JMLR: Workshop and Conference Proceedings 24:103–116, 2012 10th European Workshop on Reinforcement Learning

Evaluation and Analysis of the Performance of the EXP3 Algorithm in Stochastic Environments

Yevgeny Seldin SELDIN@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Tübingen, Germany
University College London, London, UK

Csaba Szepesvári SZEPESVA@UALBERTA.CA
University of Alberta, Edmonton, Canada

Peter Auer AUER@UNILEOBEN.AC.AT
Montanuniversität Leoben, Austria

Yasin Abbasi-Yadkori ABBASIYA@UALBERTA.CA
University of Alberta, Edmonton, Canada

Editor: Marc Peter Deisenroth, Csaba Szepesvári, Jan Peters

Abstract

EXP3 is a popular algorithm for adversarial multiarmed bandits, suggested and analyzed in this setting by Auer et al. [2002b]. Recently there was an increased interest in the performance of this algorithm in the stochastic setting, due to its new applications to stochastic multiarmed bandits with side information [Seldin et al., 2011] and to multiarmed bandits in the mixed stochastic-adversarial setting [Bubeck and Slivkins, 2012]. We present an empirical evaluation and improved analysis of the performance of the EXP3 algorithm in the stochastic setting, as well as a modification of the EXP3 algorithm capable of achieving "logarithmic" regret in stochastic environments.

1. Introduction

Multiarmed bandits are the simplest instance of the exploration-exploitation trade-off problem, which is the basic question in reinforcement learning. There exist two main variants of this problem: stochastic and adversarial. In stochastic multiarmed bandit problems the rewards for playing each arm are generated independently from unknown distributions corresponding to each arm [Auer et al., 2002a]. In adversarial multiarmed bandit problems a sequence of rewards is generated for each arm by an adversary before the game starts [Auer et al., 2002b].

The performance of algorithms for multiarmed bandits is usually evaluated in terms of regret bounds. In the stochastic setting the regret bounds control the difference between the reward obtained by the algorithm and the expected reward that would have been obtained if the algorithm would have played the best arm in all rounds of the game. In the adversarial setting the regret bounds control the difference between the reward obtained by the algorithm and the reward that could have been obtained if the algorithm would have played in all rounds the arm corresponding to the best rewards sequence. In both stochastic and adversarial environments we can talk about in-expectation and high-probability regret bounds.

In-expectation regret bounds are concerned with the expected regret of the algorithm (with respect to its internal randomization and, in the stochastic case, the stochasticity of the environment). High-probability regret bounds provide stronger high-probability guarantees on individual roll-outs of the game. This difference is important to keep in mind when comparing different results.

There exist stochastic and deterministic strategies for stochastic multiarmed bandits. Deterministic strategies are based on computing upper confidence bounds for each of the arms at each round of the game and playing the arm with the highest upper confidence bound. The simplest algorithm from this family is UCB1 algorithm of Auer et al. [2002a], which uses Hoeffding's inequality for computing the upper confidence bounds. Several improvements of this algorithm were proposed, which use tighter concentration inequalities for computing the upper confidence bounds and/or more careful algorithms [Audibert and Bubeck, 2009; Audibert et al., 2009; Auer and Ortner, 2010; Garivier and Cappé, 2011; Maillard et al., 2011].

Stochastic algorithms for stochastic multiarmed bandits include Thompson sampling [Thompson, 1933; Chapelle and Li, 2011; Kaufmann et al., 2012] and EwS algorithm of Maillard [2011]. Stochastic policies have some practical advantages over deterministic UCB-type algorithms. In particular, they are easier to apply in the situation of delayed feedback [Chapelle and Li, 2011].

The most well-known algorithm for adversarial multiarmed bandits is the EXP3.P suggested by Auer et al. [2002b]. In each round of the game EXP3.P picks an arm according to a Gibbs distribution based on upper confidence bounds for each arm, and plays it.

Most analyses of algorithms for stochastic multiarmed bandits provide in-expectation regret bounds. Typically these bounds achieve regret of the form $O(\sum_{\{a:\Delta(a)>0\}} \frac{\ln t}{\Delta(a)})$, termed "logarithmic" regret, where $\Delta(a)$ is the gap between the expected reward of arm a and the expected reward of the "best" arm (the one corresponding to the highest expected reward). In the adversarial case EXP3.P yields $\tilde{O}(\sqrt{KT})$ high-probability regret bound, where the time horizon T is assumed to be known and $\tilde{O}$ hides some logarithmic factors [Auer et al., 2002b]. Unknown time horizons are treated by using the "doubling trick".

There is another algorithm for adversarial multiarmed bandits suggested in Auer et al. [2002b], called EXP3, that picks arms according to a Gibbs distribution based on empirical importance-weighted rewards of the arms instead of using upper confidence bounds. To be more precise, the Gibbs distribution is mixed in a carefully designed proportion with a uniform distribution, which performs the exploration.

We note that there is an important distinction between EXP3 and stochastic algorithms for stochastic multiarmed bandits mentioned earlier (Thompson sampling and EwS): EXP3 is based on importance-weighted sampling, whereas the other two algorithms are based on unweighted rewards. Importance-weighted sampling provides certain advantages when moving to more complex problems, such as stochastic multiarmed bandits with side information [Seldin et al., 2011]. It is also used as a part of the strategy in mixed adversarial and stochastic settings [Bubeck and Slivkins, 2012]. This motivates our study of the performance of EXP3 in stochastic environments.

Seldin et al. [2012] observed that the performance of EXP3 in stochastic environments is comparable to the performance of UCB1 in the "initial phase" of the game (the "initial phase" corresponds to $t < \ln(\Delta)^2/\Delta^4$, where Δ is the minimal positive gap). However,

EXP3 cannot keep up with UCB1 after the "initial phase" and the paper provided only suboptimal high-probability $\tilde{O}(K^{1/3}t^{2/3})$ regret bound for the algorithm. We note that Auer et al. [2002b] provided a $\tilde{O}(\sqrt{KT})$ in-expectation regret bound for EXP3 in the adversarial setting, but no high-probability statement could be derived. The question of whether a high-probability $\tilde{O}(\sqrt{Kt})$ regret bound can be derived in the stochastic setting remained open. Although there is a high-probability $\tilde{O}(\sqrt{KT})$ regret bound for the EXP3.P algorithm in the adversarial setting (which directly implies that the same bound holds in the stochastic setting), Seldin et al. [2012] showed that in practice in the stochastic setting EXP3.P is significantly inferior to EXP3, which provided the motivation for a deeper study of EXP3.

