An Improved Parametrization and Analysis of the EXP3++ Algorithm for Stochastic and Adversarial Bandits

Yevgeny Seldin   SELDIN@DI.KU.DK
University of Copenhagen

Gábor Lugosi   GABOR.LUGOSI@GMAIL.COM
ICREA and Pompeu Fabra University

Abstract

We present a new strategy for gap estimation in randomized algorithms for multiarmed bandits and combine it with the EXP3++ algorithm of Seldin and Slivkins (2014). In the stochastic regime the strategy reduces dependence of regret on a time horizon from $(\ln t)^3$ to $(\ln t)^2$ and eliminates an additive factor of order $\Delta e^{1/\Delta^2}$, where Δ is the minimal gap of a problem instance. In the adversarial regime regret guarantee remains unchanged.

1. Introduction

Stochastic (i.i.d.) and adversarial multiarmed bandits are two of the most basic problems in online learning (Thompson, 1933; Robbins, 1952; Lai and Robbins, 1985; Auer et al., 2002a, 1995, 2002b). In recent years there has been an increased interest in algorithms that can be applied in both settings (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016). This line of work can be seen as part of a growing area of "general-purpose" algorithms that are applicable to multiple online learning settings simultaneously (Seldin, 2015). The advantage of such algorithms is in their ability to exploit problem simplicity (such as i.i.d. environment) without compromising on the worst case guarantees.

There exist two basic approaches to deriving algorithms applicable to both stochastic and adversarial multiarmed bandits. The first starts with an algorithm for stochastic bandits and equips it with a mechanism for detecting deviations from the i.i.d. assumption. If such a deviation is detected, the algorithm switches into an adversarial operation mode (Bubeck and Slivkins, 2012; Auer and Chiang, 2016). The switch is irreversible and, therefore, this approach relies on a knowledge of time horizon. It allows to achieve $O\left(\sum_{a:\Delta(a)>0} \frac{\ln T}{\Delta(a)}\right)$ regret guarantee in the stochastic regime and $O\left(\sqrt{KT \ln T}\right)$ regret guarantee in the adversarial regime, where a indexes the arms, $\Delta(a)$ is the suboptimality gap of arm a, T is the number of game rounds, and K is the number of arms (Auer and Chiang, 2016). We note that in absence of the knowledge of time horizon the approach has to be combined with the doubling trick, which leads to deterioration of regret guarantee in the stochastic regime to $O\left(\sum_{a:\Delta(a)>0} \frac{(\ln t)^2}{\Delta(a)}\right)$ (we use capital T in results that assume a known time horizon and small t otherwise).

The second approach is to start with an algorithm for adversarial bandits and modify its exploration strategy to allow for gap detection. This approach has a number of advantages and disadvantages. On the positive side it has a single operation mode that naturally takes care of both regimes; it

does not rely on the knowledge of time horizon; it has a better regret guarantee of $O\left(\sqrt{Kt\ln K}\right)$ in the adversarial regime; and it can handle additional intermediate regimes, such as moderately contaminated stochastic regime and adversarial regime with a gap (Seldin and Slivkins, 2014). On the negative side its current regret guarantee in the stochastic regime is weaker, $O\left(\sum_{a:\Delta(a)>0}\frac{(\ln t)^3}{\Delta(a)}\right)$ with an exponentially large additive constant, and it does not provide high-probability regret guarantee in the adversarial regime, but only a guarantee on the expected regret. In our contribution we modify the second approach and improve its regret guarantee in the stochastic regime by a multiplicative factor of $\ln t$, as well as eliminate the exponentially large additive constant.

The work of Seldin and Slivkins (2014) is based on an observation that the EXP3 algorithm with losses for adversarial multiarmed bandits (Auer et al., 2002b; Bubeck and Cesa-Bianchi, 2012) has a degree of freedom in the choice of exploration strategy. Seldin and Slivkins have proposed a generalized EXP3++ algorithm based on a combination of two independent mechanisms. The first mechanism controls the performance of the algorithm in adversarial environments through a standard EXP3-like playing strategy in the form of a Gibbs distribution over actions. The second mechanism exploits the residual degree of exploration freedom for detection and exploitation of suboptimality gaps. The two mechanisms operate in parallel with almost no interference and achieve improved regret guarantee in the stochastic regime without impairing the adversarial regret guarantee.

We propose a new generic strategy for gap estimation in the stochastic regime that can be combined with almost any randomized playing strategy, including the EXP3++ algorithm. The new strategy is based on unweighted losses, as opposed to importance-weighted losses used in the main result of Seldin and Slivkins (2014, Theorem 3). It improves over the attempt of Seldin and Slivkins to use unweighted losses for gap estimation (Seldin and Slivkins, 2014, Theorem 4), both in terms of regret bound and in terms of underlying assumptions (the regret bound is improved by a multiplicative factor of order $\frac{\ln t}{\Delta}$ and the assumption on known time horizon is eliminated).

The proposed approach is modular: we provide an algorithm for gap estimation in the i.i.d. regime and then combine it with the EXP3++ algorithm, which provides protection against an adversary. The key features of the contribution are summarized below:

+ We propose a novel generic strategy for gap estimation by randomized algorithms in i.i.d. regimes. Our strategy can be combined with any randomized algorithm that has the necessary freedom in the choice of exploration distribution.

+ In combination with the EXP3++ the regret of the combined algorithm in the i.i.d. regime is of order $O\left(\sum_{a:\Delta(a)>0}\frac{(\ln t)^2}{\Delta(a)}\right)$, which is an improvement by a multiplicative factor of $\ln t$ compared to Seldin and Slivkins (2014). In the adversarial regime the regret guarantee is unchanged, $O\left(\sqrt{Kt\ln K}\right)$.

− The new approach does *not* provide an improved regret guarantee in the moderately contaminated stochastic regime and adversarial regime with a gap defined in Seldin and Slivkins (2014). For both regimes only the worst-case adversarial regret guarantee holds.

• Without the assumption on known time horizon the regret guarantee in the stochastic regime is of the same order as Auer and Chiang (2016) and the regret guarantee in the adversarial regime is stronger by a factor of $\sqrt{\ln t}$. However, our approach does not provide high-probability

guarantee in the adversarial regime, as do Auer and Chiang (2016). But on the positive side our approach is modular, it does not depend on the time horizon, and has a single operation mode for both stochastic and adversarial regimes, which makes it a bit more elegant.

In the following we start with outlining the problem setting in Section 2 and cite the EXP3++ algorithm and known results about it in Section 3. We present our gap estimation strategy in Section 4 and its combination with EXP3++ in Section 5. The corresponding proofs are given in Sections 6 and 7 and we finish with a discussion in Section 8.

