PAC-Bayes-Bernstein Inequality for Martingales and its Application to Multiarmed Bandits

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

Nicolò Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Italy

Peter Auer AUER@UNILEOBEN.AC.AT
Chair for Information Technology, University of Leoben, Austria

François Laviolette FRANCOIS.LAVIOLETTE@IFT.ULAVAL.CA
Université Laval, Québec, Canada

John Shawe-Taylor JST@CS.UCL.AC.UK
University College London, UK

Editor: Editor's name

Abstract

We develop a new tool for data-dependent analysis of the exploration-exploitation trade-off in learning under limited feedback. Our tool is based on two main ingredients. The first ingredient is a new concentration inequality that makes it possible to control the concentration of weighted averages of multiple (possibly uncountably many) simultaneously evolving and interdependent martingales.[1] The second ingredient is an application of this inequality to the exploration-exploitation trade-off via importance weighted sampling. We apply the new tool to the stochastic multiarmed bandit problem, however, the main importance of this paper is the development and understanding of the new tool rather than improvement of existing algorithms for stochastic multiarmed bandits. In the follow-up work we demonstrate that the new tool can improve over state-of-the-art in structurally richer problems, such as stochastic multiarmed bandits with side information (Seldin et al., 2011a).

Keywords: PAC-Bayesian Analysis, Bernstein's Inequality, Martingales, Multiarmed Bandits, Model Order Selection, Exploration-Exploitation Trade-off

1. Introduction

Learning under limited feedback and the exploration-exploitation trade-off are the fundamental questions in fields like reinforcement and active learning. The existing theoretical analysis of the exploration-exploitation trade-off in problems that go beyond multiarmed bandits is mainly focused on the worst-case scenarios (Strehl et al., 2009; Jaksch et al., 2010; Beygelzimer et al., 2011, 2009). But the worst-case analysis is overly pessimistic if the environment is not adversarial and cannot exploit the opportunities provided by benign conditions. We present a new analysis framework that lays the foundation for data-dependent analysis of the exploration-exploitation trade-off.

1. See also our follow-up work on PAC-Bayesian inequalities for martingales (Seldin et al., 2011b)

Our framework is based on PAC-Bayesian analysis. The PAC-Bayesian analysis was introduced over a decade ago (Shawe-Taylor and Williamson, 1997; Shawe-Taylor et al., 1998; McAllester, 1998; Seeger, 2002) and has since made a significant contribution to the analysis and development of supervised learning methods. PAC-Bayesian bounds provide an explicit and often intuitive and easy-to-optimize trade-off between model complexity and empirical data fit, where the complexity can be nailed down to the resolution of individual hypotheses via the definition of the prior. The PAC-Bayesian analysis was applied to derive generalization bounds and new algorithms for linear classifiers and maximum margin methods (Langford and Shawe-Taylor, 2002; McAllester, 2003; Germain et al., 2009), structured prediction (McAllester, 2007), and clustering-based classification models (Seldin and Tishby, 2010), to name just a few. However, the application of PAC-Bayesian analysis beyond the supervised learning domain remained surprisingly limited. In fact, the only additional domain known to us is density estimation (Seldin and Tishby, 2010; Higgs and Shawe-Taylor, 2010).

Application of PAC-Bayesian analysis to non-i.i.d. data was partially addressed only recently by Ralaivola et al. (2010) and Lever et al. (2010). The solution of Ralaivola et al. is based on breaking the sample into independent (or almost independent) subsets (which also reduces the effective sample size to the number of independent subsets). Such an approach is inapplicable in reinforcement learning due to strong dependence of the learning process on all of its history. Lever et al. treated dependent samples in the context of analysis of U-statistics. They employed Hoeffding's canonical decomposition of U-statistics into forward martingales and applied PAC-Bayesian analysis directly to these martingales. The approach presented here is both tighter and more general.

We present a generalization of PAC-Bayesian analysis to martingales. Our generalization makes it possible to consider model order selection simultaneously with the exploration-exploitation trade-off. Some potential advantages of applying PAC-Bayesian analysis in reinforcement learning were recently pointed out by several researchers, including Tishby and Polani (2010) and Fard and Pineau (2010). Tishby and Polani suggested to use the mutual information between states and actions in a policy as a natural regularizer in reinforcement learning. They showed that regularization by mutual information can be incorporated into Bellman equations and thereby computed efficiently. Tishby and Polani conjectured that PAC-Bayesian analysis can be applied to justify such a regularization and provide generalization guarantees for it.

Fard and Pineau derived a PAC-Bayesian analysis of batch reinforcement learning. However, batch reinforcement learning does not involve the exploration-exploitation trade-off.

One of the reasons for the difficulty of applying PAC-Bayesian analysis to address the exploration-exploitation trade-off is limited feedback (the fact that we only observe the reward for the action taken, but not for all other actions). In supervised learning (and also in density estimation) the empirical error of each hypothesis in a hypotheses class can be evaluated on all the samples and, therefore, the size of the sample available for evaluation of all the hypotheses is the same (and usually relatively large). In the situation of limited feedback the samples from one action cannot be used to evaluate another action and the sample size of "bad" actions has to increase sublinearly in the number of game rounds. In a precursory report (Seldin et al., 2011c) we overcame this difficulty by applying PAC-Bayesian analysis to importance weighted sampling (Sutton and Barto, 1998). Importance

weighted sampling is commonly used in the analysis of non-stochastic bandits (Auer et al., 2002b), but has not previously been applied to the analysis of stochastic bandits.

