PAC-Bayesian Analysis of Martingales and Multiarmed Bandits

Yevgeny Seldin
Max Planck Institute
Tübingen, Germany
seldin@tuebingen.mpg.de

François Laviolette
Université Laval
Québec, Canada
francois.laviolette@ift.ulaval.ca

John Shawe-Taylor
University College London
jst@cs.ucl.ac.uk

Jan Peters
Max Planck Institute
Tübingen, Germany
jan.peters@tuebingen.mpg.de

Peter Auer
Chair for Information Technology
University of Leoben, Austria
auer@unileoben.ac.at

Abstract

We present two alternative ways to apply PAC-Bayesian analysis to sequences of dependent random variables. The first is based on a new lemma that enables to bound expectations of convex functions of certain dependent random variables by expectations of the same functions of independent Bernoulli random variables. This lemma provides an alternative tool to Hoeffding-Azuma inequality to bound concentration of martingale values. Our second approach is based on integration of Hoeffding-Azuma inequality with PAC-Bayesian analysis. We also introduce a way to apply PAC-Bayesian analysis in situation of limited feedback. We combine the new tools to derive PAC-Bayesian generalization and regret bounds for the multiarmed bandit problem. Although our regret bound is not yet as tight as state-of-the-art regret bounds based on other well-established techniques, our results significantly expand the range of potential applications of PAC-Bayesian analysis and introduce a new analysis tool to reinforcement learning and many other fields, where martingales and limited feedback are encountered.

1 Introduction

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. The power of PAC-Bayesian approach lies in the successful marriage of 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 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).

Even within supervised learning the applications of PAC-Bayesian analysis were restricted to i.i.d. data for a long time. The issue of treating non-independent samples was partially addressed only recently by Ralaivola et al. (2010) and Lever et al. (2010) (their approaches are also suitable for density estimation (Higgs and Shawe-Taylor, 2010). The solution of Ralaivola et al. (2010) essentially boils down to 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 to martingales due to strong dependence of the cumulative sum on all of its components. Lever et al. (2010) employed Hoeffding's canonical decomposition of U-statistics into forward martingales and applied PAC-Bayesian analysis directly to these martingales. Our second approach to handling sequences of dependent samples by combining PAC-Bayesian analysis with Hoeffding-Azuma inequality is based on similar ideas. Our first approach to sequences of dependent samples

is based on the new lemma that allows to bound expectations of functions of certain sequentially dependent random variables by expectations of the same functions of independent random variables.

One of the most prominent and important fields of application of martingales is reinforcement learning. 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 (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 therefore can be 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 and Pineau (2010) suggested a PAC-Bayesian analysis of batch reinforcement learning. They used the analysis to design an algorithm that is able to leverage the prior knowledge when it is informative and confirms the data distribution and ignores it when it is irrelevant. In the first case Bayesian learning algorithms perform well and in the second case PAC learning algorithms perform better, whereas Fard and Pineau showed that their algorithm performs on par with the best out of the two in all situations. However, the analysis of Fard and Pineau does not address the exploration-exploitation trade-off, which is the key feature of reinforcement learning. In their batch analysis they assume that every action was sampled in every state some minimal number of times and the bound decreases at the rate of a square root of the minimum over states and actions of the number of times an action was sampled in a state. Clearly, such an analysis is not applicable in online setting, since we do not want to sample "bad" actions many times, but then the bound does not improve with time.

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 (that is the reason why the bound of Fard and Pineau (2010) depends on the minimum of the number of times any action was taken in any state, which is the minimal sample size available for evaluation of all state-action pairs). In online setting the sample size of "bad" actions has to increase sublinearly in the number of game rounds, which results in slow or even no convergence of the bound. We resolve this issue by applying weighted sampling strategy (Sutton and Barto, 1998), which is commonly used in the analysis of non-stochastic bandits (Auer et al., 2002b), but has not been applied to the analysis of stochastic bandits previously.

The usage of weighted sampling introduces two new difficulties. One is the dependence between 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. We handle this dependence using our new PAC-Bayesian approaches to sequences of dependent variables. At the moment both approaches yield comparable bounds, however each of the approaches has its own potential advantages that can be exploited in future work.