2. Problem Setting

The problem setting follows Seldin and Slivkins (2014). We study the multiarmed bandit game. At round t of the game the algorithm chooses an *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 two of which, stochastic and adversarial, 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 (the so called oblivious learning model). Under this assumption we can assume that the loss sequences are determined before the game starts (but not revealed to the algorithm). We also make the standard assumption that the losses are bounded in the $[0,1]$ interval.

The performance of an algorithm is quantified by the *expected 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:

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

The expectation is taken over the possible randomness of the algorithm and the loss generation model. In the i.i.d. setting the ℓ_s^a-s are random variables and the definition coincides with the definition of pseudo regret (Bubeck and Cesa-Bianchi, 2012). In the adversarial setting the ℓ_s^a-s are considered deterministic and the second expectation can be omitted. In some literature $R(t)$ is termed *excess of cumulative predictive risk* (Wintenberger, 2017). Since $R(t)$ is the only notion of regret considered in the paper we will often call it simply *regret* (omitting the word "expected"). The goal of the algorithm is to minimize $R(t)$.

We consider two standard loss generation models, the *adversarial regime* and the *stochastic regime*.

Adversarial regime   In this regime the loss sequences are generated by an unrestricted adversary (who is oblivious to the algorithm's actions). 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 an arm a. An 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* $\Delta(a) = \mu(a) - \mu(a^*)$.

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

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

3

3. Known Results

In our work we are using the EXP3++ algorithm of Seldin and Slivkins (2014), which is provided in Algorithm 1.

Algorithm 1 EXP3++.

1: *Remark: See text for definition of η_t and $\xi_t(a)$; $\mathbb{1}\{\cdot\}$ is used to denote the indicator function*
2: $\forall a$: $\tilde{L}_0(a) = 0$
3: **for** $t = 1, 2, ...$ **do**
4:   $\forall a$: $\varepsilon_t(a) = \min\left\{ \frac{1}{2K}, \frac{1}{2}\sqrt{\frac{\ln K}{tK}}, \xi_t(a) \right\}$
5:   $\forall a$: $\rho_t(a) = e^{-\eta_t \tilde{L}_{t-1}(a)} / \sum_{a'} e^{-\eta_t \tilde{L}_{t-1}(a')}$
6:   $\forall a$: $\tilde{\rho}_t(a) = \left(1 - \sum_{a'} \varepsilon_t(a')\right) \rho_t(a) + \varepsilon_t(a)$
7:   Draw action A_t according to $\tilde{\rho}_t$ and play it
8:   Observe and suffer the loss $\ell_t^{A_t}$
9:   $\forall a :\ \tilde{\ell}_t^a = \frac{\ell_t^{A_t}}{\tilde{\rho}_t(a)} \mathbb{1}\{A_t = a\}$
10:   $\forall a :\ \tilde{L}_t(a) = \tilde{L}_{t-1}(a) + \tilde{\ell}_t^a$
11: **end for**

We note that since we are not changing the EXP3++ algorithm, but only modify the definition of the exploration parameters $\xi_t(a)$, the following result of Seldin and Slivkins (2014) is valid.

Theorem 1 (Seldin and Slivkins, 2014) *For $\eta_t = \frac{1}{2}\sqrt{\frac{\ln K}{tK}}$ and any $\xi_t(a) \geq 0$ the regret of the EXP3++ in the adversarial regime 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 and Cesa-Bianchi, 2012).

4. Gap Estimation in Randomized Playing Strategies

Our first contribution is a generic algorithm for gap estimation in stochastic environments. The algorithm can be combined with any randomized playing strategy, including the EXP3++. It is detailed in Algorithm 2. Line 8 is the "plug-in" point, where the algorithm can be combined with any randomized playing strategy. In combination with the EXP3++ we replace Line 8 in Algorithm 2 with Line 5 from Algorithm 1. (Note that lines 9-11 in Algorithm 2 are identical to lines 6-8 in Algorithm 1 and thus the two mechanisms can operate in parallel without interfering with each other.)

We use $\hat{L}_t(a)$ to denote unweighted cumulative loss of arm a up to (and including) round t. (It should not be confused with $\tilde{L}_t(a)$, which denotes cumulative importance-weighted loss and defined in Lines 9-10 of Algorithm 1.)

We provide the following guarantee for empirical gap estimates $\hat{\Delta}_t^{\text{LCB}}(a)$ in Line 6 of Algorithm 2.

Algorithm 2 Gap Estimation in Randomized Playing Strategies.

1: *Remark: see text for definition of $\xi_t(a)$*
2: Play each arm once and update $\hat{L}_K(a)$ and $N_K(a)$
3: **for** $t = K+1, K+2, ...$ **do**
4: $\quad \forall a:$ $\text{UCB}_t(a) = \min \left\{ 1, \frac{\hat{L}_{t-1}(a)}{N_{t-1}(a)} + \sqrt{\frac{\alpha \ln(tK^{1/\alpha})}{2N_{t-1}(a)}} \right\}$
5: $\quad \forall a:$ $\text{LCB}_t(a) = \max \left\{ 0, \frac{\hat{L}_{t-1}(a)}{N_{t-1}(a)} - \sqrt{\frac{\alpha \ln(tK^{1/\alpha})}{2N_{t-1}(a)}} \right\}$
6: $\quad \forall a:$ $\hat{\Delta}_t^{\text{LCB}}(a) = \max \left\{ 0, \text{LCB}_t(a) - \min_{a'} \text{UCB}_t(a') \right\}$ $\qquad$ // *Note that $0 \leq \hat{\Delta}_t^{\text{LCB}}(a) \leq 1$*
7: $\quad \forall a:$ $\varepsilon_t(a) = \min \left\{ \frac{1}{2K}, \frac{1}{2}\sqrt{\frac{\ln K}{tK}}, \xi_t(a) \right\}$
8: $\quad$ Let $\rho_t(a)$ be any distribution over $\{1, \ldots, K\}$ $\qquad$ // *The plug-in point for other algorithms*
9: $\quad \forall a:$ $\tilde{\rho}_t(a) = (1 - \sum_{a'} \varepsilon_t(a')) \rho_t(a) + \varepsilon_t(a)$
10: $\quad$ Draw action A_t according to $\tilde{\rho}_t$ and play it
11: $\quad$ Observe and suffer the loss $\ell_t^{A_t}$
12: $\quad \forall a:$ $\hat{L}_t(a) = \hat{L}_{t-1}(a) + \ell_t^a \mathbb{1}\{A_t = a\}$
13: $\quad \forall a:$ $N_t(a) = N_{t-1}(a) + \mathbb{1}\{A_t = a\}$
14: **end for**

Proposition 2 *For any a and t, the gap estimates $\hat{\Delta}_t^{LCB}(a)$ of Algorithm 2 in the i.i.d. regime satisfy:*

$$\mathbb{P}\left\{ \hat{\Delta}_t^{LCB}(a) \geq \Delta(a) \right\} \leq \frac{1}{t^{\alpha-1}}.$$