The usage of importance weighted sampling introduces two new difficulties. One is sequential dependence of the samples: the rewards observed in the past influence distribution over actions played in the future and through this distribution the variance of the subsequent weighted sample variables. The second problem introduced by weighted sampling is the growing variance of the weighted sample variables. In Seldin et al. (2011c) we handled this dependence by combining PAC-Bayesian analysis with Hoeffding-Azuma-type inequalities for martingales. The bounds achieved by such a combination provide $O(\frac{1}{\varepsilon_t\sqrt{t}})$ convergence rate, where t is the time step and ε_t is the minimal probability of sampling any action at time step t. The combination with Bernstein-type inequality for martingales presented here achieves $O(\frac{1}{\sqrt{\varepsilon_t t}})$ convergence rate. This improvement makes it possible to tighten the regret bounds from $O(K^{1/2}t^{3/4})$ to $O(K^{1/3}t^{2/3})$, where K is the number of arms. In Section 3 we suggest possible ways to tighten the analysis further to get $O(\sqrt{Kt})$ regret bounds. These further improvements will be studied in detail in future work.

We repeat that our main goal is not improvement of existing bounds for stochastic multiarmed bandits, which are already tight up to $\sqrt{\ln(K)}$ factors (Audibert and Bubeck, 2009; Auer and Ortner, 2010), but rather development of a new powerful tool for reinforcement learning and for other domains with richer structure. The multiarmed bandits serve us as a testbed for the development of this new tool. One example of a problem with a richer structure are multiarmed bandits with side information (a.k.a. contextual bandits). Beygelzimer et al. (2011) suggested $O\left(\sqrt{Kt\ln(N/\delta)}\right)$ and $O\left(\sqrt{t(d\ln t-\ln\delta)}\right)$ regret bounds for learning with expert advice in multiarmed bandits with side information, where N is the number of experts (in case it is finite) and d is the VC-dimension of the set of experts (in case it is infinite). In the follow-up paper Seldin et al. (2011a) we show that PAC-Bayesian analysis makes it possible to replace $\ln(N)$ and d factors with $KL(\rho\|\mu)$, where KL is the KL-divergence, $\rho(h)$ is a distribution over the experts played by the algorithm, and $\mu(h)$ is a prior distribution over the experts. Such an approach is much more flexible, since it allows individual treatment of different experts (or policies) via the definition of the prior μ.

The paper is organized as follows: Section 2 surveys the main results of the paper, Section 3 suggests possible ways to tighten the analysis further, and Section 4 discusses the results. Proofs are provided in the appendix.

2. Main Results

We start with a general concentration result for martingales based on combination of PAC-Bayesian analysis with a Bernstein-type inequality for martingales. Then, we apply this result to derive an instantaneous (per-round) bound on the distance between expected and empirical regret for the multiarmed bandit problem. This result is in turn applied to derive an instantaneous regret bound for the multiarmed bandits.

2.1. PAC-Bayes-Bernstein Inequality for Martingales

In order to present our concentration result for martingales we need a few definitions. Let $\mathcal{H}$ be an index (or a hypothesis) space, possibly uncountably infinite. Let $\{X_1(h), X_2(h), \cdots : h \in \mathcal{H}\}$ be martingale difference sequences, meaning that $\mathbb{E}[X_t(h)|\mathcal{T}_{t-1}] = 0$, where $\mathcal{T}_t = \{X_\tau(h) : 1 \leq \tau \leq t \text{ and } h \in \mathcal{H}\}$ is a set of martingale differences observed up to time t (the history). ($\{X_t(h)\}_{h\in\mathcal{H}}$ do not have to be independent, we only need the requirement on the conditional expectation to be satisfied.) Let $M_t(h) = \sum_{\tau=1}^{t} X_\tau(h)$ be martingales corresponding to the martingale difference sequences and let $V_t(h) = \sum_{\tau=1}^{t} \mathbb{E}[X_\tau(h)^2|\mathcal{T}_{\tau-1}]$ be cumulative variances of the martingales. For a distribution ρ over $\mathcal{H}$ define weighted averages of the martingales and their cumulative variances with respect to ρ as $M_t(\rho) = \mathbb{E}_{\rho(h)}[M_t(h)]$ and $V_t(\rho) = \mathbb{E}_{\rho(h)}[V_t(h)]$.

Theorem 1 (PAC-Bayes-Bernstein Inequality) *Let $\{C_1, C_2, \dots\}$ be an increasing sequence set in advance, such that $|X_t(h)| \leq C_t$ for all h with probability 1. Let $\{\mu_1, \mu_2, \dots\}$ be a sequence of "reference" ("prior") distributions over $\mathcal{H}$, such that μ_t is independent of $\mathcal{T}_t$ (but can depend on t). Let $\{\lambda_1, \lambda_2, \dots\}$ be a sequence of positive numbers set in advance that satisfy:*

$$\lambda_t \leq \frac{1}{C_t}. \tag{1}$$

Then for all possible distributions ρ_t over $\mathcal{H}$ given t and for all t simultaneously with probability greater than $1-\delta$:

$$|M_t(\rho_t)| \leq \frac{KL(\rho_t\|\mu_t) + 2\ln(t+1) + \ln\frac{2}{\delta}}{\lambda_t} + (e-2)\lambda_t V_t(\rho_t). \tag{2}$$

