1}{\underline{\varepsilon}_t(a^*)} + 1$. Since $\underline{\varepsilon}_t(a^*) \leq \varepsilon_t(a^*) \leq 1/(2K) \leq 1/4$ we can simplify the upper bound by using $\frac{1}{\underline{\varepsilon}_t(a^*)} + 1 \leq \frac{1.25}{\underline{\varepsilon}_t(a^*)}$. Further note that

$$
\begin{aligned}
&\sum_{s=1}^{t} \mathbb{E}_s\left[\left(\Delta(a) - \left(\tilde{\ell}_s^a - \tilde{\ell}_s^{a^*}\right)\right)^2\right] \\
&\quad \leq \sum_{s=1}^{t} \mathbb{E}_s\left[\left(\tilde{\ell}_s^a - \tilde{\ell}_s^{a^*}\right)^2\right] \\
&\quad = \sum_{s=1}^{t} \left(\mathbb{E}_s\left[\left(\tilde{\ell}_s^a\right)^2\right] + \mathbb{E}_s\left[\left(\tilde{\ell}_s^{a^*}\right)^2\right]\right) \\
&\quad \leq \sum_{s=1}^{t} \left(\frac{1}{\tilde{p}_s(a)} + \frac{1}{\tilde{p}_s(a^*)}\right) \leq \sum_{s=1}^{t} \left(\frac{1}{\varepsilon_s(a)} + \frac{1}{\varepsilon_s(a^*)}\right) \\
&\quad \leq \sum_{s=1}^{t} \left(\frac{1}{\underline{\varepsilon}_s(a)} + \frac{1}{\underline{\varepsilon}_s(a^*)}\right) = \nu_t(a) + \nu_t(a^*)
\end{aligned}
$$

with probability 1. Let $\overline{E}$ denote the complement of event E. Then by Bernstein's inequality $\mathbb{P}\left[\overline{E_t^a}\right] \leq b_t$. The number of times arm a is played up to round t is bounded as:

$$
\begin{aligned}
\mathbb{E}\left[N_t(a)\right] &= \sum_{s=1}^{t} \mathbb{P}\left[A_s = a\right] \\
&= \sum_{s=1}^{t} \mathbb{P}\left[A_s = a \middle| E_{s-1}^a\right] \mathbb{P}\left[E_{s-1}^a\right] \\
&\qquad + \mathbb{P}\left[A_s = a \middle| \overline{E_{s-1}^a}\right] \mathbb{P}\left[\overline{E_{s-1}^a}\right] \\
&\leq \sum_{s=1}^{t} \mathbb{P}\left[A_s = a \middle| E_{s-1}^a\right] \mathbb{1}_{\left\{E_{s-1}^a\right\}} + \mathbb{P}\left[\overline{E_{s-1}^a}\right] \\
&\leq \sum_{s=1}^{t} \mathbb{P}\left[A_s = a \middle| E_{s-1}^a\right] \mathbb{1}_{\left\{E_{s-1}^a\right\}} + e^{-b_{s-1}}.
\end{aligned}
$$

For the terms of the sum above we have:

$$
\begin{aligned}
\mathbb{P}\left[A_t = a \middle| E_{t-1}^a\right] \mathbb{1}_{\left\{E_{s-1}^a\right\}} &= \tilde{\rho}_t(a) \mathbb{1}_{\left\{E_{s-1}^a\right\}} \\
&\leq \left(\rho_t(a) + \varepsilon_t(a)\right) \mathbb{1}_{\left\{E_{s-1}^a\right\}} \\
&= \left(\varepsilon_t(a) + \frac{e^{-\eta_t \tilde{L}_{t-1}}(a)}{\sum_{a'} e^{-\eta_t \tilde{L}_{t-1}(a')}}\right) \mathbb{1}_{\left\{E_{s-1}^a\right\}} \\
&\leq \left(\varepsilon_t(a) + e^{-\eta_t\left(\tilde{L}_{t-1}(a) - \tilde{L}_{t-1}(a^*)\right)}\right) \mathbb{1}_{\left\{E_{s-1}^a\right\}} \\
&\leq \varepsilon_t(a) \mathbb{1}_{\left\{E_{s-1}^a\right\}} + e^{-\eta_t g_{t-1}(a)},
\end{aligned}
$$

Where in the last inequality we used the facts that event E_t^a holds and that since $\underline{\varepsilon}_t(a)$ is a non-increasing sequence $\nu_t(a) \leq \frac{t}{\underline{\varepsilon}_t(a)}$. Substitution of this result back into the computation of $\mathbb{E}\left[N_t(a)\right]$ completes the proof. □

Proof of Theorem 2. The proof is based on Lemma 11. Let $b_t = \ln(t\Delta(a)^2)$ and $\underline{\varepsilon}_t(a) = \varepsilon_t(a)$. For any $c \geq 18$ and any $t \geq t^*$, where t^* is the minimal integer for which $t^* \geq \frac{4c^2 K \ln(t^*\Delta(a)^2)^2}{\Delta(a)^4 \ln(K)}$, we have:

$$
\begin{aligned}
g_t(a) &= t\Delta(a) - \sqrt{2tb_t\left(\frac{1}{\varepsilon_t(a)} + \frac{1}{\varepsilon_t(a^*)}\right)} - \frac{1.25b_t}{3\varepsilon_t(a^*)} \\
&\geq t\Delta(a) - 2\sqrt{\frac{tb_t}{\varepsilon_t(a)}} - \frac{1.25b_t}{3\varepsilon_t(a)} \\
&= t\Delta(a)\left(1 - \frac{2}{\sqrt{c}} - \frac{1.25}{3c}\right) \geq \frac{1}{2}t\Delta(a).
\end{aligned}
$$

