PAC-Bayesian Analysis of the Exploration-Exploitation Trade-off

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

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

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

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

Jan Peters JAN.PETERS@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

We develop a coherent framework for integrative simultaneous analysis of the exploration-exploitation and model order selection trade-offs. We improve over our preceding results on the same subject (Seldin et al., 2011) by combining PAC-Bayesian analysis with Bernstein-type inequality for martingales. Such a combination is also of independent interest for studies of multiple simultaneously evolving martingales.

1. Introduction

The trade-off between exploration and exploitation is a fundamental question in reinforcement learning. Model order selection, which is a trade-off between model complexity and its empirical data fit, is even a more basic question in machine learning. To the best of our knowledge, we develop the first framework that enables to consider these two trade-offs simultaneously from a finite sample perspective. The importance of simultaneous consideration of the two trade-offs can be illustrated by the following simple example. Imagine we have a web page, where we can show a visitor a single advertisement out of a pool of advertisements. Assume that we are given access to additional side information about the visitors, which we are allowed to use in our choice of advertisements (this is generally known as contextual bandits problem). Further, imagine that the amount of available (contextual) side information is very large (and potentially unlimited). Considering all side information from the beginning will result in an overcomplicated model that will take prohibitively many trials to learn. Instead, similar to supervised learning, we should start with a simple model and increase its complexity as our experience grows. However, unlike in supervised learning, we have to learn under limited feedback. This means that the model order selection trade-off has to be considered simultaneously with the exploration-exploitation trade-off. We develop an integrative framework that provides finite sample guarantees for both trade-offs simultaneously.

Our solution is based on extending PAC-Bayesian analysis of supervised learning with i.i.d. samples to problems with limited feedback and sequentially dependent samples. PAC-Bayesian analysis was introduced over a decade ago (Shawe-Taylor & 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. The power of PAC-Bayesian approach lies in successful marriage of the flexibility and intuitiveness of Bayesian models with the rigor of PAC analysis. 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 prior definition. The PAC-Bayesian analysis was applied to derive generalization bounds and new algorithms for linear classifiers and maximum margin methods (Langford & Shawe-Taylor, 2002; McAllester, 2003; Germain et al., 2009), structured prediction (McAllester, 2007), and clustering-based classification models (Seldin & 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 & Tishby, 2010; Higgs & Shawe-Taylor, 2010).

Some potential advantages of applying PAC-Bayesian analysis in reinforcement learning were recently pointed out by several researchers, including Tishby & Polani (2010) and Fard & Pineau (2010). Tishby & Polani (2010) suggested that the mutual information between states and actions in a policy can be used 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 form of regularization and provide generalization guarantees for it.

Fard & Pineau (2010) suggested 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 the limited feedback (the fact that we only observe the reward for the action taken, but not for all the rest). In supervised learning (and also in density estimation) the empirical error for each hypothesis within 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 sample 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 (Seldin et al., 2011) we resolved this issue by applying weighted sampling strategy (Sutton & Barto, 1998), which is commonly used in the analysis of non-stochastic bandits (Auer et al., 2002), but has not been applied to the analysis of stochastic bandits previously.

The usage of weighted sampling introduces two new difficulties. One is sequential dependence of the samples: the rewards we observe influence the distribution over actions we play and through this distribution influence the variance of the subsequent weighted sample variables. In (Seldin et al., 2011) we handled this dependence by combining PAC-Bayesian analysis with Hoeffding-Azuma-type inequalities for martingales.

The second problem introduced by weighted sampling is the growing variance of the weighted sample variables. We did not succeed to take full control over the variance in (Seldin et al., 2011) and the bound we obtained there depended on $1/\varepsilon_t$, where ε_t is the minimal probability for sampling any action at time step t. Here we improve this dependence to $1/\sqrt{\varepsilon_t}$ by combining PAC-Bayesian analysis with Bernstein-type inequality for martingales. This improvement enables 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 and t is the game round. The combination PAC-Bayesian analysis with Bernstein-type inequality for martingales is also of independent interest for studies of multiple simultaneously evolving martingales.