Furthermore, for any choice of $\xi_t(a)$, such that $\xi_t(a) \geq \frac{\beta \ln t}{t \hat{\Delta}_t^{LCB}(a)^2}$, for $\alpha \geq 3$, for $\beta \geq 64(\alpha+1) \geq$ 256, and $t \geq t_{min}(a) := \min \left\{ t : t \geq \frac{4K\beta(\ln t)^2}{\Delta(a)^4 \ln K} \right\}$ (this is the first time when $\frac{\beta \ln t}{t \Delta(a)^2} \leq \frac{1}{2}\sqrt{\frac{\ln K}{tK}}$) the gap estimates satisfy

$$\mathbb{P}\left\{ \hat{\Delta}_t^{LCB}(a) \leq \frac{1}{2}\Delta(a) \right\} \leq \left(\frac{\ln t}{t\Delta(a)^2} \right)^{\alpha-2} + \frac{2}{Kt^{\alpha-1}} + 2\left(\frac{1}{t}\right)^{\frac{\beta}{8}}.$$

A proof of this proposition is provided in Section 6. The main message of Proposition 2 is that for an appropriate choice of $\xi_t(a)$ the gap estimates $\hat{\Delta}_t^{\text{LCB}}(a)$ satisfy $\frac{1}{2}\Delta(a) \leq \hat{\Delta}_t^{\text{LCB}}(a) \leq \Delta(a)$ with high probability. Thus, $\hat{\Delta}_t^{\text{LCB}}(a)$ can be used as a reliable estimate of $\Delta(a)$ for any higher level purpose.

5. Reparametrization and Improved Regret Guarantee for EXP3++

We combine Algorithm 2 with the EXP3++ algorithm to achieve an improved regret guarantee in the stochastic regime.

Theorem 3 *Let $\xi_t(a) = \frac{\beta \ln t}{t \hat{\Delta}_t^{LCB}(a)^2}$, where $\hat{\Delta}_t^{LCB}(a)$ is the empirical gap estimate from Algorithm 2. Then for $\alpha = 3$ and $\beta = 256$ the expected regret of EXP3++ in the stochastic regime satisfies*

$$R_t = O \left(\sum_{a:\Delta(a)>0} \frac{(\ln t)^2}{\Delta(a)} \right) + \tilde{O} \left(\sum_{a:\Delta(a)>0} \frac{K}{\Delta(a)^3} \right),$$

where the $\tilde{O}$ notation hides logarithmic factors.

A proof of this theorem is provided in Section 7. We note that the regret guarantee of EXP3++ in the adversarial regime scales with $\sqrt{t}$. Therefore, the "logarithmic" regret guarantee in the stochastic regime becomes interesting when $\frac{(\ln t)^2}{\Delta(a)} < \sqrt{t}$ or $t \geq \tilde{\Omega} \left(\frac{1}{\Delta(a)^2} \right)$ (where the tilde notation hides logarithmic factors). The second term in the regret bound in Theorem 3 comes from the initial period of the game, where reliable estimate of the gaps cannot be achieved. The value of this term is only slightly suboptimal.

Theorem 3 improves the regret bound of Seldin and Slivkins (2014, Theorem 3) by a multiplicative factor of $\ln t$ and eliminates an exponentially large additive constant of order $\Delta e^{1/\Delta^2}$. We note that asymptotically the regret bound in Theorem 3 matches the oracle bound in Seldin and Slivkins (2014, Theorem 2), where knowledge of the gaps $\Delta(a)$ is assumed.

6. Proof of Proposition 2

The proof is based on four steps. In the first step we show that $\hat{\Delta}_t^{LCB}(a) \leq \Delta(a)$ with high probability. In the second step we derive a high-probability lower bound on the exploration parameters $\varepsilon_t(a)$. In the third step we derive a high-probability lower bound on the number of times $N_t(a)$ each arm is played. Finally, in the last step we show that $\hat{\Delta}_t^{LCB}(a) \geq \frac{1}{2}\Delta(a)$ with high probability.

Step 1: An upper bound for $\hat{\Delta}_t^{\textbf{LCB}}(a)$. The following property of upper and lower confidence bounds follows by standard arguments, as in Auer et al. (2002a). (The proof is standard and provided in the appendix for completeness.)

Lemma 4 *For any a and $t \geq K$:*

$$\mathbb{P}\{\text{UCB}_t(a) \leq \mu(a)\} \leq \frac{1}{Kt^{\alpha-1}},$$

$$\mathbb{P}\{\text{LCB}_t(a) \geq \mu(a)\} \leq \frac{1}{Kt^{\alpha-1}}.$$

Corollary 5 *For any a and $t \geq K$:*

$$\mathbb{P}\left\{\hat{\Delta}_t^{LCB}(a) \geq \Delta(a)\right\} \leq \frac{1}{t^{\alpha-1}}.$$

Proof

$$\mathbb{P}\left\{\hat{\Delta}_t^{\text{LCB}}(a) \geq \Delta(a)\right\} \leq \mathbb{P}\{\text{LCB}_t(a) \geq \mu(a)\} + \sum_{a' \neq a} \mathbb{P}\{\text{UCB}_t(a') \leq \mu(a')\} \leq \frac{1}{t^{\alpha-1}}.$$

■

Step 2: A lower bound for $\varepsilon_t(a)$. We have $\tilde{\rho}_t(a) \geq \varepsilon_t(a) = \min\left\{\frac{1}{2K}, \frac{1}{2}\sqrt{\frac{\ln K}{tK}}, \xi_t(a)\right\} \geq$
$\min\left\{\frac{1}{2K}, \frac{1}{2}\sqrt{\frac{\ln K}{tK}}, \frac{\beta \ln t}{t\hat{\Delta}_t^{\text{LCB}}(a)^2}\right\}$, where the last inequality is by the choice of $\xi_t(a)$. Note that $\hat{\Delta}_t^{\text{LCB}}(a)$
is a random variable. We derive a high-probability lower bound on the exploration probabilities.

Definition 6 *We define the following events:*

$$\mathcal{E}(a,t) = \left\{\forall s \in \{1,\ldots,t\} : \varepsilon_s(a) \geq \frac{\beta \ln t}{t\Delta(a)^2}\right\}$$

$$\mathcal{E}(a^*,a,t) = \left\{\forall s \in \{1,\ldots,t\} : \varepsilon_s(a^*) \geq \frac{\beta \ln t}{t\Delta(a)^2}\right\}.$$

By using Corollary 5 we get control over the probability of $\mathcal{E}(a,t)$ and $\mathcal{E}(a^*,a,t)$.