(The choice of t^* ensures that for all suboptimal actions a we have $\varepsilon_t(a) = \xi_t(a)$, which slightly simplifies the calculations. Also note that since $\varepsilon_t(a^*) = \min\left\{\frac{1}{2K}, \beta_t\right\}$, asymptotically $1/\varepsilon_t(a)$ term in $g_t(a)$ dominates $1/\varepsilon_t(a^*)$ term and with a bit more careful bounding c can be made almost as small as 2.) By substitution of the lower bound on $g_t(a)$ into Lemma 11 we have:

$$
\begin{aligned}
\mathbb{E}\left[N_t(a)\right] &\leq t^* + \frac{\ln(t)}{\Delta(a)^2} + \frac{c\ln(t)^2}{\Delta(a)^2} \\
&\qquad + \sum_{s=1}^{t}\left(e^{-\frac{\Delta(a)}{4}\sqrt{\frac{(s-1)\ln(K)}{K}}}\right) \\
&\leq \frac{c\ln(t)^2}{\Delta(a)^2} + \frac{\ln(t)}{\Delta(a)^2} + O\left(\frac{K}{\Delta(a)^2}\right) + t^*,
\end{aligned}
$$

where we used Lemma 10 to bound the sum of the exponents. Note that t^* is of order $\tilde{O}\left(\frac{K}{\Delta(a)^4}\right)$. □

Proof of Theorem 3. Note that since by our definition $\hat{\Delta}_t(a) \leq 1$ the sequence $\underline{\varepsilon}_t(a) = \underline{\varepsilon}_t = \min\left\{\frac{1}{2K}, \beta_t, \frac{c\ln(t)^2}{t}\right\}$ satisfies the condition of Lemma 11. Also note that for t large enough, so that $t \geq \frac{4c^2K\ln(t)^4}{\ln K}$, we have $\underline{\varepsilon}_t = \frac{c\ln(t)^2}{t}$. Let $b_t = \ln(t)$ and let t^* be large enough, so that for all $t \geq t^*$ we have $t \geq \frac{4c^2K\ln(t)^4}{\ln K}$ and $t \geq e^{\frac{1}{\Delta(a)^2}}$. We are going to bound the three terms in the bound on $\mathbb{E}\left[N_t(a)\right]$ in Lemma 11. Bounding $\sum_{s=t^*}^{t} e^{-b_s}$ is easy. For bounding $\sum_{s=t^*}^{t} \varepsilon_s(a)\mathbb{1}_{\left\{E_{s-1}^a\right\}}$ we note that when E_t^a holds and $c \geq 18$ we have:

$$
\begin{aligned}
\hat{\Delta}_t(a) &\geq \frac{1}{t}\left(\tilde{L}_t(a) - \min_{a'}\tilde{L}_t(a')\right) \geq \frac{1}{t}\left(\tilde{L}_t(a) - \tilde{L}_t(a^*)\right) \\
&\geq \frac{1}{t}g_t(a) = \frac{1}{t}\left(t\Delta(a) - 2\sqrt{\frac{tb_t}{\underline{\varepsilon}_t}} - \frac{1.25b_t}{3\underline{\varepsilon}_t}\right) \qquad (4) \\
&= \frac{1}{t}\left(t\Delta(a) - \frac{2t}{\sqrt{c\ln(t)}} - \frac{1.25t}{3c\ln t}\right) \\
&\geq \Delta(a)\left(1 - \frac{2}{\sqrt{c}} - \frac{1.25}{3c}\right) \geq \tfrac{1}{2}\Delta(a),
\end{aligned}
$$

where in (4) we used the fact that E_t^a holds and in the last line we used the fact that for $t \geq t^*$ we have $\sqrt{\ln t} \geq 1/\Delta(a)$. Thus

$$\varepsilon_t \mathbb{1}_{\{E_{s-1}^a\}} \leq \frac{c(\ln t)^2}{t\hat{\Delta}_t(a)^2} \leq \frac{4c^2(\ln t)^2}{t\Delta(a)^2}$$

and $\sum_{s=t^*}^{t} \varepsilon_s \mathbb{1}_{\{E_{s-1}^a\}} = O\left(\frac{\ln(t)^3}{\Delta(a)^2}\right)$. Finally, for the last term in Lemma 11 we have already shown as an intermediate step in the calculation of the bound on $\hat{\Delta}_t(a)$ that for $t \geq t^*$ we have $g_t(a) \geq \frac{1}{2}\Delta(a)$. Therefore, the last term is of order $O\left(\frac{K}{\Delta(a)^2}\right)$. By taking all these calculations together we obtain the result of the theorem. Note that the result holds for any $\eta_t \geq \beta_t$. □

The Contaminated Stochastic Regime

Proof of Theorem 5. The key element of the previous proof was a high-probability lower bound on $\tilde{L}_t(a) - \tilde{L}_t(a^*)$. We show that we can obtain a similar lower bound in the contaminated setting too. Let $\mathbb{1}_{t,a}^{\times}$ denote the indicator function of contamination in "location" (t, a) ($\mathbb{1}_{t,a}^{\times}$ takes value 1 if contamination occurred and 0 otherwise). Let $m_t^a = \mathbb{1}_{t,a}^{\times} \ell_t^a + \left(1 - \mathbb{1}_{t,a}^{\times}\right) \mu(a)$, in other words, if either a was contaminated on round t then m_t^a is the adversarially assigned value of the loss of arm a on round t and otherwise it is the expected loss. Let $M_t(a) = \sum_{s=1}^{t} m_s^a$ then $(M_t(a) - M_t(a^*)) - \left(\tilde{L}_t(a) - \tilde{L}_t(a^*)\right)$ is a martingale. By definition of moderately contaminated after τ rounds process, for $t \geq \tau$ and any suboptimal action a the total number of rounds up to t where either a itself or a^* were contaminated is at most $t\Delta(a)/2$. Therefore, $M_t(a) - M_t(a^*) \geq (t - t\Delta(a)/2)\,\Delta(a) - t\Delta/2 \geq t\Delta(a)/2$. Define event B_t^a:

$$\frac{t\Delta(a)}{2} - \left(\tilde{L}_t(a) - \tilde{L}_t(a^*)\right) \leq 2\sqrt{\nu_t b_t} + \frac{1.25 b_t}{3\underline{\varepsilon}_t}, \quad (B_t^a)$$

where $\underline{\varepsilon}_t$ is defined in the proof of Theorem 3 and $\nu_t = \sum_{s=1}^{t} \frac{1}{\underline{\varepsilon}_s}$. Then by Bernstein's inequality $\mathbb{P}\left[\overline{B_t^a}\right] \leq b_t$. The remainder of the proof is identical to the proof of Theorem 3 with $\Delta(a)$ replaced by $\Delta(a)/2$. □

The Adversarial Regime with a Gap

The proof of Theorem 6 is based on the following lemma, which is an analogue of Theorems 3 and 5.