At the end of Section 2 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 emphasize that although this paper is focused on the multiarmed bandit problem, our main goal is not improving existing bounds for stochastic multiarmed bandits, which are already tight up to $\sqrt{\ln(K)}$ factors (Audibert & Bubeck, 2009; Auer & Ortner, 2010), but rather developing a new powerful tool for reinforcement learning in domains with a richer structure. For example, Beygelzimer et al. (2010) 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 the bandit setting, 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). We believe that PAC-Bayesian analysis should enable to replace $\ln(N)$ and d factors with $KL(\rho\|\mu)$, where $\rho(h)$ is a distribution over experts played by the algorithm and $\mu(h)$ is a prior distribution over experts that, for example, can reflect their complexity, and KL is the KL-divergence. Such an approach is much more flexible, since it allows individual treatment of different experts (or policies) via the prior definition μ and can be applied to both finite and infinite policy spaces (or expert sets). Our experience in supervised learning shows that PAC-Bayesian analysis is also handful for treating tree-shaped graphical models (since KL-divergence decomposes into sum of KL-s according to the tree structure). This property can also be useful for contextual bandits and other reinforcement learning problems.

The subsequent sections are organized as follows: Section 2 surveys the main results of the paper and Section 3 discusses the results. All the proofs are provided in the appendix.

2. Main Results

We start with a general concentration result for martingales, which is based on combination of PAC-Bayesian analysis with Bernstein-type inequality for martingales. We apply this result to derive an instantaneous (per-round) generalization bound 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 $\{X_1(h), X_2(h), ...\}$ be martingale difference sequences, meaning that $\mathbb{E}[X_t(h)|\mathcal{T}_{t-1}] = 0$, where $\mathcal{T}_t = \{X_\tau(h)\}_{\substack{1\leq\tau\leq t,\\ h\in\mathcal{H}}}$ is a set of martingale differences observed up to time t. ($\{X_t(h)\}_{h\in\mathcal{H}}$ do not have to be independent, we only need that the requirement on the conditional expectation is satisfied.) Let $M_t(h) = \sum_{\tau=1}^t X_\tau(h)$ be martingales. 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 $M_t(\rho) = \mathbb{E}_{\rho(h)}[M_t(h)]$ and $V_t(\rho) = \mathbb{E}_{\rho(h)}[V_t(h)]$.

Theorem 1 (PAC-Bayes-Bernstein Inequality). *Assume that $|X_t(h)| \leq C$ for all t and h. Let $\{\mu_1, \mu_2, ...\}$ 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 $\{\bar{V}_1, \bar{V}_2, ...\}$ be a sequence of arbitrary numbers, such that $\bar{V}_t$ is independent of $\mathcal{T}_t$ (but can depend on t) and satisfy:*

$$\sqrt{\frac{L_t}{(e-2)\bar{V}_t}} \leq \frac{1}{C}, \tag{1}$$

where

$$L_t = 2\ln(t+1) + \ln\frac{2}{\delta}.$$

Then for all possible distributions ρ_t over $\mathcal{H}$ given t and for all t simultaneously:

$$|M_t(\rho_t)| \leq \sqrt{(e-2)}\left(\begin{array}{r} KL(\rho_t\|\mu_t)\sqrt{\frac{\bar{V}_t}{L_t}} \\ +V_t(\rho_t)\sqrt{\frac{L_t}{\bar{V}_t}} + \sqrt{L_t\bar{V}_t} \end{array}\right). \tag{2}$$

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. Let $\{A_1, A_2, ...\}$ be the sequence of actions played independently at random according to $\{\pi_1, \pi_2, ...\}$ respectively. Let $\{R_1, R_2, ...\}$ be the sequence of observed rewards. Denote by $\mathcal{T}_t = \{\{A_1, .., A_t\}, \{R_1, .., R_t\}\}$ the set of taken actions and observed rewards up to round t (by definition $\mathcal{T}_{t-1} \subset \mathcal{T}_t$).

For $t \geq 1$ and $a \in \{1, .., K\}$ define a set of random variables R_t^a:

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

Define:

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

Observe that $\mathbb{E}\hat{R}_t(a) = R(a)$.

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

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

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

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

be the cumulative variance of this martingale.

Let $\{\varepsilon_1, \varepsilon_2, ...\}$ be a decreasing sequence that satisfies $\varepsilon_t \leq \min_a \pi_t(a)$. In the appendix we prove the following upper bound on $W_t(a)$.