Bound (2) is minimized by $\lambda_t = \sqrt{\frac{KL(\rho_t\|\mu_t)+2\ln(t+1)+\ln\frac{2}{\delta}}{(e-2)V_t(\rho_t)}}$. For this value of λ_t we would get

$$|M_t(\rho_t)| \leq 2\sqrt{(e-2)V_t(\rho_t)\left(KL(\rho_t\|\mu_t) + 2\ln(t+1) + \ln\frac{2}{\delta}\right)}, \tag{3}$$

however, λ_t has to be set in advance and cannot depend on the sample. Therefore, we have to make our best guess of what the values of $KL(\rho_t\|\mu_t)$ and $V_t(\rho_t)$ are going to be, which is actually possible in the case that we study below. In the follow-up paper we show that by taking an exponentially spaced grid of λ_t-s and a union bound over this grid it is possible to derive a bound, which is almost as good as (3) (Seldin et al., 2011b), but this extension is not required in the current work.

2.2. Application to the Multiarmed Bandit Problem

In order to apply our result to the multiarmed bandit problem we need some more definitions. Let $\mathcal{A}$ be a set of actions (arms) of size $|\mathcal{A}| = K$ and let $a \in \mathcal{A}$ denote the actions. Denote by $R(a)$ the expected reward of action a. 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 independently at random according to $\{\pi_1, \pi_2, \dots\}$ respectively. Let $\{R_1, R_2, \dots\}$ be the sequence of observed rewards. Denote by $\mathcal{T}_t = \{\{\pi_1, \dots, \pi_t\}, \{A_1, \dots, A_t\}, \{R_1, \dots, R_t\}\}$ the set of played policies, taken actions, and observed rewards up to round t.

For $t \geq 1$ and $a \in \{1, \dots, K\}$ define a set of random variables R_t^a (the importance weighted samples):

$$R_t^a = \begin{cases} \frac{1}{\pi_t(a)} R_t, & \text{if } A_t = a \\ 0, & \text{otherwise.} \end{cases}$$

Define:

$$\hat{R}_t(a) = \frac{1}{t} \sum_{\tau=1}^{t} R_\tau^a.$$

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

Let a^* be the "best" action (the action with the highest expected reward, if there are multiple "best" actions pick any of them). Define the expected and empirical per-round regrets as:

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

Observe that $t(\hat{\Delta}_t(a) - \Delta(a))$ form a martingale. Let

$$V_t(a) = \sum_{\tau=1}^{t} \mathbb{E}[([R_\tau^{a^*} - R_\tau^a] - [R(a^*) - R(a)])^2 | \mathcal{T}_{\tau-1}]$$

be the cumulative variance of this martingale.

Let $\{\varepsilon_1, \varepsilon_2, \dots\}$ be a decreasing sequence that satisfies $\varepsilon_t \leq \min_a \pi_t(a)$ (we say that $\pi_t(a)$ is *bounded from below* by ε_t). In the appendix we prove the following upper bound on $V_t(a)$.

Lemma 2 *For all t and a:*

$$V_t(a) \leq \frac{2t}{\varepsilon_t}.$$

For a distribution ρ over $\mathcal{A}$ define the expected and empirical regret of ρ as $\Delta(\rho) = \mathbb{E}_{\rho(a)}[\Delta(a)]$ and $\hat{\Delta}_t(\rho) = \mathbb{E}_{\rho(a)}[\hat{\Delta}_t(a)]$. The following theorem follows immediately from Theorem 1 and Lemma 2 by taking a uniform prior over the actions.

Theorem 3 *For any sequence of sampling distributions $\{\pi_1, \pi_2, \dots\}$ that are bounded from below by a decreasing sequence $\{\varepsilon_1, \varepsilon_2, \dots\}$ that satisfies*

$$\frac{\ln(K) + 2\ln(t+1) + \ln\frac{2}{\delta}}{2(e-2)t} \leq \varepsilon_t, \tag{4}$$

where π_t can depend on $\mathcal{T}_{t-1}$, for all possible distributions ρ_t given t and for all $t \geq 1$ simultaneously with probability greater than $1 - \delta$:

$$\left|\Delta(\rho_t) - \hat{\Delta}_t(\rho_t)\right| \leq 2\sqrt{\frac{2(e-2)\left(\ln(K) + 2\ln(t+1) + \ln\frac{2}{\delta}\right)}{t\varepsilon_t}}. \tag{5}$$

Proof For a uniform prior $\mu_t(a) = \frac{1}{K}$ we have $KL(\rho_t\|\mu_t) \leq \ln(K)$. By Lemma 2, for any ρ_t the weighted cumulative variance is bounded by $V_t(\rho_t) \leq \frac{2t}{\varepsilon_t}$. By taking $\lambda_t = \sqrt{\frac{\ln(K)+2\ln(t+1)+\ln\frac{2}{\delta}}{2(e-2)t}}$ and substituting the bounds on $KL(\rho_t\|\mu_t)$ and $V_t(\rho_t)$ into (2) we obtain (5). (We considered the martingales $t(\Delta(a) - \hat{\Delta}_t(a))$, which provided a factor of t in the denominator.) The technical condition (4) follows from the requirement (1) on λ_t. ■