Lemma 12. *Under the parametrization given in Theorem 3, the number of times a suboptimal arm a is played by* EXP3++$^{\texttt{AVG}}$ *in an adversarial regime with a gap satisfies:*

$$\mathbb{E}\left[N_t(a)\right] \leq \max\left\{t^*, \tau, e^{1/(\Delta(\tau,a))^2}\right\} + O\left(\frac{\ln(t)^3}{\Delta(\tau,a)^2}\right).$$

Proof. Again, the only modification we need is a high-probability lower bound on $\tilde{L}_t(a) - \tilde{L}_t(a_\tau^*)$. We note that $(\lambda_t(a) - \lambda_t(a_\tau^*)) - \left(\tilde{L}_t(a) - \tilde{L}_t(a_\tau^*)\right)$ is a martingale and that by definition for $t \geq \tau$ we have $(\lambda_t(a) - \lambda_t(a_\tau^*)) \geq t\Delta(\tau, a)$. Define the events W_t^a:

$$t\Delta(\tau,a) - \left(\tilde{L}_t(a) - \tilde{L}_t(a_\tau^*)\right) \leq 2\sqrt{\nu_t b_t} + \frac{1.25 b_t}{3\underline{\varepsilon}_t}, \quad (W_t^a)$$

where $\underline{\varepsilon}_t$ and ν_t are as in the proof of Theorem 5. By Bernstein's inequality $\mathbb{P}\left[\overline{W_t^a}\right] \leq b_t$. The remainder of the proof is identical to the proof of Theorem 3. □

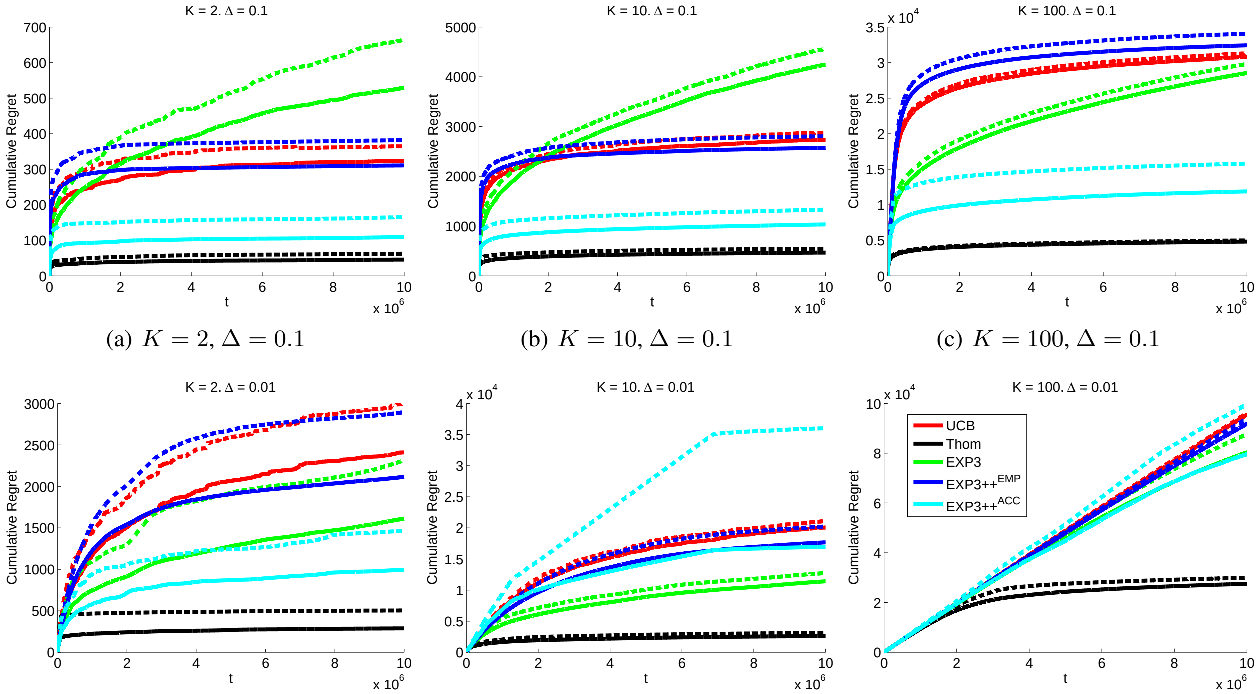
Proof of Theorem 6. Note that by definition $\Delta(\tau, a)$ is a non-decreasing sequence of τ. Since Lemma 12 is a deterministic result it holds for all τ simultaneously and we are free to choose the one that minimizes the bound. □

5. Empirical Evaluation: Stochastic Regime

We consider the stochastic multiarmed bandit problem with Bernoulli rewards. For all the suboptimal arms the rewards are Bernoulli with bias 0.5 and for the single best arm the reward is Bernoulli with bias $0.5 + \Delta$. We run the experiments with $K = 2$, $K = 10$, and $K = 100$, and $\Delta = 0.1$ and $\Delta = 0.01$ (in total, six combinations of K and Δ). We run each game for 10^7 rounds and make ten repetitions of each experiment. The solid lines in the graphs in Figure 1 represent the mean performance over the experiments and the dashed lines represent the mean plus one standard deviation (std) over the ten repetitions of the corresponding experiment.