Lemma 1. *For all a:*

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

For a distribution ρ over $\mathcal{A}$ define $\Delta(\rho) = \mathbb{E}_{\rho(a)}[\Delta(a)]$ and $\hat{\Delta}_t(\rho) = \mathbb{E}_{\rho(a)}[\hat{\Delta}_t(a)]$. The following theorem follows immediately from Theorem 1 and Lemma 1 by taking $\bar{V}_t = \frac{2t}{\varepsilon_t}$.

Theorem 2. *For any sequence of sampling distributions $\{\pi_1, \pi_2, ...\}$ that are bounded from below by a decreasing sequence $\{\varepsilon_1, \varepsilon_2, ...\}$ that satisfies*

$$\frac{L_t}{2(e-2)t} \leq \varepsilon_t, \tag{3}$$

where π_t can depend on $\mathcal{T}_{t-1}$, and for any sequence of "reference" distributions $\{\mu_1, \mu_2, ...\}$ over $\mathcal{A}$, such that μ_t is independent of $\mathcal{T}_t$ (but can depend on t), for all possible distributions ρ_t given t and for all $t \geq 1$ simultaneously with probability greater than $1-\delta$:

$$\left|\Delta(\rho_t) - \hat{\Delta}_t(\rho_t)\right| \leq \sqrt{\frac{2(e-2)}{t\varepsilon_t}} \left(\frac{KL(\rho_t \| \mu_t)}{\sqrt{L_t}} + 2\sqrt{L_t}\right). \tag{4}$$

Theorem 2 provides an improvement over the corresponding Theorems 2 and 3 in (Seldin et al., 2011) by decreasing the dependence on ε_t from $1/\varepsilon_t$ to $1/\sqrt{\varepsilon_t}$. This in turn allows to improve the regret bound, which is shown next.

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

$$\pi_{t+1}(a) = \tilde{\rho}_t^{exp}(a) = (1 - K\varepsilon_{t+1})\rho_t^{exp}(a) + \varepsilon_{t+1}, \tag{5}$$

where

$$\rho_t^{exp}(a) = \frac{1}{Z(\rho_t^{exp})} e^{\gamma_t \hat{R}_t(a)} \tag{6}$$

and

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

Then for $t \geq \max\left\{K, K^{4(e-2)}\sqrt{\frac{\delta}{2}}\right\}$ and satisfying (3) (which means that $2\ln(t+1) + \ln\frac{2}{\delta} \leq 2(e-2)\left(\frac{t}{K}\right)^{2/3}$) the per-round regret $R(a^*) - R(\tilde{\rho}_t^{exp})$ is bounded by:

$$R(a^*) - R(\tilde{\rho}_t^{exp}) \leq \frac{K^{1/3}}{(t+1)^{1/3}} \left(\begin{array}{c} (16(e-2)+1)\sqrt{\ln K} \\ +2\sqrt{2(e-2)L_t} + 1 \end{array}\right)$$

with probability greater than $1-\delta$ for all rounds t simultaneously. This translates into a total regret of $\tilde{O}(K^{1/3}t^{2/3})$ (where $\tilde{O}$ hides logarithmic factors).

Theorem 3 improves the dependence on t and K from $\tilde{O}(K^{1/2}t^{3/4})$ in (Seldin et al., 2011) to $\tilde{O}(K^{1/3}t^{2/3})$. This improvement is due to better concentration result in Theorem 2 (which is based on Theorem 1).

We note that there is still room for improvement, which we believe will enable to achieve regret bounds of $\tilde{O}(\sqrt{Kt})$. The main source of looseness is the usage of the crude global upper bound $\frac{2t}{\varepsilon_t}$ on the cumulative variances that holds for any distribution ρ_t. It is possible to show that if we play according to the distributions $\{\tilde{\rho}_1^{exp}, .., \tilde{\rho}_t^{exp}\}$, then for "good" actions a (those for which $\Delta(a) \leq \frac{1}{\gamma_t}$) the cumulative variance $W_t(a)$ 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 $C\varepsilon_t/K$, then the cumulative variance $W_t(\rho_t^{exp})$ would be bounded by CKt. This is, in fact, true for "very bad" actions (those, for which $\Delta(a)$ is close to 1) and it is also possible to show that it holds for μ_t^{exp} (and hence $W_t(\mu_t^{exp}) \leq CKt$), 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 $\rho_t^{exp}(a) \leq C\varepsilon_t/K$ for most of the rounds ($1-\varepsilon_t$ fraction should suffice) and then we will be able to achieve $\tilde{O}(\sqrt{Kt})$ regret. This research direction will be explored in more details in future work.