Remarks: Theorem 3 provides an improvement over the corresponding Theorems 2 and 3 in the precursory report (Seldin et al., 2011c) by decreasing the dependence on ε_t from $1/\varepsilon_t$ to $1/\sqrt{\varepsilon_t}$. This in turn makes it possible to improve the regret bound, which is shown next. Interestingly, the uniform prior μ_t yields a tighter (and also simpler) bound than a distribution-dependent prior used in Seldin et al. (2011c). It also broadens the range of playing strategies for which the regret bound given in Theorem 4 holds. We note that the uniform prior neutralizes the power of PAC-Bayesian analysis to discriminate between different hypotheses. For problems with richer structure studied in the follow-up paper (Seldin et al., 2011a), more interesting priors can be defined that yield advantages over alternative approaches. The multiarmed bandit problem studied here is, nevertheless, important for the development of the new tool.

We note that in the next theorem we take $\varepsilon_t = K^{-2/3}t^{-1/3}$ and the technical condition (4) is satisfied for t that is slightly larger than $K(\ln(K) + \ln\frac{2}{\delta})^{3/2}$.

Theorem 4 *Let $\varepsilon_t = K^{-2/3}t^{-1/3}$ and take any γ_t, such that $\gamma_t \geq K^{-1/3}t^{1/3}\sqrt{\ln K}$. For $t < K$ let $\pi_t(a) = \frac{1}{K}$ for all a and for $t \geq K$ let*

$$\pi_{t+1}(a) = \tilde{\rho}_t^{exp}(a) = (1 - K\varepsilon_{t+1})\rho_t^{exp}(a) + \varepsilon_{t+1},$$

where

$$\rho_t^{exp}(a) = \frac{1}{Z(\rho_t^{exp})}e^{\gamma_t\hat{R}_t(a)}$$

and

$$Z(\rho_t^{exp}) = \sum_a e^{\gamma_t\hat{R}_t(a)}.$$

Then the expected per-round regret $\Delta(\tilde{\rho}_t^{exp}) = R(a^*) - R(\tilde{\rho}_t^{exp})$ is bounded by:

$$\Delta(\tilde{\rho}_t^{exp}) \leq \frac{K^{1/3}}{(t+1)^{1/3}}\left(1 + \sqrt{\ln K} + 2\sqrt{2(e-2)\left(\ln(K) + 2\ln(t+1) + \ln\frac{2}{\delta}\right)}\right)$$

with probability greater than $1-\delta$ simultaneously for all rounds t, where t satisfies (4) (which means that $t \geq K\left(\frac{\ln(K)+2\ln(t+1)+\ln\frac{2}{\delta}}{2(e-2)}\right)^{3/2}$, note that t also appears on the right hand side). This translates into a total regret of $\tilde{O}(K^{1/3}t^{2/3})$ (where $\tilde{O}$ hides logarithmic factors).

For $\gamma_t = \varepsilon_t^{-1}$ the playing strategy in Theorem 4 is known as the EXP3 algorithm for adversarial bandits (Auer et al., 2002b), which is applied here to stochastic bandits. When γ_t tends to infinity, we obtain the ε-greedy algorithm for stochastic bandits (Auer et al., 2002a). Theorem 4 covers the spectrum of all possible intermediate strategies.

3. Towards a Tighter Regret Bound

We note that there is still a room for improvement, which we believe will enable to achieve regret bounds of order $\tilde{O}(\sqrt{Kt})$. The main source of looseness is the usage of the crude global upper bound $\frac{2t}{\varepsilon_t}$ on the cumulative variances in Lemma 2 that holds for any distribution ρ_t. While this bound seems to be tight for the ε-greedy strategy, we believe that it can be tightened for the EXP3 algorithm. It is possible to show that if we play according to the distributions $\{\tilde{\rho}_1^{exp}, \ldots, \tilde{\rho}_t^{exp}\}$, then for "good" actions a (those for which $\Delta(a) \leq \frac{1}{\gamma_t}$) the cumulative variance $V_t(a)$ is bounded by CKt for some constant C. If we could show that for "bad" actions a (those for which $\Delta(a) > \frac{1}{\gamma_t}$) the probability ρ_t^{exp} of picking such actions is bounded by $C\varepsilon_t$, then the cumulative variance $V_t(\rho_t^{exp})$ would be bounded by CKt. This is, in fact, true for "very bad" actions (those, for which $\Delta(a)$ is close to 1), but it does not hold for actions with $\Delta(a)$ close to $\frac{1}{\gamma_t}$. However, we can possibly show that for such actions $\rho_t^{exp}(a) \leq C\varepsilon_t$ for most of the rounds ($1 - \varepsilon_t$ fraction will suffice) and then we will be able to achieve $\tilde{O}(\sqrt{Kt})$ regret. In the experiment that follows we provide an empirical evidence that this conjecture holds in practice.

Another possible approach is to apply the EXP3.P algorithm of Auer et al. (2002b). However, in the experiment that follows we show that in the stochastic setting EXP3 algorithm achieves much lower regret than EXP3.P. It is, therefore, worth exploring the first route. We also note that Auer et al. (2002b) do not provide an explicit bound on the variance of EXP3.P, which is required for our bound. This would have to be done for the second way of achieving $\tilde{O}(\sqrt{Kt})$ regret bound.