In the experiments EXP3++ is parametrized by $\xi_t(a) = \frac{\ln(t\hat{\Delta}_t(a)^2)}{32t\hat{\Delta}_t(a)^2}$, where $\hat{\Delta}_t(a)$ is the empirical estimate of $\Delta(a)$ defined in (2). In order to demonstrate that in the stochastic regime the exploration parameters are in full control of the performance we run the EXP3++ algorithm with two different learning rates. EXP3++$^{\texttt{EMP}}$ corresponds to $\eta_t = \beta_t$ and EXP3++$^{\texttt{ACC}}$ corresponds to $\eta_t = 1$. Note that only the EXP3++$^{\texttt{EMP}}$ has a performance guarantee in the adversarial regime.

We compare EXP3++ algorithm with the EXP3 algorithm (as described in Bubeck & Cesa-Bianchi (2012)), the UCB1 algorithm of Auer et al. (2002a), and Thompson's sampling. Since it was demonstrated empirically in Seldin et al. (2013) that in the above experiments the performance of Thompson sampling is comparable or superior to the performance of EwS and KL-UCB, the latter two algorithms are excluded from the comparison. For the EXP3++ and the EXP3 algorithms we transform the rewards into losses via $\ell_t^a = 1 - r_t^a$ transformation, other algorithms operate directly on the rewards.



(a) $K = 2, \Delta = 0.1$ (b) $K = 10, \Delta = 0.1$ (c) $K = 100, \Delta = 0.1$

(d) $K = 2, \Delta = 0.01$ (e) $K = 10, \Delta = 0.01$ (f) $K = 100, \Delta = 0.01$

Figure 1. **Comparison of UCB1, Thompson sampling ("Thom"), EXP3, and EXP3++ algorithms in the stochastic regime.** The legend in figure (f) corresponds to all the figures. EXP3++$^{\text{EMP}}$ is the Empirical EXP3++ algorithm and EXP3++$^{\text{ACC}}$ is an Accelerated Empirical EXP3++, where we take $\eta_t = 1$. Solid lines correspond to means over 10 repetitions of the corresponding experiments and dashed lines correspond to the means plus one standard deviation.

The results are presented in Figure 1. We see that in all the experiments the performance of EXP3++$^{\text{EMP}}$ is almost identical to the performance of UCB1. However, unlike UCB1 and Thompson's sampling, EXP3++$^{\text{EMP}}$ is secured against the possibility that the game is controlled by an adversary. In the supplementary material we show that any deterministic algorithm is vulnerable against an adversary.

The EXP3++$^{\text{ACC}}$ algorithm can be seen as a teaser for future work. It performs better than EXP3++$^{\text{EMP}}$, but it does not have the adversarial regime performance guarantee. However, we do not exclude the possibility that by some more sophisticated simultaneous control of η_t and $\varepsilon_t(a)$-s it may be possible to design an algorithm that will have both better performance in the stochastic regime and a regret guarantee in the adversarial regime. An example of such sophisticated control of the learning rate in the full information games can be found in de Rooij et al. (2014).

6. Discussion

We presented a generalization of the EXP3 algorithm, the EXP3++ algorithm, which augments the EXP3 algorithm with a new control lever in the form exploration parameters $\varepsilon_t(a)$ that are tuned individually for each arm. We have shown that the new control lever is extremely useful in detecting and exploiting the gap in a wide range of regimes, while the old control lever always keeps the worst-case performance of the algorithm under control. Due to the central role of the EXP3 algorithm in the adversarial analysis that stretches far beyond the adversarial bandits and due to the simplicity of our generalization we believe that our result will lead to a multitude of new algorithms for other problems that exploit the gaps without compromising on the worst-case performance guarantees. There is also room for further improvement of the presented technique that we plan to pursue in future work.

Acknowledgments

The authors would like to thank Sébastien Bubeck and Wouter Koolen for useful discussions and Csaba Szepesvári for bringing up the reference to Cesa-Bianchi & Fischer (1998). This research was supported by an Australian Research Council Australian Laureate Fellowship (FL110100281).

References

Agrawal, Shipra and Goyal, Navin. Further optimal regret bounds for Thompson sampling. In *AISTATS*, 2013.

Audibert, Jean-Yves and Bubeck, Sébastien. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2009.

Auer, Peter, Cesa-Bianchi, Nicolò, Freund, Yoav, and Schapire, Robert E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the Annual Symposium on Foundations of Computer Science*, 1995.

Auer, Peter, Cesa-Bianchi, Nicolò, and Fischer, Paul. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47, 2002a.

Auer, Peter, Cesa-Bianchi, Nicolò, Freund, Yoav, and Schapire, Robert E. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1), 2002b.

Babaioff, Moshe, Dughmi, Shaddin, Kleinberg, Robert, and Slivkins, Aleksandrs. Dynamic pricing with limited supply. In *13th ACM Conf. on Electronic Commerce (EC)*, 2012.

Boucheron, Stéphane, Lugosi, Gábor, and Massart, Pascal. *Concentration Inequalities A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

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

Bubeck, Sébastien and Cesa-Bianchi, Nicolò. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5, 2012.

Bubeck, Sébastien and Slivkins, Aleksandrs. The best of both worlds: stochastic and adversarial bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2012.

Cappé, Olivier, Garivier, Aurélien, Maillard, Odalric-Ambrym, Munos, Rémi, and Stoltz, Gilles. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3), 2013.

Cesa-Bianchi, Nicolò and Fischer, Paul. Finite-time regret bounds for the multiarmed bandit problem. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1998.

Cesa-Bianchi, Nicolò and Lugosi, Gábor. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

de Rooij, Steven, van Erven, Tim, Grünwald, Peter D., and Koolen, Wouter M. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 2014.

Kaufmann, Emilie, Korda, Nathaniel, and Munos, Rémi. Thompson sampling: An optimal finite time analysis. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2012.