The second problem introduced by weighted sampling is the growing variance of the weighted sample variables. Martingale bounding techniques used in this work do not enable to take full control over the variance, which explains the gap between our results and state-of-the-art bounds for mutliarmed bandits (Auer et al., 2002a, Auer and Ortner, 2010). Tighter bounds can be achieved by combining PAC-Bayesian analysis with Bernstein-type inequality for martingales (Beygelzimer et al., 2010). Such a combination will be presented in future work.

The subsequent sections are organized as follows: Section 2 surveys the main results of the paper, Section 3 presents our bound on expectation of convex functions of sequentially dependent random variables and illustrates its application to derivation of an alternative to Hoeffding-Azuma inequality, Section 4 provides a PAC-Bayesian analysis of the weighted sampling strategy based on the bound from Section 3, Section 5 provides PAC-Bayesian analysis of the weighted sampling strategy based on martingales, Section 6 derives a regret bound for the multiarmed bandit problem, and Section 7 concludes the results.

2 Main Results

One of the foundation stones of our paper is the following lemma that enables to bound expectations of convex functions of certain sequentially dependent random variables by expectations of the same functions of independent Bernoulli random variables. The lemma generalizes a preceding result of Maurer (2004) for independent random variables and might have a wide interest on its own right far

beyond the PAC-Bayesian analysis. The lemma can be used to derive an alternative to Hoeffding-Azuma inequality (Hoeffding, 1963, Azuma, 1967). This alternative can be much tighter in certain situations (see our derivation and discussion of Lemma 7 in the next section).

Lemma 1 *Let $X_1, .., X_N$ be dependent random variables belonging to the $[0,1]$ interval and distributed by $p(x_i|X_1, .., X_{i-1})$, such that $\mathbb{E}[X_i|X_1, .., X_{i-1}] = p$ for all i. Let $Y_1, .., Y_N$ be independent Bernoulli random variables, such that $\mathbb{E}Y_i = p$ for all i. Then for any convex function $f : [0,1]^N \to \mathbb{R}$:*

$$\mathbb{E}f(X_1, .., X_N) \leq \mathbb{E}f(Y_1, .., Y_N).$$

We present the subsequent results in the context of the multiarmed bandit problem, which is probably the most common problem in machine learning, where sequentially dependent variables are encountered. 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 indicator random variables $\{I_t^a\}_{t,a}$:

$$I_t^a = \begin{cases} 1, & \text{if } A_t = a \\ 0, & \text{otherwise.} \end{cases}$$

Define a set of random variables $R_t^a = \frac{1}{\pi_t(a)} I_t^a R_t$. In other words:

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

Define: $\hat{R}_t(a) = \frac{1}{t} \sum_{\tau=1}^{t} R_\tau^a$. For a distribution ρ over $\mathcal{A}$ define $R(\rho) = \mathbb{E}_{\rho(a)} R(a)$ and $\hat{R}_t(\rho) = \mathbb{E}_{\rho(a)} \hat{R}_t(a)$.

For two distributions ρ and μ, let $KL(\rho\|\mu)$ denote the KL-divergence between ρ and μ. For two Bernoulli random variables with biases p and q let $kl(p\|q) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}$ be an abbreviation for $KL([p, 1-p]\|[q, 1-q])$.

We present two alternative results, the first applies Lemma 1 to handle sequences of dependent random variables and the second is based on combination of PAC-Bayesian analysis with Hoeffding-Azuma inequality. Then we compare the results and present a regret bound for the multiarmed bandit problem based on the first solution.

2.1 PAC-Bayesian Analysis of Sequentially Dependent Variables Based on Lemma 1

Our first PAC-Bayesian theorem provides a bound on the divergence between $\hat{R}_t(\rho_t)$ and $R(\rho_t)$ for any playing strategy ρ_t throughout the game.