3. Discussion

We presented an improved PAC-Bayesian analysis of martingales that is based on combination of PAC-Bayesian bound with Bernstein-type inequality for martingales. The new bound enables to provide better finite sample generalization and regret guarantees for exploration-exploitation and model order selection trade-offs simultaneously. There are several important and fascinating research directions that take root at our result.

First, our concentration result for martingales can be of interest in any study of multiple simultaneously evolving and possibly interdependent martingales, especially when the number of martingales is uncountably infinite and standard union bounds cannot be applied. Just as an example, our result can be applied to derive new generalization bounds for active learning (Beygelzimer et al., 2009).

Another important direction is to tighten Theorems 2 and 3, so that the regret bound will match state-of-the-art regret bounds obtained by alternative techniques. We believe that the ideas mentioned at the end of the previous section can make it possible.

Once we have a bound that matches state-of-the-art regret bounds we can extend the technique to richer problems with large or infinite number of states, such as contextual bandits (Beygelzimer et al., 2010), or large or infinite number of actions, such as Gaussian process bandits (Srinivas et al., 2010). Through definition of appropriate priors over hypothesis spaces, PAC-Bayesian approach should enable to obtain bounds that involve natural measures of model complexity, such as mutual information between states and actions in contextual bandits. Such a measure of model complexity is more flexible than plain number of experts or VC-dimension used in (Beygelzimer et al., 2010) since it allows to differentiate between complexities of individual hypotheses. A similar analysis was already performed and proved successful in the context of co-clustering in supervised and unsupervised learning (Seldin & Tishby, 2010).

A. Proofs

In this appendix we provide the proofs of Theorems 1 and 3 and Lemma 1.

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, see the proof of Theorem 1 in (Beygelzimer et al., 2010) for a proof.

Lemma 2 (Bernstein's inequality)**.** *Let $X_1, .., X_t$ be a martingale difference sequence (meaning that $\mathbb{E}[X_\tau | X_1, .., X_{\tau-1}] = 0$ for all τ), such that $X_\tau \leq C$ for all τ. Let $M_t = \sum_{\tau=1}^t X_\tau$ be the corresponding martingale and $V_t = \sum_{\tau=1}^t \mathbb{E}[X_\tau^2 | X_1, .., X_{\tau-1}]$ 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 & Varadhan, 1975; Dupuis & Ellis, 1997; Gray, 2011) and forms the basis of PAC-Bayesian analysis. See (Banerjee, 2006) for a proof.

Lemma 3 (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 $\delta_t = \frac{1}{t(t+1)}\delta \leq \frac{1}{(t+1)^2}\delta$. (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 $1 - \frac{\delta}{2}$:

$$\begin{aligned}
&\lambda_t M_t(\rho_t) - (e-2)\lambda_t^2 V_t(\rho_t) \\
&\quad = \mathbb{E}_{\rho_t(h)}[\lambda_t M_t(h) - (e-2)\lambda_t^2 V_t(h)] && (7)\\
&\quad \leq KL(\rho_t \| \mu_t) + \ln \mathbb{E}_{\mu_t(h)}[e^{\lambda_t M_t(\mu_t) - (e-2)\lambda_t^2 V_t(\mu_t)}] && (8)\\
&\quad \leq KL(\rho_t \| \mu_t) + 2\ln(t+1) + \ln\frac{2}{\delta} \\
&\qquad + \ln \mathbb{E}_{\mathcal{T}_t}\mathbb{E}_{\mu_t(h)}[e^{\lambda_t M_t(h) - (e-2)\lambda_t^2 V_t(h)}] && (9)\\
&\quad = KL(\rho_t \| \mu_t) + L_t \\
&\qquad + \ln \mathbb{E}_{\mu_t(h)}\mathbb{E}_{\mathcal{T}_t}[e^{\lambda_t M_t(h) - (e-2)\lambda_t^2 V_t(h)}] && (10)\\
&\quad \leq KL(\rho_t \| \mu_t) + L_t, && (11)
\end{aligned}$$

where (7) is by definition of $M_t(\rho_t)$ and $V_t(\rho_t)$, (8) is by Lemma 3, (9) holds with probability greater than $1 - \frac{\delta}{2}$ by Markov's inequality and a union bound over t, (10) is due to the fact that μ_t is independent of $\mathcal{T}_t$ and by definition of L_t, and (11) is by Lemma 2.