Lai, Tze Leung and Robbins, Herbert. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6, 1985.

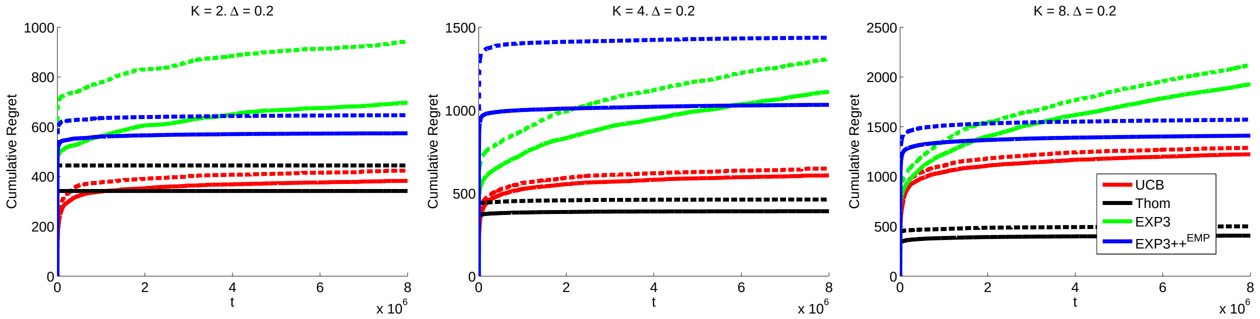
Maillard, Odalric-Ambrym. *Apprentissage Séquentiel: Bandits, Statistique et Renforcement*. PhD thesis, INRIA Lille, 2011.

Robbins, Herbert. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

Seldin, Yevgeny, Szepesvári, Csaba, Auer, Peter, and Abbasi-Yadkori, Yasin. Evaluation and analysis of the performance of the EXP3 algorithm in stochastic environments. In *JMLR Workshop and Conference Proceedings*, volume 24 (EWRL), 2013.

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

Thompson, William R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25, 1933.



(a) $K = 2, \Delta = 0.2$ (b) $K = 4, \Delta = 0.2$ (c) $K = 8, \Delta = 0.2$

Figure 2. **Perturbed Stochastic Environment.** Comparison of `UCB1`, Thompson sampling ("Thom"), `EXP3`, and `EXP3++`$^{\texttt{EMP}}$. The legend in figure (c) corresponds to all the figures. Solid lines correspond to means over 10 repetitions of the corresponding experiments and dashed lines correspond to the means plus one standard deviation.

A. Vulnerability of Deterministic Algorithms in the Adversarial Regime

We show that in the adversarial regime any deterministic algorithm can be forced to suffer linear regret. Let $\mathbb{A}$ be any deterministic algorithm. Note that given a sequence of losses up to time t the adversary knows which arm $\mathbb{A}$ will play on round $t + 1$. So the adversary can incrementally design a sequence of losses, such that the arm played by $\mathbb{A}$ always has loss 1 and all other arms have loss 0. On T rounds the loss of $\mathbb{A}$ will be T, whereas the loss of the best arm in hindsight will be at most T/K and the regret will be at least T/2.

B. Empirical Evaluation: Moderately Contaminated Stochastic Environment

We simulate moderately contaminated stochastic environment by drawing the first 2,000 rounds of the game according to one stochastic model and then switching the best arm and continuing the game until $T = 8,000,000$. We note that the contamination is not fully adversarial, but drawn from a different stochastic model. We run this experiment with $\Delta = 0.2$ and $K = 2$, 4, and 8 arms. The resutls are presented in Figure 2. It is hard to see the first 2,000 rounds on the graph, but their effect on all the algorithms is clearly visible. Despite the initial corrupted rounds the `EXP3++`$^{\texttt{EMP}}$ algorithm successfully returns to the stochastic operation mode and achieves better results than `EXP3`.

C. Proof of Lemma 10

Proof. It is easy to check by differentiation that

$$\int e^{-c\sqrt{t}}dt = -\frac{2}{c}\sqrt{t}e^{-c\sqrt{t}} - \frac{2}{c^2}e^{-c\sqrt{t}}.$$

Thus, we have:

$$\begin{aligned}\sum_{t=0}^{\infty} e^{-c\sqrt{t}} &= O\left(\int_{t=0}^{\infty} e^{-c\sqrt{t}}dt\right)\\ &= O\left(-\frac{2}{c}\sqrt{t}e^{-c\sqrt{t}} - \frac{2}{c^2}e^{-c\sqrt{t}}\Big|_0^{\infty}\right)\\ &= O\left(\frac{2}{c^2}\right).\end{aligned}$$

□

D. Definition of `EXP3++`$^{\texttt{LCBT}}$ and Proof of Theorem 4

As discussed in Section 3, if we assume a known time horizon T, then we can eliminate the additive term $e^{1/\Delta(a)^2}$ in the regret bound for `EXP3++`. For this purpose we define and analyze a version of `EXP3++`, called `EXP3++`$^{\texttt{LCBT}}$, in which we replace the empirical gap estimate (2) with a lower confidence bound on the gap.

D.1. Algorithm specification

Fix arm a and round t. Recall that $N_t(a)$ denotes the number of times this arm has been played by the algorithm up to round t, and let $\bar{\mu}_t(a)$ be the corresponding average loss. The difference $|\bar{\mu}_t(a) - \mu(a)|$ is bounded from above by the confidence term

$$\texttt{conf}_t(a) = \min\left(1, \sqrt{8\log(T)/N_t(a)}\right), \qquad (5)$$

with probability at least $1 - T^{-2}$. The minimal expected loss $\mu^* = \min_a \mu(a)$ can w.h.p. be upper-bounded by

$$\mu_t^* = \min_{\text{arms } a} \left(\, \bar{\mu}_t(a) + \texttt{conf}_t(a)\, \right).$$

The gap $\Delta(a)$ can w.h.p. be lower-bounded by

$$\Delta_t^{\texttt{LB}}(a) = \max\left(0,\ \bar{\mu}_t(a) - \texttt{conf}_t(a) - \mu_t^*\right). \qquad (6)$$

Using this lower bound, we define algorithm $\texttt{EXP3++}^{\texttt{LCBT}}$ to be a version of $\texttt{EXP3++}$ with

$$\xi_t(a) = \frac{\log^2 T}{t\,(\Delta_t^{\texttt{LB}}(a))^2}. \tag{7}$$

D.2. Regret bound and proof sketch

For convenience, we repeat the formulation of Theorem 4 in Theorem 13 below. Theorem 13 provides a regret bound for $\texttt{EXP3++}^{\texttt{LCBT}}$ in the stochastic regime with known time horizon.