Theorem 2 *For any sequence of sampling distributions $\{\pi_1, \pi_2, ...\}$ that are not zero for any $a \in \mathcal{A}$, where π_t can depend on $\mathcal{T}_{t-1}$, and for any sequence of "reference" ("prior") 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 $1 - \delta$:*

$$kl(\pi_t^{lmin} \hat{R}_t(\rho_t) \| \pi_t^{lmin} R(\rho_t)) \leq \frac{KL(\rho_t\|\mu_t) + 3\ln(t+1) - \ln\delta}{t}, \tag{1}$$

where

$$\pi_t^{lmin} \leq \min_{\substack{a, \\ 1 \leq \tau \leq t}} \pi_\tau(a).$$

The number π_t^{lmin} lower bounds sampling probabilities for all the actions up to time t ($lmin$ stands for "left minimum" or minimum of $\pi_\tau(a)$ up to ["left to"] time t).

The KL-divergence $kl(p\|q)$ bounds the absolute difference between p and q as

$$|p - q| \leq \sqrt{kl(p\|q)/2} \tag{2}$$

(Cover and Thomas, 1991). Combined with (1) this relation yields (with probability greater than $1 - \delta$):

$$\left| R(\rho_t) - \hat{R}_t(\rho_t) \right| \leq \frac{1}{\pi_t^{lmin}} \sqrt{\frac{KL(\rho_t\|\mu_t) + 3\ln(t+1) - \ln\delta}{2t}}. \tag{3}$$

2.2 Combination of PAC-Bayesian Analysis with Hoeffding-Azuma Inequality

The result presented next is based on a combination of PAC-Bayesian analysis with Hoeffding-Azuma inequality. We introduce one more definition:

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

where $w_\tau^t \geq 0$ for all t and τ and $\sum_{\tau=1}^t w_\tau^t > 0$ for all t. $\hat{R}_t^{w^t}(a)$ is a weighted sum of the samples. For a special case, where $w_\tau^t = \frac{1}{t}$ for all τ, $\hat{R}_t^{w^t}(a) = \hat{R}_t(a)$.

Theorem 3 *For any sequence of sampling distributions $\{\pi_1, \pi_2, ...\}$ that are not zero for any $a \in \mathcal{A}$, where π_t can depend on $\mathcal{T}_{t-1}$, and for any sequence of "reference" ("prior") distributions $\{\mu_1, \mu_2, ...\}$ over $\mathcal{A}$, such that μ_t is independent of $\mathcal{T}_t$ (but can depend on t), for any sequence of positive parameters $\{\lambda_1, \lambda_2, ...\}$ and for any sequence of weighting vectors $\{w^1, w^2, ...\}$, such that λ_t and w^t are 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 $1 - \delta$:*

$$\left|\hat{R}_t^{w^t}(a) - R(a)\right| \leq \frac{KL(\rho_t\|\mu_t) + \frac{1}{2}\lambda_t^2 \sum_{\tau=1}^t \left(\frac{w_\tau^t}{\pi_\tau^{min}}\right)^2 + 2\ln(t+1) + \ln\frac{2}{\delta}}{\lambda_t \sum_{\tau=1}^t w_\tau^t}, \tag{4}$$

where

$$\pi_t^{min} \leq \min_a \pi_t(a).$$

For the special case $w_\tau^t = \frac{1}{t}$ we obtain that with probability greater than $1 - \delta$:

$$\left|\hat{R}_t(a) - R(a)\right| \leq \frac{KL(\rho_t\|\mu_t) + \frac{1}{2}\frac{\lambda_t^2}{t^2} \sum_{\tau=1}^t \frac{1}{(\pi_\tau^{min})^2} + 2\ln(t+1) + \ln\frac{2}{\delta}}{\lambda_t}. \tag{5}$$

By taking

$$\lambda_t = \sqrt{2t^2 \left(2\ln(t+1) + \ln\frac{2}{\delta}\right) / \left(\sum_{\tau=1}^t \frac{1}{(\pi_\tau^{min})^2}\right)}$$

we obtain:

$$\left|\hat{R}_t(a) - R(a)\right| \leq \sqrt{\frac{\frac{1}{t}\left(\sum_{\tau=1}^t \frac{1}{(\pi_\tau^{min})^2}\right)}{2t}} \left(\frac{KL(\rho_t\|\mu_t)}{\sqrt{\ln(t+1) + \ln\frac{2}{\delta}}} + \sqrt{\ln(t+1) + \ln\frac{2}{\delta}}\right). \tag{6}$$