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 $1 - \delta$:

$$|M_t(\rho_t)| \leq \frac{KL(\rho_t \| \mu_t) + (e-2)\lambda_t^2 V_t(\rho_t) + L_t}{\lambda_t}. \quad (12)$$

By taking

$$\lambda_t = \sqrt{\frac{L_t}{(e-2)\bar{V}_t}}$$

and substituting into (12) we obtain (2). The technical condition (1) follows from the requirement that $\lambda_t \in [0, \frac{1}{C}]$. □

A.2. Proof of Lemma 1

Proof of Lemma 1.

$$\begin{aligned}
W_t(a) &= \sum_{\tau=1}^t \mathbb{E}[([R_\tau^{a^*} - R_\tau^a] - [R(a^*) - R(a)])^2 | \mathcal{T}_{\tau-1}] \\
&= \left(\sum_{\tau=1}^t \mathbb{E}[(R_\tau^{a^*} - R_\tau^a)^2 | \mathcal{T}_{\tau-1}]\right) - t\Delta(a)^2 && (13)\\
&\leq \left(\sum_{\tau=1}^t \left(\frac{\pi_\tau(a)}{\pi_\tau(a)^2} + \frac{\pi_\tau(a^*)}{\pi_\tau(a^*)^2}\right)\right) - t\Delta(a)^2 && (14)\\
&= \left(\sum_{\tau=1}^t \left(\frac{1}{\pi_\tau(a)} + \frac{1}{\pi_\tau(a^*)}\right)\right) - t\Delta(a)^2 \\
&\leq \frac{2t}{\varepsilon_t}, && (15)
\end{aligned}$$

where (13) is due to the fact that $\mathbb{E}[R_\tau^a | \mathcal{T}_{\tau-1}] = R(a)$, (14) is due to the fact that $R_t \leq 1$ and (15) 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 3

Proof of Theorem 3. We take the same prior $\mu_t(a)$ that was used in (Seldin et al., 2011)

$$\mu_t^{exp}(a) = \frac{1}{Z(\mu_t^{exp})} e^{\gamma_t R(a)}, \tag{16}$$

where $Z(\mu_t^{exp}) = \sum_a e^{\gamma_t R(a)}$ is the normalization factor.

We reuse the same regret decomposition we had in (Seldin et al., 2011), but write it in a new form using Δ-s:

$$\begin{aligned}\Delta(\tilde{\rho}_t^{exp}) &= \Delta(\rho_t^{exp}) + [R(\rho_t^{exp}) - R(\tilde{\rho}_t^{exp})]\\ &\leq [\Delta(\rho_t^{exp}) - \hat{\Delta}_t(\rho_t^{exp})] + \hat{\Delta}_t(\rho_t^{exp}) + K\varepsilon_{t+1} \qquad (17)\\ &\leq [\Delta(\rho_t^{exp}) - \hat{\Delta}_t(\rho_t^{exp})] + \frac{\ln K}{\gamma_t} + K\varepsilon_{t+1}, \qquad (18)\end{aligned}$$

where in (17) we used the bound on $[R(\rho_t^{exp}) - R(\tilde{\rho}_t^{exp})]$ obtained in (Seldin et al., 2011) and in (18) we used Lemma 4 given below. Note that due to working with Δ-s we are left to bound only one term instead of two terms we had to bound in (Seldin et al., 2011).