Lemma 7 *For $t \geq t_{min}(a)$ (where $t_{min}(a)$ is defined in Proposition 2) and $\alpha \geq 3$*

$$\mathbb{P}\left\{\overline{\mathcal{E}(a,t)}\right\} \leq \left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2},$$

$$\mathbb{P}\left\{\overline{\mathcal{E}(a^*,a,t)}\right\} \leq \left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2}.$$

Proof We start with the proof of the first inequality. Note that by definition $\hat{\Delta}_s^{\text{LCB}}(a) \leq 1$ and thus
$\frac{\beta \ln s}{s\hat{\Delta}_s^{\text{LCB}}(a)} \geq \frac{\beta \ln s}{s}$. For $s \leq t\Delta(a)^2/\ln t$ we have $\frac{\beta \ln s}{s} \geq \frac{\beta(\ln s)\ln t}{t\Delta(a)^2} \geq \frac{\beta \ln t}{t\Delta(a)^2}$ and thus $\varepsilon_s(a) \geq$
$\frac{\beta \ln t}{t\Delta(a)^2}$. Therefore, we have:

$$\mathbb{P}\left\{\overline{\mathcal{E}(a,t)}\right\} = \mathbb{P}\left\{\exists s \in \left[\frac{t\Delta(a)^2}{\ln t}, t\right] : \varepsilon_s(a) \leq \frac{\beta \ln t}{t\Delta(a)^2}\right\}$$

$$= \mathbb{P}\left\{\exists s \in \left[\frac{t\Delta(a)^2}{\ln t}, t\right] : \hat{\Delta}_t^{\text{LCB}}(a) \geq \Delta(a)\right\} \leq \sum_{s=\frac{t\Delta(a)^2}{\ln t}}^t \frac{1}{s^{\alpha-1}} \leq \frac{1}{2}\left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2}.$$

(The summation is bounded by using Lemma 11 in the appendix.) The bound for $\mathbb{P}\left\{\overline{\mathcal{E}(a^*,a,t)}\right\}$
follows the same lines using the fact that $\Delta(a^*) = 0 \leq \Delta(a)$. ■

Step 3: A lower bound for $N_t(a)$. We use the following concentration inequality.

Theorem 8 *Let $X_1,\ldots,X_n$ be Bernoulli random variables adapted to filtration $\mathcal{F}_1,\ldots,\mathcal{F}_n$ (in particular, X_i may depend on $X_1,\ldots,X_{i-1}$). Let $\mathcal{E}_\gamma$ be the event $\mathcal{E}_\gamma = \{\forall i : \mathbb{E}[X_i|\mathcal{F}_{i-1}] \geq \gamma\}$. Then*

$$\mathbb{P}\left\{\left(\sum_{i=1}^n X_i \leq \frac{1}{2}n\gamma\right) \wedge \mathcal{E}_\gamma\right\} \leq e^{-n\gamma/8}.$$

The result is based on quite standard techniques and its proof is provided in the appendix.

By Theorem 8, for $t \geq t_{\min}(a)$ we have:

$$\mathbb{P}\left\{N_t(a) \leq \frac{\beta \ln t}{2\Delta(a)^2}\right\} \leq \mathbb{P}\left\{\mathcal{E}(a,t) \wedge \left(N_t(a) \leq \frac{\beta \ln t}{2\Delta(a)^2}\right)\right\} + \mathbb{P}\left\{\overline{\mathcal{E}(a,t)}\right\}$$

$$\leq e^{-\frac{\beta \ln t}{8\Delta(a)^2}} + \frac{1}{2}\left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2}$$

$$\leq \left(\frac{1}{t}\right)^{\frac{\beta}{8\Delta(a)^2}} + \frac{1}{2}\left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2}$$

$$\leq \left(\frac{1}{t}\right)^{\frac{\beta}{8}} + \frac{1}{2}\left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2}.$$

In the same way we have $\mathbb{P}\left\{N_t(a^*) \leq \frac{\beta \ln t}{2\Delta(a)^2}\right\} \leq \left(\frac{1}{t}\right)^{\frac{\beta}{8}} + \frac{1}{2}\left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2}.$

Step 4: A lower bound for $\hat{\Delta}_t^{\mathbf{LCB}}(a)$. By Lemma 4 upper and lower confidence bounds satisfy $\mathbb{P}\left\{(\mathrm{UCB}_t(a^*) \leq \mu(a^*)) \vee (\mathrm{LCB}_t(a) \geq \mu(a))\right\} \leq \frac{2}{Kt^{\alpha-1}}$. Assuming that $\mathrm{UCB}_t(a^*) \geq \mu(a^*)$ and $\mathrm{LCB}_t(a) \leq \mu(a)$, we have:

$$\hat{\Delta}_t^{\mathrm{LCB}}(a) \geq \mathrm{LCB}_t(a) - \min_{a'} \mathrm{UCB}_t(a)$$

$$\geq \mathrm{LCB}_t(a) - \mathrm{UCB}_t(a^*)$$

$$= \frac{\hat{L}_{t-1}(a)}{N_{t-1}(a)} - \sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_t(a)}} - \frac{\hat{L}_{t-1}(a^*)}{N_{t-1}(a^*)} - \sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_t(a^*)}}$$

$$= \frac{\hat{L}_{t-1}(a)}{N_{t-1}(a)} + \sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_t(a)}} - 2\sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_t(a)}}$$

$$\quad - \left(\frac{\hat{L}_{t-1}(a^*)}{N_{t-1}(a^*)} - \sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_t(a^*)}}\right) - 2\sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_t(a^*)}}$$

$$= \mathrm{UCB}_t(a) - \mathrm{LCB}_t(a^*) - 2\sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_t(a)}} - 2\sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_t(a^*)}}$$

$$\geq \Delta(a) - 2\sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_t(a)}} - 2\sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_t(a^*)}}.$$

By Step 3, for $t \geq t_{\min}(a)$ we have $t \geq K$ and $\mathbb{P}\left\{\left(N_t(a) \leq \frac{\beta \ln t}{2\Delta(a)^2}\right) \vee \left(N_t(a^*) \leq \frac{\beta \ln t}{2\Delta(a)^2}\right)\right\} \leq$
$\left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2} + 2\left(\frac{1}{t}\right)^{\frac{\beta}{8}}$. Assuming that $N_t(a) > \frac{\beta \ln t}{2\Delta(a)^2}$ and $N_t(a^*) > \frac{\beta \ln t}{2\Delta(a)^2}$ we have:

$$\hat{\Delta}_t^{\mathrm{LCB}}(a) \geq \Delta(a) - 2\sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_t(a)}} - 2\sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_t(a^*)}}$$

$$\geq \Delta(a) - 4\sqrt{\frac{2\Delta(a)^2\alpha \ln\left(tK^{1/\alpha}\right)}{2\beta \ln t}}$$

$$\geq \Delta(a) - 4\sqrt{\frac{\Delta(a)^2(\alpha+1)\ln t}{\beta \ln t}}$$

$$= \Delta(a)\left(1 - 4\sqrt{\frac{\alpha+1}{\beta}}\right).$$

Taking everything together we obtain that for $t \geq t_{\min}(a)$ and $\beta \geq 64(\alpha+1)$ we have

$$\mathbb{P}\left\{\hat{\Delta}_t^{\mathrm{LCB}}(a) \leq \frac{1}{2}\Delta(a)\right\} \leq \left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2} + \frac{2}{Kt^{\alpha-1}} + 2\left(\frac{1}{t}\right)^{\frac{\beta}{8}}.$$