Theorem 13. *Consider the stochastic regime with a known time horizon T and minimal gap Δ. The $\texttt{EXP3++}^{\texttt{LCBT}}$ algorithm with any $\eta_t \geq \beta_t$ achieves regret*

$$R(T) \leq \frac{O(\log^3 T)}{\Delta^3}. \tag{8}$$

Let us describe the key steps of the proof.

First, we capture two useful properties of $\Delta_t^{\texttt{LB}}(a)$.

Claim 14. *The following two events hold with probability at least $1 - O(\frac{1}{T})$:*

$$\{\Delta_t^{\texttt{LB}}(a) \leq \Delta(a) \quad \forall a, t\} \tag{9}$$

$$\{N_t(a) \geq \theta(a) \Rightarrow \Delta_t^{\texttt{LB}}(a) \geq \tfrac{1}{2}\,\Delta(a) \quad \forall a, t\}, \tag{10}$$

for threshold $\theta(a) = \frac{\Theta(K \log^2 T)}{\Delta^4(a)}$.

In fact, the rest of the proof uses $\Delta_t^{\texttt{LB}}(a)$ only through the above two properties. (Accordingly, our analysis applies to any other estimator $\Delta_t^{\texttt{LB}}(a) \in [0, 1]$ that satisfies the two properties with probability at least $1 - O(\frac{1}{T})$.)

The proof for (9) is a straightforward application of Azuma-Hoeffding inequality. Proving (10) requires a little subtlety to handle $\texttt{conf}_t(a^*)$. (Here and throughout, a^* is some best arm.)

Second, a simple but crucial computation identifies

$$\Delta_t^{\texttt{est}}(a) = \tfrac{1}{t}\left(\tilde{L}_t(a) - \tilde{L}_t(a^*)\right)$$

as a lever that our analysis can use to control the probability of choosing a suboptimal arm.

Claim 15. *Fix suboptimal action a and round t. Then*

$$\tilde{\rho}_t(a) \leq \exp\left(-\tfrac{1}{2K}\sqrt{t}\;\Delta_t^{\texttt{est}}(a)\right) + \varepsilon_t(a).$$

Proof. Recall that $\tilde{\rho}_t(a) \leq \rho_t(a) + \varepsilon_t(a)$. Denoting $w_t(a) = \exp(-\eta_t \tilde{L}_t(a))$, we have

$$\begin{aligned} \rho_t(a) &= w_t(a) / \sum_{a'} w_t(a') \leq w_t(a)/w_t(a^*) \\ &\leq \exp\left(-\eta_t(\tilde{L}_t(a) - \tilde{L}_t(a^*))\right). \end{aligned} \qquad \square$$

Next, we use the lower-bounding property of $\Delta_t^{\texttt{LB}}(\cdot)$ (Claim 14(a)) to deduce a lower bound on $\Delta_t^{\texttt{est}}(\cdot)$.

Lemma 16. *The following event holds with probability at least $1 - O(\frac{1}{T})$:*

$$\Delta_t^{\texttt{est}}(a) \geq \tfrac{1}{2}\,\Delta(a) \quad \forall\, a,\, t \geq t^*(a). \tag{11}$$

Here it suffices to take $t^*(a) \triangleq \frac{\Theta(K^2 \log^4 T)}{\Delta^4(a)}$.

Lemma 16 is proved by applying Bernstein's inequality (Theorem 9) to bound the deviation of $\tilde{L}_t(a)$ for each arm a. The crux is to bound from above the

$$\Sigma_n^2 = \sum_{i=1}^n \mathbb{E}\left[(X_i)^2 \middle| \mathcal{F}_{i-1}\right]$$

term in Theorem 9. Claim 14(a) boosts the exploration parameter $\varepsilon_t(\cdot)$, which essentially allows to upper-bound Σ_n^2 by $\tilde{O}(t^2\,\Delta^2(a))$ for a sufficiently large t, rather than merely by $\tilde{O}(t^2)$.

Now, plugging Equation (11) into Claim 15 implies that for $t \geq t^*(a)$ the probability of choosing a suboptimal arm a is essentially at most $\varepsilon_t(a)$. And *that* term can be upper-bounded using Claim 14(a): $\varepsilon_t(a) \leq \tilde{O}(\frac{1}{t\,\Delta^2(a)})$ after arm a has been chosen at least $\theta(a)$ times. Putting this together, we see that after some initial period the algorithm enters the regime where arm a is chosen with probability at most $\tilde{O}(\frac{1}{t\,\Delta^2(a)})$ in each round t, which implies that the total expected number of times arm a is chosen in this regime is at most $\tilde{O}(\frac{1}{\Delta^2(a)})$. The initial period includes at least $t^*(a)$ rounds and at least $\theta(a)$ plays of arm a, whichever comes latest. It is not hard to see that arm a can be selected at most $\max(t^*(a),\,\theta(a))$ times during this period. With some computations, the above argument proves that $\mathbb{E}\left[N_t(a)\right] \leq \frac{O(K^2 \log^4 t)}{\Delta^4(a)}$ for each suboptimal arm a. By Equation (1), this implies the claimed regret bound.