1.1. Summary of the Main Contributions

In Theorem 1 we provide a $\tilde{O}(\sqrt{KT})$ high-probability regret bound for application of the EXP3 algorithm in stochastic environments. We note that we analyze EXP3 algorithm with time-varying learning rate, which is a more elegant approach than the doubling trick.

Theorem 1 is based on Lemma 2, which shows that in stochastic environments the empirical importance-weighted rewards of all arms are "well-behaved" with high probability. We note that in adversarial environments the empirical importance-weighted rewards are not "well-behaved" and additional tools are required for their control, such as EXP3.P.

We also propose and analyze a modification of the EXP3 algorithm that we name EXP3ELM, which is based on a combination of EXP3 with action elimination (see Algorithm 2 box). EXP3ELM performs identically to EXP3 in the "initial phase" of the game and comparably to UCB1 after the "initial phase". In Theorem 3 we prove a "logarithmic" regret bound for EXP3ELM.

We also provide an empirical evaluation of the performance of EXP3 and EXP3ELM.

2. Problem Setting and Definitions

Let $\mathcal{A}$ be a set of K actions (arms) and let $a \in \mathcal{A}$ denote the actions. Denote by $R(a)$ the expected reward of action a and let $p(\cdot|a)$ be the unknown reward distribution underlying a. We assume that the support of $p(\cdot|a)$ is contained in the [0,1] interval. Let π_t be a distribution over $\mathcal{A}$ that is played at round t of the game (a policy). Let $\{A_1, A_2, \dots\}$ be the sequence of actions played, such that A_t is distributed according to π_t and π_t is computed based on the history of past observations, $\mathcal{H}_{t-1} \equiv (A_1, \dots, A_{t-1}, R_1, \dots, R_{t-1})$, where $R_1, R_2, \dots$ is the sequence of observed rewards, so that R_t is distributed according to $p(\cdot|A_t)$.

For $t \geq 1$ and $a \in \{1, \dots, K\}$ define a set of random variables R_t^a (the importance weighted rewards) and their cumulative sum $\hat{R}_t(a)$ as:

$$R_t^a \equiv \frac{R_t}{\pi_t(a)}\mathbb{I}_{\{A_t=a\}} = \begin{cases} \frac{1}{\pi_t(a)}R_t, & \text{if } A_t = a; \\ 0, & \text{otherwise;} \end{cases} \qquad \hat{R}_t(a) \equiv \sum_{\tau=1}^{t} R_\tau^a.$$

Note that $\mathbb{E}[R_t^a|\mathcal{H}_{t-1}] = R(a)$ and $\mathbb{E}[\hat{R}_t(a)] = tR(a)$.

Let $a^* \equiv \arg\max_a R(a)$ be the "best" action in the game (if there are multiple "best" actions, pick any of them arbitrarily). We define the regret and empirical regret of an action

$t = 0; \quad \varepsilon_0 = \frac{1}{K}; \quad \forall a\colon \hat{R}_0(a) = 0.$
for $t = 1, 2, \ldots$ **do**
  $\varepsilon_t = \min\left\{\frac{1}{K}, \sqrt{\frac{\ln K}{Kt}}\right\}.$
  $\forall a\colon \tilde{\rho}_t(a) = (1 - K\varepsilon_t)\frac{e^{\varepsilon_{t-1}\hat{R}_{t-1}(a)}}{\sum_{a'} e^{\varepsilon_{t-1}\hat{R}_{t-1}(a')}} + \varepsilon_t.$
  Draw A_t according to $\tilde{\rho}_t$ and play it.
  Observe reward R_t.
  $\forall a\colon \hat{R}_t(a) = \hat{R}_{t-1}(a) + \frac{R_t}{\tilde{\rho}_t(a)}\mathbb{I}_{\{a=A_t\}}.$
end

Algorithm 1: EXP3.

Input: Confidence parameter δ.
Initialization: $t = 0; \quad \varepsilon_0 = \frac{1}{K}; \quad \mathcal{A}_0 = \mathcal{A};$
$B = 4(e-2)\left(2\ln K + \ln\frac{2}{\delta}\right); \quad \forall a\colon \hat{R}_0(a) = 0;$
for $t = 1, 2, \ldots$ **do**
  $\varepsilon_t = \min\left\{\frac{1}{K}, \sqrt{\frac{\ln K}{Kt}}\right\}.$
  $\forall a \in \mathcal{A}_t$:
  $\tilde{\rho}_t(a) = (1 - |\mathcal{A}_t|\varepsilon_t)\frac{e^{\varepsilon_{t-1}\hat{R}_{t-1}(a)}}{\sum_{a' \in \mathcal{A}_t} e^{\varepsilon_{t-1}\hat{R}_{t-1}(a')}} + \varepsilon_t.$
  Draw $A_t \in \mathcal{A}_t$ according to $\tilde{\rho}_t$ and play it.
  Observe reward R_t.
  $\forall a\colon \hat{R}_t(a) = \hat{R}_{t-1}(a) + \frac{R_t}{\tilde{\rho}_t(a)}\mathbb{I}_{\{a=A_t\}}.$
  $\forall a \in \mathcal{A}_t\colon V_{R_t}(a) = V_{R_{t-1}}(a) + \frac{1}{\tilde{\rho}_t(a)}.$
  $\mathcal{A}_t = \mathcal{A}_{t-1} \setminus$
  $\left\{a : \hat{R}_t^{max} - \hat{R}_t(a) > \sqrt{B(V_{R_t}(\hat{a}_t^*) + V_{R_t}(a))}\right\}.$
end

Algorithm 2: EXP3ELM.