7. Proof of Theorem 3

In order to obtain a regret bound, for each suboptimal arm a we have to bound $\sum_{s=1}^{t}\mathbb{E}\left[\rho_s(a)\right]$ and $\sum_{s=1}^{t}\mathbb{E}\left[\varepsilon_s(a)\right]$. For the former we have:

$$\rho_t(a) = \frac{e^{-\eta_t \tilde{L}_{t-1}(a)}}{\sum_{a'} e^{-\eta_t \tilde{L}_{t-1}(a')}} = \frac{e^{-\eta_t\left(\tilde{L}_{t-1}(a)-\tilde{L}_{t-1}(a^*)\right)}}{\sum_{a'} e^{-\eta_t\left(\tilde{L}_{t-1}(a')-\tilde{L}_{t-1}(a^*)\right)}} \leq e^{-\eta_t\left(\tilde{L}_{t-1}(a)-\tilde{L}_{t-1}(a^*)\right)} = e^{-\eta_t \tilde{\Delta}_t(a)},$$

where $\tilde{\Delta}_t(a) = \tilde{L}_{t-1}(a) - \tilde{L}_{t-1}(a^*)$ is the gap between cumulative importance-weighted estimates of the losses. Unfortunately, the bound on unweighted gap estimates $\hat{\Delta}_t^{\mathrm{LCB}}(a)$ provided by Proposition 2 does not directly lead to a bound on the weighted gap estimates $\tilde{\Delta}_t(a)$ and, therefore, does not provide a bound on $\rho_t(a)$. We use the following form on Bernstein's inequality for martingales to achieve this goal. Theorem 9 is a minor variation of a classical Bernstein's inequality for martingales (Freedman, 1975), where we relax the assumption on boundedness of the martingale difference sequence. The theorem follows by a simple adaptation of the proof by McDiarmid (1998, Theorem 3.15), which is sketched in the appendix.

Theorem 9 (Bernstein's inequality for martingales) *Let $X_1, \ldots, X_n$ be a martingale difference sequence with respect to filtration $\mathcal{F}_1, \ldots, \mathcal{F}_n$, where each X_j is bounded from above, and let $S_i = \sum_{j=1}^{i} X_j$ be the associated martingale. Let $\nu_n = \sum_{j=1}^{n}\mathbb{E}\left[(X_j)^2\big|\mathcal{F}_{j-1}\right]$ and $c_n = \max_{1 \leq j \leq n}\{X_j\}$. Then for any $\delta > 0$:*

$$\mathbb{P}\left\{\left(S_n \geq \sqrt{2\nu \ln\frac{1}{\delta}} + \frac{c \ln\frac{1}{\delta}}{3}\right) \wedge (\nu_n \leq \nu) \wedge (c_n \leq c)\right\} \leq \delta.$$

9

We apply this theorem to martingale difference sequence $X_s = \Delta(a) - \left(\tilde{\ell}_s^a - \tilde{\ell}_s^{a^*} \right)$ with respect to filtration $\mathcal{F}_1, \mathcal{F}_2, \dots$ in order to bound the martingale $t\Delta(a) - \tilde{\Delta}_t(a) = \sum_{s=1}^t X_s$. We start by bounding the magnitude of X_s-es and the sum of their conditional variances and then use Bernstein's inequality to bound $\tilde{\Delta}_t(a)$. The bound on $\tilde{\Delta}_t(a)$ is then used to bound $\rho_t(a)$. At the end we treat the second term of the regret bound, $\sum_{s=1}^t \mathbb{E}\left[\varepsilon_s(a)\right]$.

Control of the magnitude of $\max\limits_{1 \le s \le t} \{X_s\}$. We start by bounding the magnitude of the martingale difference sequence $X_s = \Delta(a) - \left(\tilde{\ell}_s^a - \tilde{\ell}_s^{a^*} \right)$. We have:

$$\Delta(a) - \left(\tilde{\ell}_s^a - \tilde{\ell}_s^{a^*} \right) \le 1 + \tilde{\ell}_s^{a^*}$$

$$\le 1 + \frac{1}{\varepsilon_s(a^*)}$$

$$= 1 + \max\left\{ 2K, 2\sqrt{\frac{sK}{\ln K}}, \frac{s\hat{\Delta}_s^{\mathrm{LCB}}(a^*)^2}{\beta \ln s} \right\}$$

$$\le 1.25 \max\left\{ 2K, 2\sqrt{\frac{sK}{\ln K}}, \frac{s\hat{\Delta}_s^{\mathrm{LCB}}(a^*)^2}{\beta \ln s} \right\}.$$

Note that $\hat{\Delta}_s^{\mathrm{LCB}}(a^*) \le 1$ and thus for $t \ge t_{\min}(a)$ and $s \le t\Delta(a)^2/\ln t$ we have $\frac{1}{\varepsilon_s(a^*)} \le \frac{t\Delta(a)^2}{\beta \ln t}$ (we have a^* on the left-hand side and a on the right-hand side). Furthermore, by Proposition 2 we have that $\mathbb{P}\left\{ \hat{\Delta}_s^{\mathrm{LCB}}(a^*) \ge \Delta(a^*) \right\} \le \frac{1}{s^{\alpha-1}}$, where $\Delta(a^*) = 0$. Thus,

$$\mathbb{P}\left\{ \exists s \in \left[\frac{t\Delta(a)^2}{\ln t}, t \right] : \hat{\Delta}_s^{\mathrm{LCB}}(a^*) \ge 0 \right\} \le \sum_{s=t\Delta(a)^2/\ln t}^t \frac{1}{s^{\alpha-1}} \le \frac{1}{2} \left(\frac{\ln t}{t\Delta(a)^2} \right)^{\alpha-2}.$$

Let $c_t = \max\limits_{1 \le s \le t} \{X_s\}$. We obtain that for $t \ge t_{\min}(a)$

$$\mathbb{P}\left\{ c_t \ge \frac{1.25 t\Delta(a)^2}{\beta \ln t} \right\} \le \frac{1}{2} \left(\frac{\ln t}{t\Delta(a)^2} \right).$$