Lemma 4. *Let $x_1 = 0$ and $x_2, .., x_n$ be $n-1$ arbitrary numbers. For any $\alpha > 0$ and $n \geq 2$:*

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

Proof. Since negative x_i-s only decrease the left hand side of (19) 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:

$$\frac{\sum_{i=1}^n x_i e^{-\alpha x_i}}{\sum_{j=1}^n e^{-\alpha x_j}} \leq \max_x \frac{(n-1)xe^{-\alpha x}}{1 + (n-1)e^{-\alpha x}}. \tag{20}$$

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 (20) we get

$$\frac{1}{\alpha} \cdot \frac{(n-1)y \ln \frac{1}{y}}{1 + (n-1)y}.$$

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 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:

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

where the last inequality follows from the fact that $\frac{n-1}{n} \leq \ln n$ for $n \geq 2$. □

In order to obtain an explicit bound on $[\Delta(\rho_t) - \hat{\Delta}_t(\rho_t)]$ we need an explicit bound on $KL(\rho_t^{exp} \| \mu_t^{exp})$. To obtain such a bound we modify the procedure that was used in (Seldin et al., 2011), which in turn was based on the procedure developed by Lever et al. (2010). Due to tighter concentration inequality in Theorem 1 we obtain a tighter bound on $KL(\rho_t^{exp} \| \mu_t^{exp})$.

The derivation procedure starts with the following lemma, which is proved similarly to Lemma 12 in (Seldin et al., 2011).

Lemma 5. *For μ_t^{exp} and ρ_t^{exp} defined by (16) and (6):*

$$KL(\rho_t^{exp} \| \mu_t^{exp}) \leq \gamma_t \left(\begin{array}{c} [\Delta(\rho_t^{exp}) - \hat{\Delta}_t(\rho_t^{exp})] \\ +[\hat{\Delta}_t(\mu_t^{exp}) - \Delta(\mu_t^{exp})] \end{array} \right).$$

Proof. We use the following definitions:

$$\begin{aligned}Z'(\mu_t^{exp}) &= \sum_a e^{-\gamma_t \Delta(a)}\\ &= \sum_a e^{-\gamma_t (R(a^*) - R(a))}\\ &= e^{-\gamma_t R(a^*)} Z(\mu_t^{exp}).\end{aligned}$$

$$\begin{aligned}Z'(\rho_t^{exp}) &= \sum_a e^{-\gamma_t \hat{\Delta}_t(a)}\\ &= \sum_a e^{-\gamma_t (\hat{R}_t(a^*) - \hat{R}_t(a))}\\ &= e^{-\gamma_t \hat{R}_t(a^*)} Z(\rho_t^{exp}).\end{aligned}$$

The following identity is easily verified from the definitions:

$$\begin{aligned}\frac{1}{Z'(\mu_t^{exp})} &= \frac{1}{Z(\mu_t^{exp})} e^{\gamma_t R(a^*)}\\ &= \mu_t(a) e^{-\gamma_t R(a)} e^{\gamma_t R(a^*)}\\ &= \mu_t(a) e^{\gamma_t \Delta(a)}.\end{aligned}$$

Now we have:

$$
\begin{aligned}
&KL(\rho_t^{exp}\|\mu_t^{exp}) = \sum_a \rho_t(a)\ln\frac{e^{\gamma_t\hat{R}_t(a)}Z(\mu_t^{exp})}{e^{\gamma_t R(a)}Z(\rho_t^{exp})}\\
&= \sum_a \rho_t(a)\ln\frac{e^{-\gamma_t\hat{\Delta}_t(a)}Z'(\mu_t^{exp})}{e^{-\gamma_t\Delta(a)}Z'(\rho_t^{exp})}\\
&= \gamma_t[\Delta(\rho_t^{exp}) - \hat{\Delta}_t(\rho_t^{exp})] - \ln\frac{\sum_a e^{-\gamma_t\hat{\Delta}_t(a)}}{Z'(\mu_t^{exp})}\\
&= \gamma_t[\Delta(\rho_t^{exp}) - \hat{\Delta}_t(\rho_t^{exp})] - \ln\sum_a \mu_t^{exp}(a)e^{\gamma_t(\Delta(a)-\hat{\Delta}_t(a))}\\
&\leq \gamma_t\left([\Delta(\rho_t^{exp}) - \hat{\Delta}_t(\rho_t^{exp})] + [\hat{\Delta}_t(\mu_t^{exp}) - \Delta(\mu_t^{exp})]\right).
\end{aligned}
$$