D.3. High-probability events

Before we delve into the detailed analysis, we use Azuma-Hoeffding inequality and Bernstein's inequality to set up several useful high-probability events.

First, we use the following corollary of Theorem 9.[2]

Theorem 17. *Let $X_1, \ldots, X_n$ be 0-1 random variables. Let $M = \sum_{t=1}^n M_t$, where $M_t = \mathbb{E}\left[X_t | X_1, \ldots, X_{t-1}\right]$ for each t. Then for any $b \geq 1$ the event*

$$|\textstyle\sum_{t=1}^n X_t - M_t| \leq b(\sqrt{M \log n} + \log n).$$

holds with probability at least $1 - n^{-\Omega(b)}$.

[2] For a self-contained proof, one can refer to Theorem 4.10 in the full version of (Babaioff et al., 2012).

Second, we approximate $\tilde{L}_t(a)$ with $t\mu(a)$, as long as $\varepsilon_t(\dot{})$ can be bounded from below.

Claim 18. *Fix arm a and round t. Suppose $\varepsilon_s(a) \geq \varepsilon_s$ for each round $s \leq t$ and some numbers $\varepsilon_1 \geq \varepsilon_2 \geq \ldots \geq \varepsilon_t \geq 0$. Denote $\nu = \sum_{s=1}^{t} \frac{1}{\varepsilon_s}$. Then for each $\lambda > 0$*

$$\mathbb{P}\left[|\tilde{L}_t(a) - t\mu(a)| \leq \sqrt{2\nu\lambda} + \frac{\sqrt{2}}{3}\frac{\lambda}{\varepsilon_t} \right] \geq 1 - 2\,e^{-\lambda}.$$

Proof. Note that $\tilde{\ell}_s(a)$, $s \leq t$ be form a martingale difference sequence such that for all rounds $s \leq t$ we have $|\tilde{\ell}_s(a)| \leq 1/\varepsilon_t$ and moreover

$$\begin{aligned}\sum_{s=1}^{t} \mathbb{E}_s\left[\tilde{\ell}_s^2(a)\right] &\leq \sum_{s=1}^{t} \mathbb{E}_s\left[\frac{1}{\tilde{p}_s(a)}\right] \\ &\leq \sum_{s=1}^{t} \mathbb{E}_s\left[\frac{1}{\varepsilon_s(a)}\right] \leq \sum_{s=1}^{t} \frac{1}{\varepsilon_s} = \nu.\end{aligned}$$

Thus, the theorem follows from Bernstein's inequality. □

Third, for each time interval $[t_0, t]$ let $n_{[t_0,t]}(a)$ be the number of times arm a is chosen in this time interval. We approximate it with $\hat{n}_{[t_0,t]}(a) = \sum_{s=t_0}^{t} \tilde{p}_s(a)$.

Claim 19. *Fix arm a, rounds $t_0 \leq t$, and $\lambda > 0$. Let $\hat{n} = \hat{n}_{[t_0,t]}(a)$. Then*

$$\mathbb{P}\left[\,|\; n_{[t_0,t]}(a) - \hat{n} \;| \leq O(\sqrt{\lambda\,\hat{n}} + \lambda)\right] \geq 1 - e^{-\lambda}. \quad (12)$$

Proof. This follows from Theorem 17. □

We will use the following high-probability event:

$$\begin{aligned}\{|\; n_{[t_0,t]}(a) &- \hat{n}_{[t_0,t]}(a) \;| \\ &\leq O(\sqrt{\log(T)\,\hat{n}_{[t_0,t]}(a)} + \log T) : \\ &\quad \forall a,\; t_0 \leq t\}. \qquad (13)\end{aligned}$$

This event holds with probability $1 - O(\frac{1}{T})$ by Claim 19.

D.4. Detailed analysis

To side-step some difficulties with handling low-probability events in early rounds, we define several high-probability events, and focus on the "clean execution" when all these events hold.

Proof of Claim 14, Equation (10). It suffices to focus on a *clean execution* of the algorithm: one in which events (9) and (13) hold.

Recall that for any arm a and any $\Delta > 0$

$$N_t(a) \geq \tfrac{8\,\log T}{\Delta^2} \Rightarrow \mathtt{conf}_t(a) \leq \tfrac{1}{4}\Delta. \qquad (14)$$

Fix suboptimal arm a and consider some round t such that $N_t(a) \geq \theta(a)$. Then $\mathtt{conf}_t(a) \leq \frac{1}{4}\Delta(a)$ by Equation (14) with $\Delta = \Delta(a)$. It remains to prove that $\mathtt{conf}_t(a^*) \leq \frac{1}{4}\Delta(a)$.

From the event in (9), $\tilde{\rho}_s(a^*) \geq \varepsilon_s(a^*) \geq \frac{A}{\sqrt{s}}$, where $A = \sqrt{\frac{\ln K}{2K}}$. It follows that

$$\begin{aligned}\hat{N}_t(a^*) &\triangleq \textstyle\sum_{s=1}^{t} \tilde{\rho}_s(a^*) \geq \sum_{s=1}^{t} \frac{A}{\sqrt{s}} \\ &\geq \textstyle\sum_{s=t/2}^{t} \frac{A}{\sqrt{2t}} \geq \frac{A}{2\sqrt{2}}\,\sqrt{t}.\end{aligned}$$

Noting that $t \geq N_t(a) \geq \theta(a)$, we have $\hat{N}_t(a^*) \geq \frac{2c\,\log T}{\Delta^4(a)}$ By (13), it follows that $N_t(a^*) \geq \frac{c\,\log T)}{\Delta^4(a)}$. Using Equation (14) with $\Delta = \Delta(a)$, we obtain $\mathtt{conf}_t(a^*) \leq \frac{1}{4}\Delta(a)$, completing the proof. □

Proof of Lemma 16. Let the threshold $t^*(a)$ be $t^*(a) \triangleq \frac{c\,K^2\,\log^4(T)}{\Delta^4(a)}$, where c is some absolute constant.