a by:

$$\Delta(a) \equiv R(a^*) - R(a), \qquad \hat{\Delta}_t(a) \equiv \hat{R}_t(a^*) - \hat{R}_t(a).$$

Note that $\hat{R}_t(a) - tR(a)$ is a martingale. The variance of this martingale satisfies:

$$\sum_{\tau=1}^{t} \mathbb{E}\left[(R_\tau^a - R(a))^2 \middle| \mathcal{H}_{\tau-1}\right] = \sum_{\tau=1}^{t} \mathbb{E}\left[(R_\tau^a)^2 \middle| \mathcal{H}_{\tau-1}\right] - tR(a)^2 \le \sum_{\tau=1}^{t} \pi_t(a)\frac{(R_t)^2}{\pi_t(a)^2} = \sum_{\tau=1}^{t} \frac{1}{\pi_\tau(a)},$$

where we used the fact that $R_t \le 1$. Let $V_{R_t}(a) \equiv \sum_{\tau=1}^{t} \frac{1}{\pi_\tau(a)}$ be the upper bound on the variance of the martingale $\hat{R}_t(a) - tR(a)$.

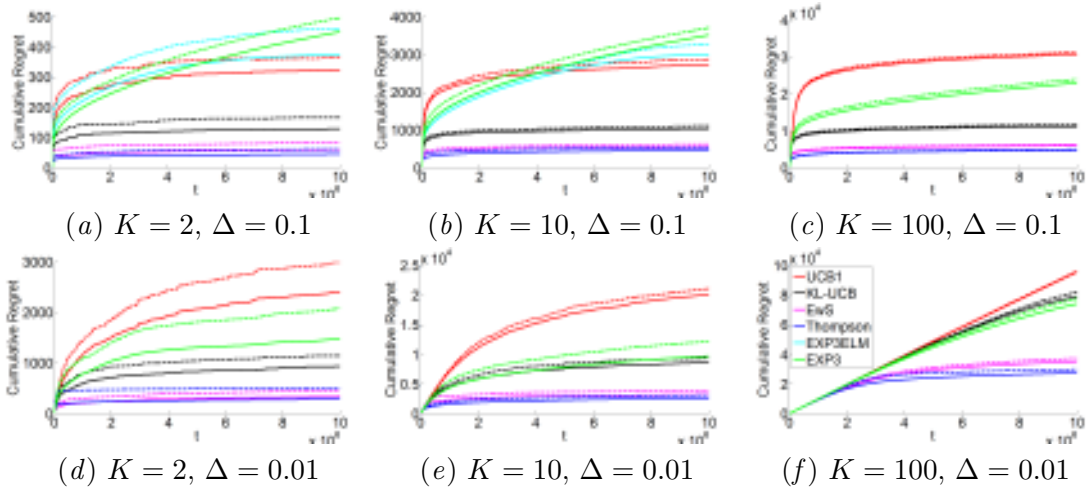
Finally, define the highest empirical reward at round t and the empirically best arm at round t as:

$$\hat{R}_t^{max} \equiv \max_a \hat{R}_t(a), \qquad \hat{a}_t^* \equiv \arg\max_a \hat{R}_t(a).$$

3. Algorithms

In this section we present two learning algorithms that we are using in this paper. The first algorithm is a minor modification of the EXP3 algorithm of Auer et al. [2002b]. The difference is that the learning rate ε_t, which is equal to the exploration rate, is changing with time.

The EXP3 algorithm keeps sampling "bad" actions at the rate of $\varepsilon_t = \sqrt{\frac{\ln K}{Kt}}$ and, therefore, its regret is $\Omega(\sqrt{Kt})$ and it cannot compete with UCB1 after the "initial phase" of the game. This shortcoming is fixed in the EXP3ELM algorithm in the Algorithm 2 box. EXP3ELM maintains a set $\mathcal{A}_t$ of "active" actions. Initially, $\mathcal{A}_0 = \mathcal{A}$ and the actions are withdrawn from the active set as soon as it becomes evident with high probability that they are suboptimal.



(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 4.1: **Comparison of EXP3, EXP3ELM, UCB1, KL-UCB, EwS, and Thompson sampling.** The legend in figure (f) corresponds to all the figures. Solid lines represent the mean performance over the experiments and dashed lines represent the mean plus one standard deviation over the ten repetitions of the corresponding experiment. In figures (c) - (f) the number of rounds was insufficient for arms elimination to take place, therefore, the behavior of EXP3ELM is identical to EXP3 and EXP3ELM is omitted from the figures. Enlarged version of the figure is provided in the supplementary material.

4. Experimental Results

Before we dive into the analysis of the algorithms we present several experimental results. We consider 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.

We compare EXP3 and EXP3ELM algorithms presented in the previous section with UCB1 [Auer et al., 2002a], KL-UCB [Garivier and Cappé, 2011; Maillard et al., 2011] (we use the version of the algorithm suggested in Garivier and Cappé [2011] with constant $c = 3$), EwS [Maillard, 2011], and Thompson sampling [Thompson, 1933; Kaufmann et al., 2012]. From the experiments we see that EXP3 is superior to UCB1 in the "initial phase" of the game and EXP3ELM is identical to EXP3 in the "initial phase" and comparable to UCB1 after the "initial phase". The KL-UCB algorithm is a much stronger competitor. Nevertheless, for the hardest "needle in a haystack" problems ($\Delta = 0.01$ and $K = 10$ and 100) EXP3 performs comparably and, in the hardest case, even slightly better. EwS and Thompson sampling are clear leaders, however, we remind that EXP3 is based on importance-weighted sampling and its study in the stochastic setting is of independent interest.