2.3 Comparison of Theorem 2 with Theorem 3

It is interesting to compare Theorems 2 and 3 resulting from the two different approaches. Inequality (3) depends on $\frac{1}{\pi_t^{lmin}} = \max_{1\leq\tau\leq t}\left\{\frac{1}{\pi_\tau^{min}}\right\}$, whereas (6) depends on $\sqrt{\frac{1}{t}\sum_{\tau=1}^t \frac{1}{(\pi_\tau^{min})^2}}$. If π_τ^{min} are approximately equal for all τ, then the two terms are approximately identical. However, a single small value of π_τ^{min} can increase the value of $\frac{1}{\pi_t^{lmin}}$ significantly for all $t \geq \tau$, while its relative contribution to the average of $\frac{1}{(\pi_\tau^{min})^2}$ will decrease with time. This property provides an advantage to Theorem 3. On the other hand, the stronger *kl* form (1) of Theorem 2 can potentially be an advantage for the bound based on Lemma 1, but we did not exploit it in this work.

Since for our choice of sampling strategy $\frac{1}{\pi_t^{lmin}} \approx \sqrt{\frac{1}{t}\sum_{\tau=1}^t \frac{1}{(\pi_\tau^{min})^2}}$ up to small constants, we present a regret bound based on Theorem 2 only. A regret bound based on Theorem 3 can be derived in a similar way and is identical to the bound presented below up to small constants.

2.4 Regret Bound for Multiarmed Bandits

We applied Theorem 2 to derive the following regret bound for the multiarmed bandit problem.

Theorem 4 *For $t < K^3$ let $\pi_t(a) = \frac{1}{K}$ for all a. Let $\gamma_t = K^{1/4}t^{1/4}$ and $\varepsilon_t = K^{-1/4}t^{-1/4}$ and for $t \geq (K^3 - 1)$ let*

$$\pi_{t+1}(a) = \tilde{\rho}_t^{exp}(a) = (1 - K\varepsilon_{t+1})\rho_t^{exp}(a) + \varepsilon_{t+1}, \tag{7}$$

where

$$\rho_t^{exp}(a) = \frac{1}{Z(\rho_t^{exp})} e^{\gamma_t \hat{R}_t(a)} \tag{8}$$

and

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

Then for $t \geq K^3$ the per-round regret $R(a^*) - R(\tilde{\rho}_t^{exp})$ (where a^* is the best action) is bounded by:

$$R(a^*) - R(\tilde{\rho}_t^{exp}) \leq \frac{K^{3/4}}{(t+1)^{1/4}} \left(2.5 + \sqrt{\frac{\ln(K) + 3\ln(t+1) - \ln\delta}{2K}} + \sqrt{\frac{3\ln(t+1) - \ln\delta}{2K}} \right)$$

with probability greater than $1-\delta$ for all rounds t simultaneously. This translates into a total regret of $\tilde{O}(K^{3/4}t^{3/4})$ (where $\tilde{O}$ hides logarithmic factors).

Note that ε_t bounds $\pi_t(a)$ from below for all a and $t \geq K^3$. Furthermore, since ε_t is a decreasing sequence it actually bounds $\pi_\tau(a)$ from below for all a and $\tau \leq t$. Hence, for the prediction strategy selected in Theorem 4 and for $t \geq K^3$ we can substitute π_t^{lmin} with ε_t in (1) and (3).

3 Proof of Lemma 1 and an Example of its Application

We start with the proof of Lemma 1 and then illustrate how it can be applied to martingales.