First, we reduce to the case when the event (9) holds deterministically. Indeed, consider *another* version of EXP3++ where $\Delta_t^{\mathtt{LB}}(a)$ is replaced by

$$\Delta_t^*(a) = \min(\Delta(a),\, \Delta_t^{\mathtt{LB}}(a)).$$

This new version satisfies (9) deterministically, and the two versions coincide with probability at least $1 - O(\frac{1}{T})$. This completes the reduction. From here on, we assume (9).

Fix round $t \leq T$ and parameter $\lambda > 0$. Denote $A = \sqrt{\frac{2K}{\ln K}}$ and $B = \frac{\Delta(a)}{\log^2(T)}$. For each arm a, we apply Claim 18 with

$$\begin{aligned}\varepsilon_s &= \min\left(\tfrac{1}{K},\, \tfrac{1}{A\sqrt{s}},\, \tfrac{1}{B\,s}\right). \\ \nu &\triangleq \textstyle\sum_{s=1}^{t} \frac{1}{\varepsilon_s} \leq \sum_{s=1}^{t} \left(K + A\sqrt{s} + Bs\right) \\ &= O(Kt + A\,t^{3/2} + B\,t^2).\end{aligned}$$

Plugging this into Claim 18, we obtain

$$\mathbb{P}\left[|\tilde{L}_t(a) - t\mu(a)| \leq \Gamma(a)\right] \leq 1 - 2\,e^{-\lambda},$$

where for any $t \geq (\lambda A)^2 + \frac{K^2}{A^2}$ and some constant $c > 0$

$$\Gamma(a) = c \cdot \left(\sqrt{\lambda A}\; t^{3/4} + (\lambda B + \sqrt{\lambda B})\; t\right).$$

Fix some suboptimal arm a. Then

$$\mathbb{P}\left[\tilde{L}_t(a) - \tilde{L}_t(a^*) < t\,\Delta(a) - \Gamma(a) - \Gamma(a^*)\right] < 2K\,e^{-\lambda}.$$

Take $\lambda = \log(KT^2)$. It suffices to prove that

$$\Gamma(a) + \Gamma(a^*) \leq \tfrac{t\,\Delta(a)}{2} \quad \forall\, t \geq t^*(a).$$

This holds because $\Gamma(a^*) = c\,\sqrt{\lambda A}\; t^{3/4}$ and

$$\begin{cases} 2c\,\sqrt{\lambda A}\; t^{3/4} \le \frac{1}{4} t\,\Delta(a) & \text{if} \quad t \ge \frac{(8c)^4\,(\lambda A)^2}{\Delta^4(a)},\\ c\,\lambda B \le c\,\sqrt{\lambda B} \le \frac{1}{8}\Delta(a) & \text{if} \quad \log^2 T \ge (8c)^2 \lambda. \end{cases}$$

□

Proof of Theorem 13. An execution of the algorithm is called *clean* if the following events hold:

- (13): $n_{[t_0,t]}(\cdot)$ is close to expectation.
- (10): estimate $\Delta_t^{\texttt{LB}}(\cdot)$ is sharp if $N_t(a) \ge \theta(a)$.
- (11): estimate $\tilde{L}_t(a) - \tilde{L}_t(a^*)$ is sharp if $t \ge t^*(a)$.

We have proved that an execution is clean with probability at least $1 - O(\frac{1}{T})$. So it suffices to focus on a clean execution from here on.

Fix a suboptimal arm a. By event (11), in each round $t \ge t^*(a)$ we have $\tilde{L}_t(a) - \tilde{L}_t(a^*) \ge \frac{1}{2}\, t\Delta(a)$. Plugging this into Claim 15, we obtain

$$\tilde{\rho}_t(a) \le \exp(-\tfrac{1}{2}\,\sqrt{t}\,\Delta(a)) + \varepsilon_t(a) \le \tfrac{1}{T} + \varepsilon_t(a).$$

Let $\theta(a)$ be the threshold from event (10). Assume that arm a is selected at least $\theta(a)$ times, i.e. that $N_t(a) \ge \theta(a)$ for some round t. Let t_0 be the smallest such round. Then for any round $t \ge t_0$ we have $\Delta_t^{\texttt{LB}}(a) \ge \frac{1}{2}\Delta(a)$, and consequently $\varepsilon_t(a) \le \frac{O(\log^2 T)}{t\,\Delta^2(a)}$.

Letting $t^\sharp = \max(t_0, t^*(a))$ we have

$$\hat{n}_{[t^\sharp,T]}(a) \triangleq \textstyle\sum_{s=t^\sharp}^{T} \tilde{\rho}_s(a) \le \frac{O(\log^3 T)}{\Delta^2(a)} \sum_{s=t^\sharp}^{T} \frac{1}{t} \le \frac{O(\log^3 T)}{\Delta^2(a)}.$$

Using event (13), we have $n_{[t^\sharp,T]}(a) \le \frac{O(\log^3 T)}{\Delta^2(a)}$.

Note that arm a can be played at most $\max(\theta(a), t^*(a))$ times before round $t^\sharp$. This is because arm a is played $\theta(a) \le t_0$ times before round t_0, and at most $t^*(a)$ times before round $t^*(a)$. Putting this together, we have

$$\begin{aligned} n_T(a) &= n_{[1,\,t^\sharp-1]}(a) + n_{[t^\sharp,T]}(a) \\ &\le \max(\theta(a), t^*(a)) + \tfrac{O(\log^3 T)}{\Delta^2(a)}. \end{aligned}$$

Plugging in $t^*(a) = \frac{O(K^2 \log^4 T)}{\Delta^4(a)}$ and $\theta(a) = \frac{O(K \log^2 T)}{\Delta^4(a)}$ we have $n_T(a) \le \frac{O(K^2 \log^4 T)}{\Delta^4(a)}$. Recall that this holds for each suboptimal arm a, in any clean execution of the algorithm. By Equation (1), this implies the claimed regret bound (8). □