5. Analysis

In this section we present an analysis of the EXP3 and EXP3ELM algorithms. For the rigorous analysis we consider a slightly modified version of the algorithms that we name EXP3C and EXP3ELMC. The modification reduces the learning rate by a factor of $1/\sqrt{C \ln K}$ (namely, we take $\varepsilon_t = \min\left\{1/\sqrt{CKt},\ 1/(2K)\right\}$), where C is defined in Theorem 1. We point out that C depends on the confidence parameter δ, but not on the time horizon. In practice, the reduction of the learning rate reduces the performance. The elimination rule in the modified EXP3ELMC algorithm uses a slightly weaker deterministic upper bound on the variance of the martingales:

$$\mathcal{A}_t = \mathcal{A}_{t-1} \setminus \left\{a : \hat{R}_t^{max} - \hat{R}_t(a) > 2\sqrt{CKt}\right\}. \tag{1}$$

We let $\hat{R}_t(EXP3C) \equiv \sum_{\tau=1}^{t} R_\tau$ denote the reward of EXP3C after t rounds. Similarly, $\hat{R}_t(EXP3ELMC)$ is the cumulative reward of EXP3ELMC after t rounds.

Theorem 1 *Let $\delta \in (0,1)$ and define $C \equiv 4(e-2)\left(\ln K + \ln \frac{4}{\delta}\right) e^2$. Then with probability greater than $1-\delta$ the regret of EXP3C is bounded as*:

$$tR(a^*) - \hat{R}_t(EXP3C) \le (1+\ln K)\sqrt{CKt} + 2(3e-4)\sqrt{\frac{Kt}{C}} + (e-2)\sqrt{\frac{1}{2}t\ln\frac{2}{\delta}} + 2K. \tag{2}$$

The key element in the proof of Theorem 1 are inequalities (3) and (4) in the following lemma, which show that the empirical rewards of all actions in EXP3 are "well-behaved" and that the empirical reward of the best action a^* always stays "at the top". (The proofs are provided at the end of this section.)

Lemma 2 *In the EXP3C algorithm with probability greater than $1-\frac{\delta}{2}$, for all t*:

$$\hat{R}_t^{max} \le tR(a^*) + \sqrt{CKt}, \tag{3}$$

$$\hat{R}_t(a^*) \ge tR(a^*) - \sqrt{CKt}, \tag{4}$$

$$V_{R_t}(a^*) \le C'Kt, \tag{5}$$

$$\frac{1}{\tilde{\rho}_t(a^*)} \le C'K, \tag{6}$$

$$\text{If } \hat{R}_{t-1}(a) \ge (t-1)R(a^*) - \sqrt{CK(t-1)},\ \text{ then } \frac{1}{\tilde{\rho}_t(a)} \le C'K, \tag{7}$$

where $C' = e^2$ and C is defined in Theorem 1.

It is important to note that in the EXP3.P algorithm for the adversarial case the control over empirical importance-weighted rewards is achieved by adding the variance of the rewards to the exponent of the Gibbs distribution in $\tilde{\rho}_t$. As it follows from Lemma 2, in the stochastic case the empirical importance-weighted rewards are controlled directly.

Finally, we show that EXP3ELM achieves "logarithmic" regret ($\ln t$ term does not appear explicitly in the bound, but it is replaced by $\ln(1/\delta)$ in the definition of C). The proof of this theorem is provided in the supplementary material.

Theorem 3 *Let $\delta \in (0,1)$ and C as defined in Theorem 1, let RHS(2) be the right hand side of (2), then with probability greater than $1-\delta$ the regret of EXP3ELMC is bounded as*:

$$tR(a^*) - \hat{R}_t(EXP3ELMC) \leq \min\left\{RHS(2),\ 4CK\sum_a \frac{1}{\Delta(a)}\right\}.$$

5.1. Proofs

Let $\rho_t(a) \equiv \frac{e^{\varepsilon_{t-1}\hat{R}_{t-1}(a)}}{Z(\rho_t)}$ denote the non-smoothed version of $\tilde{\rho}_t$, where $Z(\rho_t) \equiv \sum_a e^{\varepsilon_{t-1}\hat{R}_{t-1}(a)}$ is the normalization factor. Also, let $\rho_t^{\varepsilon}(a) \equiv \frac{e^{\varepsilon\hat{R}_{t-1}(a)}}{Z(\rho_t^{\varepsilon})}$ and $Z(\rho_t^{\varepsilon}) \equiv \sum_a e^{\varepsilon\hat{R}_{t-1}(a)}$ denote the corresponding quantities, where instead of the "native" value ε_{t-1} we use a different ε.

We use the following two results from prior work. The first result from Auer et al. [2002b] relates the rewards R_t to the logarithm of the fraction of normalization coefficients. Note that in the definition of ρ_t we have ε_{t-1}, so $Z(\rho_t^{\varepsilon_{t-2}})$ corresponds to ρ_t with ε from the definition of ρ_{t-1}.

Lemma 4 ([Auer et al., 2002b])

$$R_t \geq (1-K\varepsilon_t)\frac{1}{\varepsilon_t}\ln\frac{Z(\rho_t^{\varepsilon_{t-2}})}{Z(\rho_{t-1})} - (e-2)\varepsilon_t\frac{1}{\tilde{\rho}_t(A_t)} \tag{8}$$

The second result from [Maillard, 2011, Section 3.2] bounds the sum of logarithms of the fractions of normalization coefficients.

Lemma 5 ([Maillard, 2011])

$$\sum_{\tau=1}^{t}\frac{1}{\varepsilon_\tau}\ln\frac{Z(\rho_\tau^{\varepsilon_{\tau-2}})}{Z(\rho_{\tau-1})} \geq \hat{R}_t(a^*) - \frac{\ln K}{\varepsilon_t}.$$

We also prove the following lemma that takes care of the last term in (8). Lemma 5 together with Lemma 6 allow us to work with a varying learning rate.