□

Now we want to get an explicit upper bound on $KL(\rho_t^{exp}\|\mu_t^{exp})$. Note that for our choice of ε_t the technical condition (3) of Theorem 2 is satisfied by t large enough, so that

$$2\ln(t+1) + \ln\frac{2}{\delta} \leq 2(e-2)\left(\frac{t}{K}\right)^{2/3}.$$

(This requirement is satisfied by $t = O\left(K\left(\ln\frac{1}{\delta}\right)^{3/2}\right)$.) By Theorem 2 with probability greater than $1-\delta$:

$$
\begin{aligned}
&\Delta(\rho_t^{exp}) - \hat{\Delta}_t(\rho_t^{exp})\\
&\qquad \leq \sqrt{\frac{2(e-2)}{t\varepsilon_t}}\left(\frac{KL(\rho_t^{exp}\|\mu_t^{exp})}{\sqrt{L_t}} + 2\sqrt{L_t}\right)
\end{aligned}
\tag{21}
$$

and

$$\hat{\Delta}_t(\mu_t^{exp}) - \Delta(\mu_t^{exp}) \leq 2\sqrt{\frac{2(e-2)L_t}{t\varepsilon_t}}.$$

By substituting this into Lemma 5 we obtain:

$$
\begin{aligned}
&KL(\rho_t^{exp}\|\mu_t^{exp})\\
&\qquad \leq \gamma_t\sqrt{\frac{2(e-2)}{t\varepsilon_t}}\left(\frac{KL(\rho_t^{exp}\|\mu_t^{exp})}{\sqrt{L_t}} + 4\sqrt{L_t}\right).
\end{aligned}
$$

By reorganizing the terms:

$$KL(\rho_t^{exp}\|\mu_t^{exp})\left(1 - \gamma_t\sqrt{\frac{2(e-2)}{t\varepsilon_t L_t}}\right) \leq 4\gamma_t\sqrt{\frac{2(e-2)L_t}{t\varepsilon_t}}. \tag{22}$$

Note that for our choice of γ_t and ε_t:

$$\gamma_t\sqrt{\frac{2(e-2)}{t\varepsilon_t L_t}} = \sqrt{\frac{2(e-2)K}{2\ln(t+1)+\ln\frac{2}{\delta}}}.$$

By simple algebraic manipulations we obtain that

$$\gamma_t\sqrt{\frac{2(e-2)}{t\varepsilon_t L_t}} \leq \frac{1}{2} \tag{23}$$

for

$$t \geq K^{4(e-2)}\sqrt{\frac{\delta}{2}}.$$

By substituting (23) into (22) we obtain that:

$$KL(\rho_t^{exp}\|\mu_t^{exp}) \leq 8\gamma_t\sqrt{\frac{2(e-2)L_t}{t\varepsilon_t}}.$$

By substituting this into (21) we obtain

$$
\begin{aligned}
&\Delta(\rho_t^{exp}) - \hat{\Delta}_t(\rho_t^{exp})\\
&\qquad \leq \sqrt{\frac{2(e-2)}{t\varepsilon_t}}\left(8\gamma_t\sqrt{\frac{2(e-2)}{t\varepsilon_t}} + 2\sqrt{L_t}\right).
\end{aligned}
$$

For our choice of γ_t and ε_t:

$$\Delta(\rho_t^{exp}) - \hat{\Delta}_t(\rho_t^{exp}) \leq \frac{K^{1/3}}{t^{1/3}}\left(\begin{array}{c}16(e-2)\sqrt{\ln K}\\ +2\sqrt{2(e-2)L_t}\end{array}\right)$$

Substitution of the result into (18) concludes the proof.

□

Acknowledgments

We thank Ronald Ortner for useful discussions.

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 №231495. This publication only reflects the authors' views.

References

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

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

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

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

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

Beygelzimer, Alina, Langford, John, Li, Lihong, Reyzin, Lev, and Schapire, Robert E. Contextual bandit algorithms with supervised learning guarantees. http://arxiv.org/abs/1002.4058, 2010.

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

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

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

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

Gray, Robert M. *Entropy and Information Theory.* Springer, 2 edition, 2011.

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

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

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

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

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

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

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

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

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

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

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

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

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

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