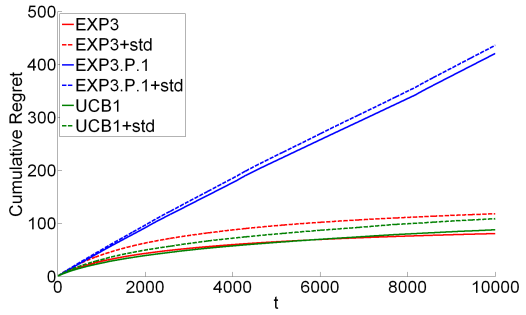
3.1. Empirical Test Study

In the following experiment we show that in the stochastic setting EXP3 algorithm achieves lower regret compared to EXP3.P.1 algorithm of Auer et al. (2002a). We also show that the variance of EXP3 algorithm is reasonably close to $2Kt$. Finally, we show that in the stochastic setting the regret of EXP3 algorithm is comparable or even lower than the regret of UCB strategy (Auer et al., 2002a) in the short run, but gets worse in the long run. We note that UCB strategy is not compatible with PAC-Bayesian analysis, since in UCB every action has its own sample size and the sample size of "bad" actions grows sublinearly with the number of game rounds. Designing a strategy that would be compatible with PAC-Bayesian analysis and achieve the regret of UCB in the long run is an important direction for future research.

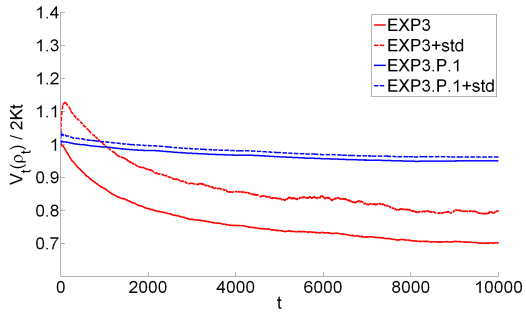
Experiment Setup

We took a 2-arm bandit problem with biases 0.5 and 0.6 for the two arms and ran EXP3 algorithm from Theorem 4 with $\varepsilon_t = 1/\sqrt{Kt}$ and $\gamma_t = \sqrt{t \ln K/K}$, EXP3.P.1 algorithm of Auer et al. (2002b) with $\delta = 0.001$, and UCB1 algorithm of Auer et al. (2002a). In the first experiment we made 1000 repetitions of the game and in each game we ran each of the algorithms for 10,000 rounds. In the second experiment we made 100 repetitions of the game and in each game we ran each of the algorithms for 10^7 rounds. In Figure 1 we show:

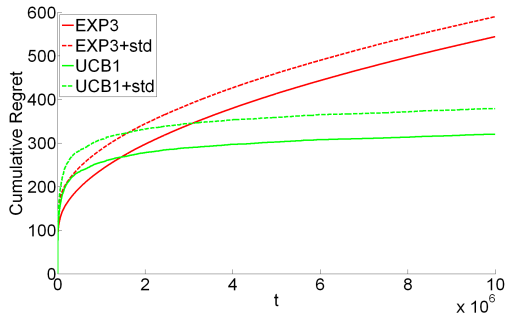
1.a Experiment 1 (10^4 rounds): Average (over 1000 repetitions of the game) cumulative regret of EXP3, EXP3.P.1, and UCB1 algorithms.



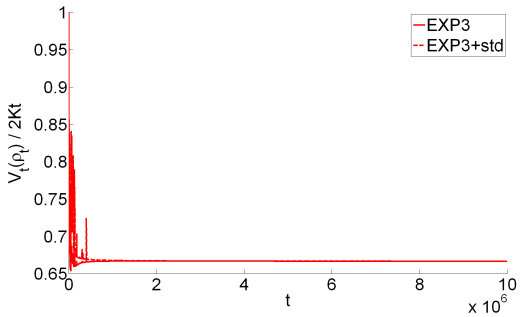
(a) Cumulative Regret, 10^4 rounds



(b) $\frac{1}{2Kt}$ · (Cumulative Variance), 10^4 rounds



(c) Cumulative Regret, 10^7 rounds



(d) $\frac{1}{2Kt}$ · (Cumulative Variance), 10^7 rounds

Figure 1: **Experimental results.** Solid lines show mean values over experiment repetitions, dotted lines show mean values plus one standard deviation (std).

1.b Experiment 1: Average cumulative variance of EXP3 and EXP3.P.1 normalized by $2Kt$, which is what we would like it to be: $\frac{1}{2Kt} \cdot \frac{1}{1000} \sum_{i=1}^{1000} V_t^i(\rho_t)$, where $i \in [1, \ldots, 1000]$ indexes the experiments.

1.c Experiment 2 (10^7 rounds): Average (over 100 repetitions of the game) cumulative regret of EXP3 and UCB1 algorithms. The regret of EXP3.P.1 algorithm was far above the regret of EXP3 and UCB1 and, therefore, was omitted from the graphs.

1.d Experiment 2: Average cumulative variance of EXP3 normalized by $2Kt$.