Lemma 6 *With probability greater than $1-\frac{\delta}{2}$, for all t:*

$$\sum_{\tau=1}^{t}\varepsilon_\tau\frac{1}{\tilde{\rho}_\tau(A_\tau)} \leq 2\sqrt{\frac{Kt}{C}} + \sqrt{\frac{1}{2}t\ln\frac{2}{\delta}}.$$

Proof of Lemma 6 We have $\mathbb{E}\left[\frac{1}{\tilde{\rho}_t(A_t)}\middle|\mathcal{H}_{t-1}\right] = \sum_a \tilde{\rho}_t(a)\frac{1}{\tilde{\rho}_t(a)} = K$, which implies that $\sum_{\tau=1}^{t}\varepsilon_\tau\left(\frac{1}{\tilde{\rho}_\tau(A_\tau)} - K\right)$ is a martingale. Since $\varepsilon_t\frac{1}{\tilde{\rho}_t(A_t)} \in (0,1]$, we have by Hoeffding-Azuma's inequality, with probability greater than $1-\frac{\delta}{2}$:

$$\sum_{\tau=1}^{t}\varepsilon_\tau\frac{1}{\tilde{\rho}_\tau(A_\tau)} \leq K\sum_{\tau=1}^{t}\varepsilon_\tau + \sqrt{\frac{1}{2}t\ln\frac{2}{\delta}} \leq 2\sqrt{\frac{Kt}{C}} + \sqrt{\frac{1}{2}t\ln\frac{2}{\delta}}.$$

By martingale stopping argument, similar to the one used in the proof of Bernstein's inequality (see supplementary material), the statement holds for all t simultaneously. ■

Now we are ready to prove Theorem 1. (The proof of Lemma 2 is provided below.)

Proof of Theorem 1 Summing over t the two sides of Lemma 4 and applying Lemmas 5, 2, and 6, we obtain with probability greater than $1-\delta$:

$$\begin{aligned}
&\sum_{\tau=1}^{t} R_t \geq \sum_{\tau=1}^{t}(1-K\varepsilon_\tau)\frac{1}{\varepsilon_\tau}\ln\frac{Z(\rho_\tau^{\varepsilon_\tau-2})}{Z(\rho_{\tau-1})} - (e-2)\sum_{\tau=1}^{t}\varepsilon_\tau\frac{1}{\tilde{\rho}_\tau(A_\tau)} \\
&= \sum_{\tau=1}^{t}\frac{1}{\varepsilon_\tau}\ln\frac{Z(\rho_\tau^{\varepsilon_\tau-2})}{Z(\rho_{\tau-1})} - K\sum_{\tau=1}^{t}\varepsilon_\tau\frac{1}{\varepsilon_\tau}\ln\frac{Z(\rho_\tau^{\varepsilon_\tau-2})}{Z(\rho_{\tau-1})} - (e-2)\sum_{\tau=1}^{t}\varepsilon_\tau\frac{1}{\tilde{\rho}_\tau(A_\tau)} \\
&\geq \hat{R}_t(a^*) - \frac{\ln K}{\varepsilon_t} - 2(e-1)K\sum_{\tau=1}^{t}\varepsilon_\tau - 2(e-2)\sqrt{\frac{Kt}{C}} - (e-2)\sqrt{\frac{1}{2}t\ln\frac{2}{\delta}} \qquad (9) \\
&\geq tR(a^*) - \sqrt{CKt} - \sqrt{CKt}\ln K - 4(e-1)\sqrt{\frac{Kt}{C}} - 2(e-2)\sqrt{\frac{Kt}{C}} - (e-2)\sqrt{\frac{1}{2}t\ln\frac{2}{\delta}} - 2K \qquad (10) \\
&= tR(a^*) - (1+\ln K)\sqrt{CKt} - 2(3e-4)\sqrt{\frac{Kt}{C}} - (e-2)\sqrt{\frac{1}{2}t\ln\frac{2}{\delta}} - 2K,
\end{aligned}$$

where in (9) we applied Lemmas 5 and 6 and used the following upper bound on $\frac{1}{\varepsilon_t}\ln\frac{Z(\rho_t^{\varepsilon_t-2})}{Z(\rho_{t-1})}$, which follows from Lemma 4 (by definition $\varepsilon_t \leq \frac{1}{2}$)

$$\frac{1}{\varepsilon_t}\ln\frac{Z(\rho_t^{\varepsilon_t-2})}{Z(\rho_{t-1})} \leq \frac{1}{1-K\varepsilon_t}\left(R_t + (e-2)\varepsilon_t\frac{1}{\tilde{\rho}_t(A_t)}\right) \leq 2(1+(e-2)) = 2(e-1),$$

and in (10) we applied Lemma 2. (By the union bound, Lemmas 2 and 6 hold simultaneously with probability greater than $1-\delta$.) ■

Proof of Lemma 2 We prove the lemma by induction. For $t=1$ all the claims of the lemma hold. We assume that the claims hold for $t-1$ and show that this implies that they also hold for t.

We start with the proof of (6). If $\rho_t(a^*) \geq \frac{1}{K}$ we are done. Otherwise:

$$\begin{aligned}
\frac{1}{\tilde{\rho}_t(a^*)} \leq \frac{1}{\rho_t(a^*)} &= \frac{\sum_{a'} e^{\varepsilon_{t-1}\hat{R}_{t-1}(a')}}{e^{\varepsilon_{t-1}\hat{R}_{t-1}(a^*)}} \\
&\leq \sum_{a'} e^{\varepsilon_{t-1}\left(\hat{R}_{t-1}^{max} - (t-1)R(a^*) + \sqrt{CK(t-1)}\right)} \leq Ke^{2\varepsilon_{t-1}\sqrt{CK(t-1)}} = C'K,
\end{aligned}$$

where we used the fact that $\hat{R}_{t-1}(a') \leq \hat{R}_{t-1}^{max}$ for all a' and the induction assumption $\hat{R}_t(a^*) \geq tR(a^*) - \sqrt{CKt}$.

Inequality (5) follows from (6) and inequality (4) follows from (5) by Bernstein's inequality (see supplementary material).

Now we prove (7). We note that the proof of (6) was based on the induction assumption (4) for $t-1$. In (7) we assumed that $\hat{R}_{t-1}(a) \geq (t-1)R(a^*) - \sqrt{CK(t-1)}$ and using this assumption the proof is identical.