Proof of Lemma 1: The proof follows the lines of the proof of Lemma 3 in Maurer (2004). Any point $\bar{x} = (x_1, .., x_N) \in [0,1]^N$ can be written as a convex combination of the extreme points $\bar{\eta} = (\eta_1, .., \eta_N) \in \{0,1\}^N$ in the following way:

$$\bar{x} = \sum_{\bar{\eta} \in \{0,1\}^N} \left(\prod_{i:\eta_i=0} (1 - x_i) \prod_{i:\eta_i=1} x_i \right) \bar{\eta}.$$

Convexity of f therefore implies

$$f(\bar{x}) \leq \sum_{\bar{\eta} \in \{0,1\}^N} \left(\prod_{i:\eta_i=0} (1 - x_i) \prod_{i:\eta_i=1} x_i \right) f(\bar{\eta}), \tag{9}$$

with equality if $\bar{x} \in \{0,1\}^N$. At the next step Maurer (2004) uses independence of X_i-s, whereas we use the fact that their conditional expectation is constant. Taking expectation of both sides of (9) we obtain:

$$\begin{aligned}
\mathbb{E}_{X_1,..,X_N}[f(\bar{X})] &\leq \mathbb{E}_{X_1,..,X_N} \left[\sum_{\bar{\eta} \in \{0,1\}^N} \left(\prod_{i:\eta_i=0} (1 - X_i) \prod_{i:\eta_i=1} X_i \right) f(\bar{\eta}) \right] \\
&= \sum_{\bar{\eta} \in \{0,1\}^N} \mathbb{E}_{X_1,..,X_N} \left[\left(\prod_{i:\eta_i=0} (1 - X_i) \prod_{i:\eta_i=1} X_i \right) \right] f(\bar{\eta}) \\
&= \sum_{\bar{\eta} \in \{0,1\}^N} \mathbb{E}_{X_1,..,X_{N-1}} \left[\mathbb{E}_{X_N} \left[\left(\prod_{i:\eta_i=0} (1 - X_i) \prod_{i:\eta_i=1} X_i \right) \middle| X_1, .., X_{N-1} \right] \right] f(\bar{\eta}) \\
&= \sum_{\bar{\eta} \in \{0,1\}^N} \mathbb{E}_{X_1,..,X_{N-1}} \left[\begin{array}{l} \left(\prod_{i:\eta_i=0, i<N} (1 - X_i) \prod_{i:\eta_i=1, i<N} X_i \right) \\ \cdot \mathbb{E}_{X_N} \left[\mathbb{I}(\eta_N = 0)(1 - X_N) \mathbb{I}(\eta_N = 1) X_N | X_1, .., X_{N-1} \right] \end{array} \right] f(\bar{\eta}) \qquad (10) \\
&= \sum_{\bar{\eta} \in \{0,1\}^N} \mathbb{E}_{X_1,..,X_{N-1}} \left[\left(\prod_{i:\eta_i=0, i<N} (1 - X_i) \prod_{i:\eta_i=1, i<N} X_i \right) \cdot \left[\mathbb{I}(\eta_N = 0)(1-p) \mathbb{I}(\eta_N = 1) p \right] \right] f(\bar{\eta}) \\
&= ... \qquad (11) \\
&= \sum_{\bar{\eta} \in \{0,1\}^N} \left(\prod_{i:\eta_i=0} (1 - p) \prod_{i:\eta_i=1} p \right) f(\bar{\eta}) \\
&= \mathbb{E}_{Y_1,..,Y_N}[f(\bar{Y})].
\end{aligned}$$

In (10) $\mathbb{I}$ is the indicator function (note that only one of $\mathbb{I}(\eta_N = 0)$ and $\mathbb{I}(\eta_N = 1)$ is 1 and the other one is 0). In (11) we apply induction in order to replace X_i-s by p, one-by-one from the last to the first, same way we did it for X_N. ■

3.1 Application to Martingales

We apply Lemma 1 to derive an alternative to Hoeffding-Azuma inequality. The derivation is based on Markov's inequality and a concentration result for independent Bernoulli variables provided below.