Observations

1. In the stochastic setting the performance of EXP3 is significantly superior to the performance of EXP3.P.1.

2. In the stochastic setting, the performance of EXP3 is comparable or even superior to the performance of UCB1 in the short run, but becomes worse than the performance of UCB1 in the long run (beyond $2 \cdot 10^6$ iterations). The reason is that the number of pulls of the suboptimal arm are roughly $\sqrt{t}$ for EXP3 and $\ln(t)/\Delta(a)^2$ for UCB. In our experiment $\Delta(a) = 0.1$ for the suboptimal arm, thus $\sqrt{t} > \ln(t)/\Delta(a)^2$ when $t > \ln(t)^2/\Delta(a)^4$, which holds when $t > 2 \cdot 10^6$.

3. In the stochastic setting, the variance of EXP3 is initially higher than the variance of EXP3.P.1, but eventually it becomes lower.

4. Initially the variance of EXP3 is just slightly above $2Kt$ (by a factor of less than 2) and eventually it stabilizes around $0.66 \cdot 2Kt$ for the problem that we considered.

4. Discussion

We presented a new framework for data-dependent analysis of the exploration-exploitation trade-off and for simultaneous analysis of model order selection and the exploration-exploitation trade-off. We note that model order selection does not come up in the multiarmed bandit problem due to simplicity of the structure of this problem. Nevertheless, the multiarmed bandit problem is a convenient playground for the development of the new tool. In the follow-up paper we show that the new technique developed here can be applied to multiarmed bandits with side information and yield an advantage over state-of-the-art (Seldin et al., 2011a).

An important direction for future research is to tighten Theorems 3 and 4, so that the regret bound will match state-of-the-art regret bounds obtained by alternative techniques. We believe that the ideas described in Section 3 can make it possible. The experiments presented in Section 3 show that empirically in the stochastic setting our algorithm is significantly superior to state-of-the-art algorithms for adversarial bandits and slightly worse than state-of-the-art algorithms for stochastic bandits. Closing the gap with state-of-the-art algorithms for stochastic bandits is another important direction for future research.

Other directions for future research include application of our framework to Markov decision processes (Fard and Pineau, 2010), active learning (Beygelzimer et al., 2009), and problems with continuous state and action spaces, such as Gaussian process bandits (Srinivas et al., 2010).

Appendix A. Proofs

In this appendix we provide the proofs of Theorems 1 and 4 and Lemma 2.

A.1. Proof of Theorem 1

The proof of Theorem 1 relies on the following two lemmas. The first one is a Bernstein-type inequality. For a proof of Lemma 5 see, for example, the proof of Theorem 1 in Beygelzimer et al. (2011).

Lemma 5 (Bernstein's inequality) *Let $X_1, \ldots, X_t$ be a martingale difference sequence (meaning that $\mathbb{E}[X_\tau | X_1, \ldots, X_{\tau-1}] = 0$ for all τ), such that $X_\tau \leq C$ for all τ with probability 1. Let $M_t = \sum_{\tau=1}^{t} X_\tau$ be a corresponding martingale and $V_t = \sum_{\tau=1}^{t} \mathbb{E}[X_\tau^2 | X_1, \ldots, X_{\tau-1}]$ be the cumulative variance of this martingale. Then for any fixed $\lambda \in [0, \frac{1}{C}]$:*

$$\mathbb{E} e^{\lambda M_t - (e-2)\lambda^2 V_t} \leq 1.$$

The second lemma originates in statistical physics and information theory (Donsker and Varadhan, 1975; Dupuis and Ellis, 1997; Gray, 2011) and forms the basis of PAC-Bayesian analysis. See (Banerjee, 2006) for a proof.

Lemma 6 (Change of measure inequality) *For any measurable function $\phi(h)$ on $\mathcal{H}$ and any distributions $\mu(h)$ and $\rho(h)$ on $\mathcal{H}$, we have*:

$$\mathbb{E}_{\rho(h)}[\phi(h)] \leq KL(\rho\|\mu) + \ln \mathbb{E}_{\mu(h)}[e^{\phi(h)}].$$

Now we are ready to state the proof of Theorem 1.

Proof of Theorem 1 Take $\phi(h) = \lambda_t M_t(h) - (e-2)\lambda_t^2 V_t(h)$ and $\delta_t = \frac{1}{t(t+1)}\delta \geq \frac{1}{(t+1)^2}\delta$. (It is well-known that $\sum_{t=1}^{\infty} \frac{1}{t(t+1)} = \sum_{t=1}^{\infty}\left(\frac{1}{t} - \frac{1}{t+1}\right) = 1$.) Then the following holds for all ρ_t and t simultaneously with probability greater than $1 - \frac{\delta}{2}$:

$$\begin{aligned}
\lambda_t M_t(\rho_t) - (e-2)\lambda_t^2 V_t(\rho_t) &= \mathbb{E}_{\rho_t(h)}[\lambda_t M_t(h) - (e-2)\lambda_t^2 V_t(h)] && (6)\\
&\leq KL(\rho_t\|\mu_t) + \ln \mathbb{E}_{\mu_t(h)}[e^{\lambda_t M_t(h) - (e-2)\lambda_t^2 V_t(h)}] && (7)\\
&\leq KL(\rho_t\|\mu_t) + 2\ln(t+1) + \ln\frac{2}{\delta} + \ln \mathbb{E}_{\mathcal{T}_t}\mathbb{E}_{\mu_t(h)}[e^{\lambda_t M_t(h) - (e-2)\lambda_t^2 V_t(h)}] && (8)\\
&= KL(\rho_t\|\mu_t) + 2\ln(t+1) + \ln\frac{2}{\delta} + \ln \mathbb{E}_{\mu_t(h)}\mathbb{E}_{\mathcal{T}_t}[e^{\lambda_t M_t(h) - (e-2)\lambda_t^2 V_t(h)}] && (9)\\
&\leq KL(\rho_t\|\mu_t) + 2\ln(t+1) + \ln\frac{2}{\delta}, && (10)
\end{aligned}$$

where (6) is by definition of $M_t(\rho_t)$ and $V_t(\rho_t)$, (7) is by Lemma 6, (8) holds with probability greater than $1 - \frac{\delta}{2}$ by Markov's inequality and a union bound over t, (9) is due to the fact that μ_t is independent of $\mathcal{T}_t$, and (10) is by Lemma 5.