Control of the sum of conditional variances $\sum_{s=1}^t \mathbb{E}\left[(X_s)^2 \middle| \mathcal{F}_{s-1}\right]$. We start by looking at individual terms in the sum. We have:

$$\mathbb{E}\left[\left(\Delta(a) - \left(\tilde{\ell}_s^a - \tilde{\ell}_s^{a^*} \right) \right)^2 \middle| \mathcal{F}_{s-1} \right] \le \mathbb{E}\left[\left(\tilde{\ell}_s^a - \tilde{\ell}_s^{a^*} \right)^2 \middle| \mathcal{F}_{s-1} \right]$$

$$= \mathbb{E}\left[\left(\tilde{\ell}_s^a \right)^2 \middle| \mathcal{F}_{s-1} \right] + \mathbb{E}\left[\left(\tilde{\ell}_s^{a^*} \right)^2 \middle| \mathcal{F}_{s-1} \right],$$

where the equality is due to the fact that by the way importance-weighted samples are defined we have $\tilde{\ell}_s^a \tilde{\ell}_s^{a^*} = 0$ and thus $\left(\tilde{\ell}_s^a - \tilde{\ell}_s^{a^*} \right)^2 = \left(\tilde{\ell}_s^a \right)^2 + \left(\tilde{\ell}_s^{a^*} \right)^2$. Further,

$$\mathbb{E}\left[\left(\tilde{\ell}_s^a \right)^2 \middle| \mathcal{F}_{s-1} \right] = \tilde{\rho}_s(a) \left(\frac{\ell_s^a}{\tilde{\rho}_s(a)} \right)^2 \le \frac{1}{\tilde{\rho}_s(a)} \le \frac{1}{\varepsilon_s(a)} = \max\left\{ 2K, 2\sqrt{\frac{sK}{\ln K}}, \frac{s\hat{\Delta}_s^{\mathrm{LCB}}(a)^2}{\beta \ln s} \right\}.$$

Note that $\hat{\Delta}_s^{\mathrm{LCB}}(a) \leq 1$ and thus for $t \geq t_{\min}(a)$ and $s \leq t\Delta(a)^2/\ln t$ we have $\frac{1}{\varepsilon_s(a)} \leq \frac{t\Delta(a)^2}{\beta \ln t}$. Furthermore,

$$\mathbb{P}\left\{\exists s \in \left[\frac{t\Delta(a)^2}{\ln t}, t\right] : \hat{\Delta}_s^{\mathrm{LCB}}(a) \geq \Delta(a)\right\} \leq \sum_{s=\frac{t\Delta(a)^2}{\ln t}}^{t} \frac{1}{s^{\alpha-1}} \leq \frac{1}{2}\left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2}.$$

We define $\nu_t = \sum_{s=1}^{t} \mathbb{E}\left[\left(\Delta(a) - \left(\tilde{\ell}_s^a - \tilde{\ell}_s^{a^*}\right)\right)^2 \Big| \mathcal{F}_{s-1}\right]$ and we have that for $t \geq t_{\min}(a)$

$$\mathbb{P}\left\{\nu_t \geq \frac{2t^2\Delta(a)^2}{\beta \ln t}\right\}$$

$$\leq \mathbb{P}\left\{\exists s \in \left[\frac{t\Delta(a)^2}{\ln t}, t\right] : \hat{\Delta}_s^{\mathrm{LCB}}(a) \geq \Delta(a)\right\} + \mathbb{P}\left\{\exists s \in \left[\frac{t\Delta(a)^2}{\ln t}, t\right] : \hat{\Delta}_s^{\mathrm{LCB}}(a^*) \geq 0\right\}$$

$$\leq \left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2}.$$

Note that the random event involving $\hat{\Delta}_s^{\mathrm{LCB}}(a^*)$ is the same as the one we have considered in Step 1. Thus, in total $\mathbb{P}\left\{c_t \geq \frac{1.25t\Delta(a)^2}{\beta \ln t}\right\} + \mathbb{P}\left\{\nu_t \geq \frac{2t^2\Delta(a)^2}{\beta \ln t}\right\} \leq \left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2}.$

Control of $\tilde{\Delta}_t(a)$. We have that

$$\mathbb{P}\left\{\tilde{\Delta}_t(a) \leq \frac{1}{2}t\Delta(a)\right\}$$

$$= \mathbb{P}\left\{t\Delta(a) - \tilde{\Delta}_t(a) \geq \frac{1}{2}t\Delta(a)\right\}$$

$$\leq \mathbb{P}\left\{c_t \geq \frac{1.25t\Delta(a)^2}{\beta \ln t}\right\} + \mathbb{P}\left\{\nu_t \geq \frac{2t^2\Delta(a)^2}{\beta \ln t}\right\}$$

$$+ \mathbb{P}\left\{\left(t\Delta(a) - \tilde{\Delta}_t(a) \geq \frac{1}{2}t\Delta(a)\right) \wedge \left(\nu_t \leq \frac{2t^2\Delta(a)^2}{\beta \ln t}\right) \wedge \left(c_t \leq \frac{1.25t\Delta(a)^2}{\beta \ln t}\right)\right\}.$$

Taking $\nu = \frac{2t^2\Delta(a)^2}{\beta \ln t}$, $c = \frac{1.25t\Delta(a)^2}{\beta \ln t}$, and $\delta = \frac{1}{t}$, for $\beta \geq 256$ we have

$$\sqrt{2\nu \ln \frac{1}{\delta}} + \frac{c \ln \frac{1}{\delta}}{3} = \sqrt{\frac{4t^2\Delta(a)^2 \ln t}{\beta \ln t}} + \frac{1.25t\Delta(a)^2 \ln t}{3\beta \ln t} \leq t\Delta(a)\left(\frac{2}{\sqrt{\beta}} + \frac{1.25}{3\beta}\right) \leq \frac{1}{2}t\Delta(a)$$

and by Bernstein's inequality the last term is bounded by $\frac{1}{t}$. Overall, for $t \geq t_{\min}(a)$:

$$\mathbb{P}\left\{\tilde{\Delta}_t(a) \leq \frac{1}{2}t\Delta(a)\right\} \leq \left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2} + \frac{1}{t}.$$

Control of $\sum_{s=1}^{t} \mathbb{E}\left[\rho_s(a)\right]$. From here we have for $\eta_t \geq \frac{1}{2}\sqrt{\frac{\ln K}{tK}}$ and $\alpha \geq 3$:

$$
\sum_{s=1}^{t} \mathbb{E}\left[\rho_s(a)\right] \leq \sum_{s=1}^{t} \mathbb{E}\left[e^{-\eta_s \tilde{\Delta}_s(a)}\right]
$$

$$
\leq t_{\min}(a) + \sum_{s=t_{\min}(a)}^{t} \left(e^{-\frac{1}{2}\eta_s s \Delta(a)} + \left(\frac{\ln s}{s\Delta(a)^2}\right)^{\alpha-2} + \frac{1}{s}\right)
$$

$$
\leq t_{\min}(a) + \frac{\left((\ln t)^2 + \ln t\right)}{\Delta(a)^2} + \ln t + 1 + \sum_{s=t_{\min}(a)}^{t} e^{-\frac{1}{4}\sqrt{\frac{s\ln K}{K}}\Delta(a)}
$$

$$
\leq \frac{\left((\ln t)^2 + \ln t\right)}{\Delta(a)^2} + \ln t + 1 + \frac{16K}{\Delta(a)^2 \ln K} + 1 + t_{\min}(a).
$$