It is now left to prove (3). If $\hat{R}_{t-1}(a) < (t-1)R(a^*) - \sqrt{CK(t-1)}$ then

$$\hat{R}_t(a) = \hat{R}_{t-1}(a) + R_t^a \leq (t-1)R(a^*) - \sqrt{CK(t-1)} + \frac{1}{\varepsilon_t} \leq tR(a^*) + \sqrt{CKt}.$$

Otherwise, we are in the case $\hat{R}_{t-1}(a) \geq (t-1)R(a^*) - \sqrt{CK(t-1)}$. Let

$$\ell(t) \equiv \max_{\tau} \left\{ \tau : \begin{array}{c} \tau<t \text{ and} \\ \hat{R}_{\tau-1}(a)<(\tau-1)R(a^*)-\sqrt{CK(\tau-1)} \\ \text{and } \hat{R}_{\tau}(a)\geq\tau R(a^*)-\sqrt{CK\tau} \end{array} \right\}.$$

$\ell(t)$ is the last time before t when $\hat{R}_\tau(a)$ crosses the $\tau R(a^*) - \sqrt{CK\tau}$ line from below. If $\hat{R}_\tau(a)$ always stays above this line, define $\ell(t) \equiv 1$.

Let $\hat{R}_{t_1}^{t_2}(a) \equiv \sum_{\tau=t_1+1}^{t_2} R_\tau^a$ be the sub-sum of the rewards of arm a from time t_1+1 to time t_2. By our assumption and the definition of $\ell(t)$ we have $\hat{R}_\tau(a) \geq (t-1)R(a^*) - \sqrt{CK(t-1)}$ for all $\ell(t) \leq \tau < t$. Furthermore, $\hat{R}_{\ell(t)}^t - (t-\ell(t))R(a)$ is a martingale and, by (7), the variance of this martingale satisfies

$$\sum_{\tau=\ell(t)+1}^{t} \mathbb{E}\left[(\hat{R}_\tau^a - R(a))^2 | \mathcal{H}_{\tau-1}\right] \leq \sum_{\tau=\ell(t)+1}^{t} \frac{1}{\tilde{\rho}_\tau(a)} \leq C'K(t-\ell(t)).$$

Therefore, by Bernstein's inequality:

$$\hat{R}_{\ell(t)}^t \leq (t-\ell(t))R(a) + \sqrt{CK(t-\ell(t))} \leq (t-\ell(t))R(a^*) + \sqrt{CKt}.$$

As well,

$$\hat{R}_{\ell(t)}(a) = \hat{R}_{\ell(t)-1}(a) + \frac{1}{\varepsilon_{\ell(t)}} \leq (\ell(t)-1)R(a^*) - \sqrt{CK(\ell(t)-1)} + \frac{1}{\varepsilon_{\ell(t)}} \leq \ell(t)R(a^*).$$

Putting it together we obtain:

$$\hat{R}_t(a) = \hat{R}_{\ell(t)}(a) + \hat{R}_{\ell(t)}^t(a) \leq \ell(t)R(a^*) + (t-\ell(t))R(a^*) + \sqrt{CKt} = tR(a^*) + \sqrt{CKt}.$$

■

5.2. Comparison with Other Regret Bounds

We stress out that our results in Theorems 1 and 3 and Lemma 2 are high-probability results, as opposed to in-expectation analysis that is provided for most other algorithms for stochastic multiarmed bandits, so the comparison is not completely straightforward. In Theorem 3 we obtain an $O((\sqrt{Kt})\ln K)$ regret in the "initial phase" of the game and $O(K \ln K \sum_a \frac{1}{\Delta(a)})$ after the "initial phase" of the game. The bound for the "initial phase" can be compared with $O(\sqrt{KT}\frac{\ln(K \ln K)}{\sqrt{\ln K}})$ in-expectation regret bound for an improved version of UCB [Auer

and Ortner, 2010]. We lose slightly less than a factor of $\sqrt{\ln K}$, however, we achieve a stronger high-probability statement. We also note that the $\sqrt{\ln K}$ factor can be slightly reduced by tuning C. The regret bound after the "initial phase" can be compared with the $O(\ln t \sum_a \frac{1}{\Delta(a)})$ in-expectation regret bound for the UCB1 algorithm [Auer et al., 2002a]. We do not have the $\ln t$ factor since we eliminate arms, but this factor will come back if we replace the elimination by more intelligent exploration schedule. On the other hand, we have $K \ln K$ factor in our bound, but, as in the case with the "initial phase", we provide a stronger high-probability statement.

It is also interesting to compare our result with the analysis of EXP3 and EXP3.P algorithms for adversarial multiarmed bandits in Auer et al. [2002b]. In Theorem 1 we provide a high-probability $O(\ln K \sqrt{Kt})$ regret bound for EXP3 in the stochastic setting. This can be compared with $O(\sqrt{KT \ln K})$ in-expectation regret bound for EXP3 and $O(\sqrt{KT \ln(KT)})$ high-probability regret bound for EXP3.P in the adversarial setting. (We note that Auer et al. [2002b] assume that the time horizon T is known and tune the learning rate based on this knowledge. A similar in-expectation regret bound for application of EXP3 with time-varying learning rate to adversarial bandits is provided by Maillard [2011].) The main message of Theorem 1 and Lemma 2 is that in the stochastic setting we can apply Gibbs sampling based on the empirical rewards instead of Gibbs sampling based on upper confidence bounds used by EXP3.P algorithm for adversarial environments.

6. Discussion

We presented a high-probability analysis of the EXP3 algorithm in stochastic environments and designed a modification of EXP3 that performs identically to EXP3 in the "initial phase" of the game and comparably to UCB1 after the "initial phase". The main distinction of the proposed algorithm from existing algorithms for stochastic multiarmed bandits is that it is based on importance-weighted sampling. We provided a comprehensive study of the behavior of importance-weighted sampling in stochastic environments.

Acknowledgments

We would like to thank Odalric-Ambrym Maillard for helpful discussions and useful references. This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886, and by the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement N^o270327. This publication only reflects the authors' views.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

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

Jean Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 2009.

Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

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 multiarmed bandit problem. *SIAM Journal of Computing*, 32(1), 2002b.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

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 Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

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

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2011.

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

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

Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2011.

Yevgeny Seldin, Peter Auer, François Laviolette, John Shawe-Taylor, and Ronald Ortner. PAC-Bayesian analysis of contextual bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Yevgeny Seldin, Nicolò Cesa-Bianchi, Peter Auer, François Laviolette, and John Shawe-Taylor. PAC-Bayes-Bernstein inequality for martingales and its application to multiarmed bandits. *JMLR Workshop and Conference Proceedings*, 26, 2012.