By applying the same argument to martingales $-M_t(h)$ and taking a union bound over the two we obtain that with probability greater than $1 - \delta$:

$$|M_t(\rho_t)| \leq \frac{KL(\rho_t\|\mu_t) + 2\ln(t+1) + \ln\frac{2}{\delta}}{\lambda_t} + (e-2)\lambda_t V_t(\rho_t),$$

which is the statement of the theorem. The technical condition (1) follows from the requirement that $\lambda_t \in [0, \frac{1}{C_t}]$. ■

A.2. Proof of Lemma 2

Proof of Lemma 2

$$\begin{aligned}
V_t(a) &= \sum_{\tau=1}^{t} \mathbb{E}[([R_\tau^{a^*} - R_\tau^a] - [R(a^*) - R(a)])^2 | \mathcal{T}_{\tau-1}]\\
&= \left(\sum_{\tau=1}^{t} \mathbb{E}[(R_\tau^{a^*} - R_\tau^a)^2 | \mathcal{T}_{\tau-1}]\right) - t\Delta(a)^2 && (11)\\
&\leq \left(\sum_{\tau=1}^{t}\left(\frac{\pi_\tau(a)}{\pi_\tau(a)^2} + \frac{\pi_\tau(a^*)}{\pi_\tau(a^*)^2}\right)\right) && (12)\\
&= \left(\sum_{\tau=1}^{t}\left(\frac{1}{\pi_\tau(a)} + \frac{1}{\pi_\tau(a^*)}\right)\right)\\
&\leq \frac{2t}{\varepsilon_t}, && (13)
\end{aligned}$$

where (11) is due to the fact that $\mathbb{E}[R_\tau^a|\mathcal{T}_{\tau-1}] = R(a)$, (12) is due to the fact that $R_t \leq 1$ and $t\Delta(a)^2 \geq 0$, and (13) is due to the fact that $\frac{1}{\pi_\tau(a)} \leq \frac{1}{\varepsilon_t}$ for all a and $1 \leq \tau \leq t$. ■

A.3. Proof of Theorem 4

Proof of Theorem 4 We use the following regret decomposition:

$$\Delta(\tilde{\rho}_t^{exp}) = [\Delta(\rho_t^{exp}) - \hat{\Delta}_t(\rho_t^{exp})] + \hat{\Delta}_t(\rho_t^{exp}) + [R(\rho_t^{exp}) - R(\tilde{\rho}_t^{exp})]. \tag{14}$$

The first term in the decomposition is bounded by Theorem 3. Before bounding the middle term in (14) we bound the last term, which is much simpler, and then return to the middle term. The bound on $[R(\rho_t^{exp}) - R(\tilde{\rho}_t^{exp})]$ is achieved by the following lemma.

Lemma 7 *Let $\tilde{\rho}$ be an ε-smoothed version of ρ, such that*

$$\tilde{\rho}(a) = (1 - K\varepsilon)\rho(a) + \varepsilon.$$

Then

$$R(\rho) - R(\tilde{\rho}) \leq K\varepsilon. \tag{15}$$

Proof

$$\begin{aligned} R(\rho) - R(\tilde{\rho}) &= \sum_a (\rho(a) - \tilde{\rho}(a))R(a) \\ &\leq \frac{1}{2}\sum_a |\rho(a) - \tilde{\rho}(a)| \qquad (16) \\ &= \frac{1}{2}\sum_a |\rho(a) - (1 - K\varepsilon)\rho(a) - \varepsilon| \\ &= \frac{1}{2}\sum_a |K\varepsilon\rho(a) - \varepsilon| \\ &\leq \frac{1}{2}K\varepsilon\sum_a \rho(a) + \frac{1}{2}K\varepsilon \\ &= K\varepsilon. \end{aligned}$$

In (16) we used the fact that $0 \leq R(a) \leq 1$ and ρ and $\tilde{\rho}$ are probability distributions. ■

In the next lemma we bound $\hat{\Delta}(\rho_t^{exp})$.