Lemma 5 (Markov's inequality) *For a random variable $X \geq 0$ with probability greater than $1 - \delta$:*

$$X \leq \frac{1}{\delta}\mathbb{E}X. \tag{12}$$

The concentration result for independent Bernoulli variables is based on the method of types in information theory (Cover and Thomas, 1991). Its proof can be found in Seeger (2003),Banerjee (2006), or Seldin and Tishby (2010).[1]

Lemma 6 *Let $X_1, .., X_N$ be i.i.d. Bernoulli random variables. Let $\hat{S} = \frac{1}{N}\sum_{i=1}^{N} X_i$ be their empirical average and $S = \mathbb{E}X_i$ the expected value. Then:*

$$\mathbb{E}_{X_1,..,X_N}[e^{Nkl(\hat{S}\|S)}] \leq N + 1. \tag{13}$$

Since KL-divergence is a convex function (Cover and Thomas, 1991) and exponent is convex and non-decreasing, $e^{Nkl(\hat{S}\|S)}$ is also a convex function. Therefore, by Lemma 1 we obtain that Lemma 6 also holds for $X_1, .., X_N$ that belong to the $[0, 1]$ interval and are sequentially dependent on each other as long as their conditional expectation $\mathbb{E}[X_i|X_1, .., X_{i-1}]$ is identical.

Alternative to Hoeffding-Azuma Inequality Based on Lemmas 1 and 6

Now we are ready to present our alternative to Hoeffding-Azuma's inequality.

Lemma 7 *Let $X_1, .., X_N$ be a martingale difference sequence (meaning that $\mathbb{E}[X_i|X_1, .., X_{i-1}] = 0$), such that $X_i \in [a_i, b_i]$ for an arbitrary $a_i \leq 0$ and $b_i \geq 0$. Let $S_1, .., S_N$ be a martingale, where $S_j = \sum_{i=1}^{j} X_i$. Let $a = \min_i a_i$ and $b = \max_i b_i$ and let $Z_i = (X_i - a)/(b - a)$. Then with probability greater than $1 - \delta$ the following holds simultaneously:*

$$kl\left(\frac{1}{N}\sum_{i=1}^{N} Z_i \| \frac{-a}{b-a}\right) \leq \frac{\ln \frac{N+1}{\delta}}{N} \tag{14}$$

and

$$|S_N| \leq (b - a)\sqrt{\frac{1}{2}N \ln \frac{N + 1}{\delta}}. \tag{15}$$

Proof of Lemma 7: By definition of Z_i we have $Z_i \in [0, 1]$ and $\mathbb{E}[Z_i|Z_1, .., Z_{i-1}] = \frac{-a}{b-a}$ is identical for all Z_i. Hence, by Markov's inequality and combination of Lemma 1 with Lemma 6 with probability greater than $1 - \delta$:

$$e^{Nkl(\frac{1}{N}\sum_{i=1}^{N} Z_i \| \frac{-a}{b-a})} \leq \frac{1}{\delta}\mathbb{E}_{Z_1,..,Z_N}[e^{Nkl(\frac{1}{N}\sum_{i=1}^{N} Z_i \| \frac{-a}{b-a})}] \leq \frac{N + 1}{\delta}.$$

Taking logarithm and normalizing by N yields (14).

By relation (2) between L_1-norm and KL-divergence (14) yields:

$$\left|\frac{1}{N}\sum_{i=1}^{N} Z_i - \frac{-a}{b-a}\right| \leq \sqrt{\frac{\ln \frac{N+1}{\delta}}{2N}}.$$

From definitions, $X_i = (b-a)Z_i + a$ and $S_N = (b-a)\sum_{i=1}^{N} Z_i + Na$. Simple algebraic manipulations yield (15). ■

[1] It is possible to prove even stronger result of a form $\sqrt{N} \leq \mathbb{E}_{X_1,..,X_N} e^{Nkl(\hat{S}\|S)} \leq 2\sqrt{N}$ for $N \geq 8$ using Stirling's approximation of the factorial (Maurer, 2004). For simplicity we use (13).

Comparison with Hoeffding-Azuma Inequality

It is instructive to compare Lemma 7 with Hoeffding-Azuma inequality, which we cite below for the comparison (Azuma, 1967, Cesa-Bianchi and Lugosi, 2006).