Control of $\sum_{s=1}^{t} \mathbb{E}\left[\varepsilon_s(a)\right]$. By Proposition 2, for $t \geq t_{\min}(a)$ we have that $\mathbb{P}\left\{\hat{\Delta}_t^{\mathrm{LCB}}(a) \leq \frac{1}{2}\Delta(a)\right\} \leq \left(\frac{\ln t}{t\Delta(a)^2}\right)^{\alpha-2} + \frac{2}{Kt^{\alpha-1}} + 2\left(\frac{1}{t}\right)^{\frac{\beta}{8}}$. Thus, for $\alpha = 3$ and $\beta = 256$ we have

$$
\sum_{s=1}^{t} \mathbb{E}\left[\varepsilon_s(a)\right] = \sum_{s=1}^{t} \mathbb{E}\left[\min\left\{\frac{1}{2K}, \frac{1}{2}\sqrt{\frac{\ln s}{sK}}, \frac{\beta \ln s}{s\hat{\Delta}_s^{\mathrm{LCB}}(a)^2}\right\}\right]
$$

$$
\leq \sum_{s=1}^{t} \mathbb{E}\left[\frac{\beta \ln s}{s\hat{\Delta}_s^{\mathrm{LCB}}(a)^2}\right]
$$

$$
\leq t_{\min}(a) + \frac{4\beta\left((\ln t)^2 + \ln t\right)}{\Delta(a)^2} + \sum_{s=t_{\min}(a)}^{t} \left(\left(\frac{\ln s}{s\Delta(a)^2}\right)^{\alpha-2} + \frac{2}{Ks^{\alpha-1}} + \left(\frac{1}{s}\right)^{\frac{\beta}{8}}\right)
$$

$$
\leq t_{\min}(a) + \frac{4\beta\left((\ln t)^2 + \ln t\right)}{\Delta(a)^2} + \frac{(\ln t)^2 + \ln t}{\Delta(a)^2} + \frac{2}{K}\left(\ln t + 1\right) + \frac{2\pi^2}{3}.
$$

By combining the bounds on $\sum_{s=1}^{t} \mathbb{E}\left[\rho_s(a)\right]$ and $\sum_{s=1}^{t} \mathbb{E}\left[\varepsilon_s(a)\right]$ we obtain that $\mathbb{E}\left[N_t(a)\right] = O\left(\sum_{a:\Delta(a)>0} \frac{(\ln t)^2}{\Delta(a)^2}\right) + \tilde{O}\left(\frac{K}{\Delta(a)^4}\right)$, which leads to the statement of the theorem.

8. Discussion

We have proposed a new algorithm for gap estimation in stochastic environments that can be combined with other randomized algorithms in a modular fashion. The algorithm provides a gap estimate $\hat{\Delta}_t^{\mathrm{LCB}}(a)$ that satisfies $\frac{1}{2}\Delta(a) \leq \hat{\Delta}_t^{\mathrm{LCB}}(a) \leq \Delta(a)$ with high probability. We have shown that the algorithm can be combined with the EXP3++ algorithm, leading to $O\left(\sqrt{Kt\ln K}\right)$ regret in the adversarial regime and $O\left(\sum_{a:\Delta(a)>0} \frac{(\ln t)^2}{\Delta(a)}\right)$ regret in the stochastic regime, where the latter is an improvement by a multiplicative factor of $\ln t$ over Seldin and Slivkins (2014).

Our work leads to a number of interesting directions for future research. First, there is a question whether the dependence of the regret guarantee on time horizon in the stochastic regime can be

reduced down to $\ln t$. We note that Auer and Chiang (2016) have a lower bound on achievable regret guarantees in the stochastic regime when simultaneously certain expected regret guarantees against an adaptive adversary or high-probability regret guarantees against an oblivious adversary are required. However, it is still unknown whether $\ln t$ regret in the stochastic regime can be achieved simultaneously with $\sqrt{t}$ expected regret against an oblivious adversary. While it does not seem possible to achieve it with the EXP3++ algorithm, some modifications of the playing rule, such as the one used in BOA (Wintenberger, 2017), could potentially do better.

A second question is whether improved regret guarantees can be achieved in the moderately contaminated stochastic regime and adversarial regime with a gap. We believe that it might not be possible with gap estimation strategies based on unweighted rewards and that in order to achieve that we should improve gap estimation based on importance-weighted rewards. The analysis technique suggested in our paper could potentially be useful for that.

There are also a number of more technical questions. For example, can we achieve high-probability regret guarantees by turning to modifications of the EXP3 algorithm, such as EXP3-IX (Neu, 2015)? Or could we replace $\frac{1}{\Delta(a)}$ factors with more refined measures of complexity, such as those in kl-UCB-type algorithms (Cappé et al., 2013)?

Acknowledgments

References

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2016.

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

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

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

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

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

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

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1), 1975.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301), 1963.

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

Colin McDiarmid. Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*. Springer, 1998.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

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

Yevgeny Seldin. The space of online learning problems. ECML-PKDD Tutorial. https://sites.google.com/site/spaceofonlinelearningproblems/, 2015.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *JMLR Workshop and Conference Proceedings*, volume 32 (ICML), 2014.

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

Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106, 2017.

Appendix A. Proof of Lemma 4

The proof is based on Hoeffding's inequality (Hoeffding, 1963).

Theorem 10 (Hoeffding's inequality) *Let $X_1, \ldots, X_n$ be i.i.d. random variables, such that $0 \leq X_i \leq 1$ and $\mathbb{E}[X_i] = \mu$ for all i. Then*

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right\} \leq \delta,$$

$$\mathbb{P}\left\{\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right\} \leq \delta.$$

Proof of Lemma 4 The proof directly follows the analysis of confidence bounds in Auer et al. (2002a). Note that $N_{t-1}(a)$ is a random variable dependent on $\hat{L}_{t-1}(a)$ and we cannot apply Hoeffding's inequality directly. Let $X_1, \ldots, X_t$ be i.i.d. random variables with the same distribution as ℓ_1^a and let $\hat{M}_s = \sum_{r=1}^{s} X_r$. Then

$$
\begin{aligned}
\mathbb{P}\left\{\mathrm{UCB}_t(a) \leq \mu(a)\right\} &= \mathbb{P}\left\{\frac{\hat{L}_{t-1}(a)}{N_{t-1}(a)} + \sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2N_{t-1}(a)}} \leq \mu(a)\right\} \\
&\leq \mathbb{P}\left\{\exists s \in \{1, \ldots, t-1\} : \frac{\hat{M}_s}{s} + \sqrt{\frac{\alpha \ln\left(tK^{1/\alpha}\right)}{2s}} \leq \mu(a)\right\} \\
&\leq \sum_{s=1}^{t-1} \mathbb{P}\left\{\mu(a) - \frac{\hat{M}_s}{s} \geq \sqrt{\frac{\ln\left(t^{\alpha}K\right)}{2s}}\right\} \\
&\leq \sum_{s=1}^{t-1} \frac{1}{Kt^{\alpha}} \\
&\leq \frac{1}{Kt^{\alpha-1}}.
\end{aligned}
$$