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

Supplementary Material

Appendix A. Proof of Theorem 3

For the proof of Theorem 3 we need one small lemma.

Lemma 7

$$\text{if } \hat{R}_{t-1}(a) \geq \hat{R}_{t-1}^{max} - 2\sqrt{CK(t-1)}, \text{ then } \frac{1}{\tilde{\rho}_t(a)} \leq C'K \tag{11}$$

and for all actions in the active set $\mathcal{A}_t$ at time t, we have $V_{R_t}(a) \leq C'Kt$.

Proof of Lemma 7 We start with the proof of (11). If $\rho_t(a) \geq \frac{1}{K}$ we are done. Otherwise,

$$\begin{aligned}\frac{1}{\tilde{\rho}_t(a)} \leq \frac{1}{\rho_t(a)} &= \frac{\sum_{a'} e^{\varepsilon_{t-1}\hat{R}_{t-1}(a')}}{e^{\varepsilon_{t-1}\hat{R}_t(a)}} \\ &\leq \sum_{a'} e^{\varepsilon_{t-1}\left(\hat{R}_{t-1}^{max} - \hat{R}_{t-1}^{max} + 2\sqrt{CK(t-1)}\right)} = \sum_{a'} e^{2\varepsilon_{t-1}\sqrt{CK(t-1)}} = C'K,\end{aligned}$$

where we used that fact that $\hat{R}_{t-1}(a') \leq \hat{R}_{t-1}^{max}$ for all a' and the assumption $\hat{R}_{t-1}(a) \geq \hat{R}_{t-1}^{max} - 2\sqrt{CK(t-1)}$. Finally, as long as an action a is in the active set, the precondition of (11) holds and, therefore, we obtain $V_{R_t}(a) = \sum_{\tau=1}^{t} \frac{1}{\tilde{\rho}_\tau(a)} \leq C'Kt$. ■

Proof of Theorem 3 By inequalities (3) and (4) and the definition of the elimination rule, with probability greater than $1-\delta$ the best action a^* is never eliminated by EXP3ELMC.

Now we bound the number of times that suboptimal actions are played. As long as action a is active, by Bernstein's inequality and the bound on $V_{R_t}(a)$ in Lemma 7, with probability greater than $1-\delta$ we have $\hat{R}_t(a) \leq tR(a) + \sqrt{CKt}$. On the other hand, $\hat{R}_t^{max} \geq \hat{R}_t(a^*) \geq tR(a^*) - \sqrt{CKt}$. Thus, with probability greater than $1-\delta$:

$$\hat{R}_t^{max} - \hat{R}_t(a) \geq t\Delta(a) - 2\sqrt{CKt}. \tag{12}$$

By the elimination rule (1), an action a is eliminated, at the latest, when the right hand side of (12) is greater than $2\sqrt{CKt}$, which means that with probability greater than $1-\delta$ each suboptimal action is eliminated after at most $\frac{4CK}{\Delta(a)^2}$ rounds. The cumulative regret for playing action a with regret $\Delta(a)$ over $\frac{4CK}{\Delta(a)^2}$ rounds is $\frac{4CK}{\Delta(a)}$. Summing over the actions we obtain the result of the theorem. ■

Appendix B. Bernstein's Inequality

We used the following form of Bernstein's inequality, which is based on a fairly standard martingale stopping argument [Freedman, 1975; Abbasi-Yadkori et al., 2011].

Theorem 8 (Bernstein's Inequality) *Let $X_1, X_2, \ldots$ be a martingale difference sequence, such that $|X_t| \leq \alpha_t$ for an increasing deterministic sequence $\alpha_1, \alpha_2, \ldots$ with probability 1. Let $M_t \equiv \sum_{\tau=1}^{t} X_\tau$ be martingale. Let $\bar{V}_1, \bar{V}_2, \ldots$ be a sequence of deterministic upper bounds on the variance $V_t \equiv \sum_{\tau=1}^{t} \mathbb{E}[X_\tau^2 | X_1, \ldots, X_{\tau-1}]$ of the martingale M_t, such that $\bar{V}_t$-s satisfy $\sqrt{\frac{\ln \frac{2}{\delta}}{(e-2)\bar{V}_t}} \leq \frac{1}{\alpha_t}$. Then with probability greater than $1-\delta$ for all t:*

$$|M_t| \leq 2\sqrt{(e-2)\bar{V}_t \ln \frac{2}{\delta}}.$$

The theorem is based on the following lemma [Freedman, 1975; Beygelzimer et al., 2011].

Lemma 9 *For M_t and V_t defined in Theorem 8 and $\lambda_t \in [0, \frac{1}{\alpha_t}]$:*

$$\mathbb{E}\left[e^{\lambda_t M_t - (e-2)\lambda_t^2 V_t}\right] \leq 1.$$

Proof of Theorem 8. Define the stopping time $t^\star$:

$$t^\star \equiv \min\left\{t : |M_t| > 2\sqrt{(e-2)\bar{V}_t \ln \frac{2}{\delta}}\right\}.$$

Define the stopped martingale difference sequence by $X_t^{t^\star} \equiv X_t \mathbb{I}_{t \leq t^\star}$. We have $|X_t^{t^\star}| \leq \alpha_t$ for all t. Clearly, the stopped martingale difference sequence is also a martingale difference sequence. Define by $M_t^{t^\star} \equiv \sum_{\tau=1}^{t} X_t^{t^\star}$ the stopped martingale and by $V_t^{t^\star}$ its conditional variance process. Clearly, $V_t^{t^\star} \leq \bar{V}_{t^\star}^{t^\star}$ for all t. Hence, we have:

$$\mathbb{E}\left[e^{\lambda_{t^\star} M_t^{t^\star} - (e-2)\lambda_{t^\star}^2 \bar{V}_{t^\star}}\right] \leq \mathbb{E}\left[e^{\lambda_{t^\star} M_t^{t^\star} - (e-2)\lambda_{t^\star}^2 V_t^{t^\star}}\right] \leq 1.$$