Lemma 8 (Hoeffding-Azuma Inequality) *Let $S_1, .., S_N$ be a zero-mean martingale satisfying $S_i - S_{i-1} \in [a_i, b_i]$, then for any $\lambda > 0$:*

$$\mathbb{E}[e^{\lambda S_N}] \leq e^{(\lambda^2/8)\sum_{i=1}^{N}(b_i-a_i)^2}.$$

It is easy to verify, using the same procedure we applied before, that Lemma 8 implies that with probability greater than $1-\delta$:

$$|S_N| \leq \frac{\frac{1}{8}\lambda^2 \sum_{i=1}^{N}(b_i-a_i)^2 + \ln\frac{2}{\delta}}{\lambda}$$

and that the above expression is minimized by $\lambda = \sqrt{8\ln\frac{2}{\delta}/\sum_{i=1}^{N}(b_i-a_i)}$ yielding:

$$|S_N| \leq \sqrt{\frac{1}{2}\left(\sum_{i=1}^{N}(b_i-a_i)^2\right)\ln\frac{2}{\delta}}. \tag{16}$$

In a special case, where $a_i = a$ for all i and $b_i = b$ for all i, this further simplifies to:

$$|S_N| \leq (b-a)\sqrt{\frac{1}{2}N\ln\frac{2}{\delta}}.$$

Now we are ready to make the comparison. If a_i-s and b_i-s are equal (or almost equal) for all i, inequality (15) matches Hoeffding-Azuma inequality up to $\ln(N+1)$ factor (which can also be halved by using a tighter bound in (13)). If a_i-s and b_i-s are not identical, inequality (15) can be potentially much worse, since a single large $(b_i - a_i)$ term will permanently increase $(b-a)$, but its relative contribution to (16) will decrease with the increase of N. However, when the empirical average is close to lower or upper limit of the domain interval the kl form of Lemma 7 in equation (14) is much tighter than the relaxed L_1 norm form in equation (15) (McAllester, 2003). Therefore, in situations, where the analysis can be carried out using the kl form of the bound, it might be preferable.

4 Proof of Theorem 2 (PAC-Bayesian Bound Based on Lemma 1)

Our proof uses the following lemma, which lays at the basis of PAC-Bayesian analysis from its inception and takes its roots back in information theory and statistical physics (Donsker and Varadhan, 1975, Dupuis and Ellis, 1997, Gray, 2011, Banerjee, 2006). The lemma allows to relate all posterior distributions ρ to a single prior distribution μ.

Lemma 9 *For any measurable function $\phi(h)$ on $\mathcal{H}$ and any distributions $\mu(h)$ and $\rho(h)$ on $\mathcal{H}$, we have*:

$$\mathbb{E}_{\rho(h)}[\phi(h)] \leq KL(\rho\|\mu) + \ln \mathbb{E}_{\mu(h)}[e^{\phi(h)}]. \tag{17}$$

Proof of Theorem 2: First, we show that $R(a) = \mathbb{E}_{\mathcal{T}_t}[\hat{R}_t(a)]$. Let $p(r|a)$ be the distribution of the reward for playing arm a and let R^a be a random variable distributed according to $p(r|a)$. Then for any t:

$$\begin{aligned} R(a) = \mathbb{E}_{p(r|a)}[R^a] &= \mathbb{E}_{p(r|a)}\left[\pi_t(a)\frac{1}{\pi_t(a)}R^a\right] = \mathbb{E}_{p(r|a)}\mathbb{E}_{\pi_t(a)}\left[\frac{1}{\pi_t(a)}I_t^a R^a\right] \\ &= \mathbb{E}_{p(r|a),\pi_t(a)}\left[\frac{1}{\pi_t(a)}I_t^a R_t\right] = \mathbb{E}_{p(r|a),\pi_t(a)}[R_t^a], \end{aligned} \tag{18}$$

where (18) holds since if $I_t^a = 1$, then R_t is distributed by $p(r|a)$, and otherwise R_t is irrelevant. Hence, we obtain that $\mathbb{E}_{\mathcal{T}_t}[\hat{R}_t(a)] = \mathbb{E}_{\mathcal{T}_t}[\frac{1}{t}\sum_{\tau=1}^{t} R_\tau^a] = R(a)$ for all a and t.