The proof of the second inequality in the lemma is analogous. ∎

Appendix B. Partial Sum of Reciprocals of Powers of Natural Numbers

Lemma 11 *For $\alpha \geq 2$ and $m \geq 1$:*

$$
\sum_{k=m}^{n} \frac{1}{k^{\alpha}} \leq \frac{1}{2m^{\alpha-1}}.
$$

Proof We have $2k^2 \geq k(k+1)$ and $\sum_{k=m}^{n} \frac{1}{k(k+1)} \leq \frac{1}{m}$ (which is obtained by writing $\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}$). Thus:

$$
\sum_{k=m}^{n} \frac{1}{k^{\alpha}} \leq \frac{1}{2} \sum_{k=m}^{n} \frac{1}{k(k+1)k^{\alpha-2}} \leq \frac{1}{2m^{\alpha-2}} \sum_{k=m}^{n} \frac{1}{k(k+1)} \leq \frac{1}{2m^{\alpha-1}}.
$$

∎

Appendix C. Proof of Theorem 8

Proof We start with a bound on a moment generating function of a single Bernoulli random variable X. For any $\lambda > 0$ we have

$$
\mathbb{E}\left[e^{-\lambda X}\right] \leq \mathbb{E}\left[1 - \lambda X + \frac{\lambda^2 X^2}{2}\right] = 1 - \left(\lambda - \frac{\lambda^2}{2}\right)\mathbb{E}[X] \leq e^{-\left(\lambda - \frac{\lambda^2}{2}\right)\mathbb{E}[X]}.
$$

And, as a consequence,

$$\mathbb{E}\left[e^{\left(\lambda-\frac{\lambda^2}{2}\right)\mathbb{E}[X]-\lambda X}\right] \leq 1.$$

For $\lambda \in (0,2]$ we have $\lambda - \frac{\lambda^2}{2} \geq 0$. Therefore, if $\mathbb{E}[X] \geq \gamma$ then $\mathbb{E}\left[e^{\left(\lambda-\frac{\lambda^2}{2}\right)\gamma-\lambda X}\right] \leq \mathbb{E}\left[e^{\left(\lambda-\frac{\lambda^2}{2}\right)\mathbb{E}[X]-\lambda X}\right] \leq 1$.

Let $Z = \mathbb{1}\{\mathcal{E}_\gamma\}$ denote the indicator random variable of the event $\mathcal{E}_\gamma$. Then for $\lambda \in (0,2]$ we have

$$\begin{aligned}
\mathbb{E}\left[Ze^{\left(\lambda-\frac{\lambda^2}{2}\right)n\gamma-\lambda\sum_{i=1}^{n}X_i}\right] &= \mathbb{E}\left[Ze^{\left(\lambda-\frac{\lambda^2}{2}\right)(n-1)\gamma-\lambda\sum_{i=1}^{n-1}X_i}\mathbb{E}\left[Ze^{\left(\lambda-\frac{\lambda^2}{2}\right)\gamma-\lambda X_n}\Big|\mathcal{F}_{n-1}\right]\right] \\
&\leq \mathbb{E}\left[Ze^{\left(\lambda-\frac{\lambda^2}{2}\right)(n-1)\gamma-\lambda\sum_{i=1}^{n-1}X_i}\mathbb{E}\left[e^{\left(\lambda-\frac{\lambda^2}{2}\right)\mathbb{E}[X_n|\mathcal{F}_{n-1}]-\lambda X_n}\Big|\mathcal{F}_{n-1}\right]\right] \\
&\leq \mathbb{E}\left[e^{\left(\lambda-\frac{\lambda^2}{2}\right)(n-1)\gamma-\lambda\sum_{i=1}^{n-1}X_i}\right] \\
&\leq \cdots \\
&\leq 1.
\end{aligned}$$

By combining this result with Markov's inequality we have that for any $\lambda \in (0,2]$

$$\begin{aligned}
\mathbb{P}\left\{\left(\sum_{i=1}^{n}X_i \leq n\gamma - \frac{\lambda}{2}n\gamma - \frac{\ln\frac{1}{\delta}}{\lambda}\right) \wedge \mathcal{E}_\gamma\right\} &= \mathbb{P}\left\{\left(\left(\lambda - \frac{\lambda^2}{2}\right)n\gamma - \lambda\sum_{i=1}^{n}X_i \geq \ln\frac{1}{\delta}\right) \wedge \mathcal{E}_\gamma\right\} \\
&= \mathbb{P}\left\{Ze^{\left(\lambda+\frac{\lambda^2}{2}\right)n\gamma-\lambda\sum_{i=1}^{n}X_i} \geq \frac{1}{\delta}\right\} \\
&\leq \delta\mathbb{E}\left[Ze^{\left(\lambda+\frac{\lambda^2}{2}\right)n\gamma-\lambda\sum_{i=1}^{n}X_i}\right] \\
&\leq \delta.
\end{aligned}$$

By taking $\lambda = \sqrt{\frac{n\gamma}{2\ln\frac{1}{\delta}}}$ we obtain

$$\mathbb{P}\left\{\left(\sum_{i=1}^{n}X_i \leq n\gamma - \sqrt{2n\gamma\ln\frac{1}{\delta}}\right) \wedge \mathcal{E}_\gamma\right\} \leq \delta.$$

Finally, taking $\delta = e^{-n\gamma/8}$ leads to $\lambda = 2$ and completes the proof. ∎

Appendix D. Proof sketch of Theorem 9 (Bernstein's inequality)

The proof is analogous to the proof of McDiarmid (1998, Theorem 3.15). McDiarmid assumes that X_j-s are bounded by c and the proof is based on defining an indicator random variable $Z = \mathbb{1}\{\nu_n \leq \nu\}$ and bounding $\mathbb{P}\{(S_n \geq \alpha) \wedge (\nu_n \leq \nu)\} = \mathbb{P}\{Ze^{\lambda S_n} \geq e^{\lambda\alpha}\}$ for $\lambda > 0$. We remove the assumption and define an indicator random variable $Z' = \mathbb{1}\{(\nu_n \leq \nu) \wedge (c_n \leq c)\}$. Then

$\mathbb{P}\left\{(S_n \geq \alpha) \wedge (\nu_n \leq \nu) \wedge (c_n \leq c)\right\} = \mathbb{P}\left\{Z'e^{\lambda S_n} \geq e^{\lambda \alpha}\right\}$ for $\lambda > 0$ and the rest of the proof is identical.