Lemma 8

$$\hat{\Delta}(\rho_t^{exp}) \leq \frac{\ln K}{\gamma_t}. \tag{17}$$

Proof Observe that by multiplying nominator and denominator in the definition of ρ_t^{exp} by $e^{-\gamma_t \hat{R}_t(a^*)}$ we obtain:

$$\rho_t^{exp}(a) = \frac{e^{\gamma_t \hat{R}_t(a)}}{Z(\rho_t^{exp})} = \frac{e^{-\gamma_t \hat{\Delta}_t(a)}}{Z'(\rho_t^{exp})},$$

where $Z'(\rho_t^{exp}) = \sum_a e^{-\gamma_t \hat{\Delta}_t(a)}$. The empirical regret $\hat{\Delta}_t(\rho_t^{exp})$ then obtains the form:

$$\hat{\Delta}_t(\rho_t^{exp}) = \sum_a \rho_t(a)\hat{\Delta}_t(a) = \frac{\sum_a \hat{\Delta}_t(a) e^{-\gamma_t \hat{\Delta}_t(a)}}{\sum_a e^{-\gamma_t \hat{\Delta}_t(a)}}.$$

The lemma follows from Lemma 9 below and the observation that $\hat{\Delta}_t(a^*) = 0$. ■

Lemma 9 *Let $x_1 = 0$ and $x_2, \ldots, x_n$ be $n-1$ arbitrary numbers. For any $\alpha > 0$ and $n \geq 2$:*

$$\frac{\sum_{i=1}^n x_i e^{-\alpha x_i}}{\sum_{j=1}^n e^{-\alpha x_j}} \leq \frac{\ln(n)}{\alpha}. \tag{18}$$

Proof Since negative x_i-s only decrease the left hand side of (18) we can assume without loss of generality that all x_i-s are positive. Due to symmetry, the maximum is achieved when all x_i-s (except x_1) are equal:

$$\frac{\sum_{i=1}^n x_i e^{-\alpha x_i}}{\sum_{j=1}^n e^{-\alpha x_j}} \leq \max_x \frac{(n-1)xe^{-\alpha x}}{1+(n-1)e^{-\alpha x}}. \tag{19}$$

We apply change of variables $y = e^{-\alpha x}$, which means that $x = \frac{1}{\alpha}\ln\frac{1}{y}$. By substituting this into the right hand side of (19) we get

$$\frac{(n-1)xe^{-\alpha x}}{1+(n-1)e^{-\alpha x}} = \frac{1}{\alpha}\cdot\frac{(n-1)y\ln\frac{1}{y}}{1+(n-1)y}.$$

In order to prove the bound we have to show that $\frac{(n-1)y\ln\frac{1}{y}}{1+(n-1)y} \leq \ln n$.

By taking Taylor's expansion of $\ln z$ around $z = n$ we have:

$$\ln z \leq \ln n + \frac{1}{n}(z-n) = \ln n + \frac{z}{n} - 1.$$

Thus:

$$\begin{aligned}
\frac{(n-1)y\ln\frac{1}{y}}{1+(n-1)y} &\leq \frac{(n-1)y(\ln n + \frac{1}{ny} - 1)}{1+(n-1)y} \\
&\leq \frac{y(n-1)\ln n + \frac{n-1}{n}}{(n-1)y+1} \\
&\leq \frac{(y(n-1)+1)\ln n}{y(n-1)+1} \qquad (20) \\
&= \ln n,
\end{aligned}$$

where (20) follows from the fact that $\ln z \leq z-1$ for any positive z, and hence $\ln\frac{1}{n} \leq \frac{1}{n} - 1$, which means that $\ln n \geq 1 - \frac{1}{n} = \frac{n-1}{n}$ for all $n > 0$. ■

Substitution of (5), (15), (17), and the choice of ε_t and γ_t in theorem formulation into (14) concludes the proof.

■

Acknowledgments

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^o231495. This publication only reflects the authors' views.

References

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.

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 multi-armed bandit problem. *SIAM Journal of Computing*, 32(1), 2002b.

Arindam Banerjee. On Bayesian bounds. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2006.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

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.

Monroe D. Donsker and S.R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28, 1975.

Paul Dupuis and Richard S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley-Interscience, 1997.

Mahdi Milani Fard and Joelle Pineau. PAC-Bayesian model selection for reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

Robert M. Gray. *Entropy and Information Theory*. Springer, 2^{nd} edition, 2011.

Matthew Higgs and John Shawe-Taylor. A PAC-Bayes bound for tailored density estimation. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2010.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11, 2010.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems (NIPS)*, 2002.

Guy Lever, François Laviolette, and John Shawe-Taylor. Distribution-dependent PAC-Bayes priors. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2010.

David McAllester. Some PAC-Bayesian theorems. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 1998.

David McAllester. Simplified PAC-Bayesian margin bounds. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2003.

David McAllester. Generalization bounds and consistency for structured labeling. In Gökhan Bakir, Thomas Hofmann, Bernhard Schölkopf, Alexander Smola, Ben Taskar, and S.V.N. Vishwanathan, editors, *Predicting Structured Data*. The MIT Press, 2007.

Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic PAC-Bayes bounds for non-IID data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 2010.

Matthias Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 2002.

Yevgeny Seldin and Naftali Tishby. PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11, 2010.

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)*, 2011a.

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. 2011b. In review. Preprint available at http://arxiv.org/abs/1110.6886.

Yevgeny Seldin, François Laviolette, John Shawe-Taylor, Jan Peters, and Peter Auer. PAC-Bayesian analysis of martingales and multiarmed bandits. http://arxiv.org/abs/1105.2416, 2011c.

John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 1997.

John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5), 1998.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 2009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In Vassilis Cutsuridis, Amir Hussain, John G. Taylor, and Daniel Polani, editors, *Perception-Reason-Action Cycle: Models, Algorithms and Systems*. Springer, 2010.