From here, by Markov's inequality, with probability greater than $1 - \frac{\delta}{2}$:

$$\lambda_{t^\star} M_t^{t^\star} - (e-2)\lambda_{t^\star}^2 \bar{V}_{t^\star} \leq \ln \frac{2}{\delta} + \ln \mathbb{E}\left[e^{\lambda_{t^\star} M_t^{t^\star} - (e-2)\lambda_{t^\star}^2 \bar{V}_{t^\star}}\right] \leq \ln \frac{2}{\delta}.$$

By applying the same argument to the negative martingale $-M_t$ and taking $\lambda_{t^\star} \equiv \sqrt{\frac{\ln \frac{2}{\delta}}{(e-2)\bar{V}_{t^\star}}}$ we obtain that with probability greater than $1-\delta$:

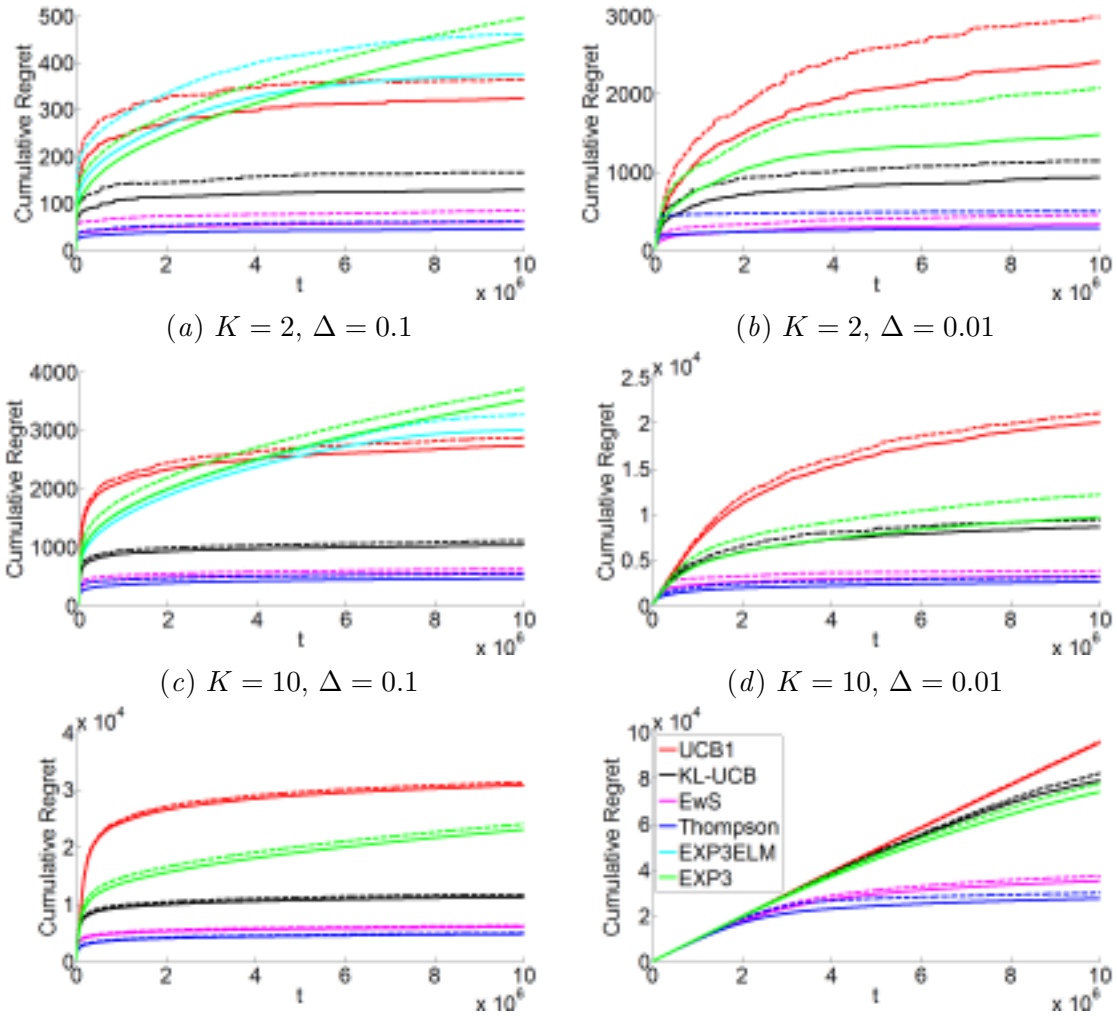
$$|M_t^{t^\star}| \leq \frac{\ln \frac{2}{\delta}}{\lambda_{t^\star}} + (e-2)\lambda_{t^\star}^2 \bar{V}_{t^\star} = 2\sqrt{(e-2)\bar{V}_{t^\star} \ln \frac{2}{\delta}}$$

Finally, we have:

$$\mathbb{P}\left\{\exists t : |M_t| > 2\sqrt{(e-2)\bar{V}_t \ln \frac{2}{\delta}}\right\} = \mathbb{P}\left\{t^\star < \infty\right\} = \mathbb{P}\left\{|M_{t^\star}^{t^\star}| > 2\sqrt{(e-2)\bar{V}_{t^\star} \ln \frac{2}{\delta}},\ t^\star < \infty\right\}$$
$$\leq \mathbb{P}\left\{|M_{t^\star}^{t^\star}| > 2\sqrt{(e-2)\bar{V}_{t^\star} \ln \frac{2}{\delta}}\right\} \leq \delta$$

■

Appendix C. Enlarged Version of Figure 4.1



(a) $K = 2$, $\Delta = 0.1$ (b) $K = 2$, $\Delta = 0.01$

(c) $K = 10$, $\Delta = 0.1$ (d) $K = 10$, $\Delta = 0.01$

(e) $K = 100$, $\Delta = 0.1$ (f) $K = 100$, $\Delta = 0.01$

Figure C.1: **Enlarged version of Figure 4.1.** The columns in this figure correspond to the rows in Figure 4.1. Please, see the caption of Figure 4.1 for explanations.