Note that $\hat{R}_t(a)$ is a sum of t random variables belonging to the $[0, \frac{1}{\pi_t^{lmin}}]$ interval. By scaling $R(a)$ and $\hat{R}_t(a)$ by a factor of π_t^{lmin} we scale the random variables to the $[0,1]$ interval, where Lemmas 1 and 6 can be applied.

We apply PAC-Bayesian analysis to the scaled version of $R(a)$ and $\hat{R}_t(a)$ for a fixed t:

$$\begin{aligned} t \cdot kl(\pi_t^{lmin}\hat{R}_t(\rho_t)\|\pi_t^{lmin}R(\rho_t)) &= t \cdot kl(\mathbb{E}_{\rho_t(a)}[\pi_t^{lmin}\hat{R}_t(a)]\|\mathbb{E}_{\rho_t(a)}[\pi_t^{lmin}R(\rho)]) \\ &\leq \mathbb{E}_{\rho_t(a)}[t \cdot kl(\pi_t^{lmin}\hat{R}_t(a)\|\pi_t^{lmin}R(a))] \quad (19) \\ &\leq KL(\rho_t\|\mu_t) + \ln \mathbb{E}_{\mu_t(a)}[e^{t \cdot kl(\pi_t^{lmin}\hat{R}_t(a)\|\pi_t^{lmin}R(a))}], \quad (20) \end{aligned}$$

where (19) is due to convexity of kl and (20) is by Lemma 9.

The second term in (20) can be bounded with high probability:

$$\begin{aligned} \mathbb{E}_{\mu_t(a)}[e^{t \cdot kl(\pi_t^{lmin}\hat{R}_t(a)\|\pi_t^{lmin}R(a))}] &\leq \frac{1}{\delta_t}\mathbb{E}_{\mathcal{T}_t}\mathbb{E}_{\mu_t(a)}[e^{t \cdot kl(\pi_t^{lmin}\hat{R}_t(a)\|\pi_t^{lmin}R(a))}] \quad (21) \\ &= \frac{1}{\delta_t}\mathbb{E}_{\mu_t(a)}\mathbb{E}_{\mathcal{T}_t}[e^{t \cdot kl(\pi_t^{lmin}\hat{R}_t(a)\|\pi_t^{lmin}R(a))}] \quad (22) \\ &\leq \frac{1}{\delta_t}(t+1), \quad (23) \end{aligned}$$

where (21) holds with probability greater than $1-\delta_t$ by Markov's inequality (Lemma 5), the interchange of expectations in (22) is possible since μ_t is independent of $\mathcal{T}_t$, and (23) is by Lemma 1 and Lemma 6. Substitution of (23) into (20) yields with probability greater than $1-\delta_t$:

$$kl(\pi_t^{lmin}\hat{R}_t(\rho_t)\|\pi_t^{lmin}R(\rho_t)) \leq \frac{KL(\rho_t\|\mu_t) + \ln\frac{t+1}{\delta_t}}{t}.$$

Finally, by setting $\delta_t = \frac{\delta}{t(t+1)} \geq \frac{\delta}{(t+1)^2}$ and applying union bound we obtain (1) for all t simultaneously (it is well-known that $\sum_{t=1}^{\infty}\frac{1}{t(t+1)} = \sum_{t=1}^{\infty}\left(\frac{1}{t} - \frac{1}{t+1}\right) = 1$). ■

The key ingredient that made the proof of Theorem 2 possible was Lemma 1, which enabled us to bound $\mathbb{E}_{\mathcal{T}_t}[e^{t \cdot kl(\pi_t^{lmin}\hat{R}_t(a)\|\pi_t^{lmin}R(a))}]$ even though the variables $\{R_1^a, .., R_t^a\}$ are dependent.

5 Proof of Theorem 3 (PAC-Bayesian Analysis Based on Hoeffding-Azuma Inequality)

In this section we provide an alternative PAC-Bayesian bound for $|\hat{R}_t^{w^t}(\rho_t) - R(\rho_t)|$ by using Hoeffding-Azuma inequality.

Proof of Theorem 3: Let

$$M_t^i(a) = \frac{1}{t}\sum_{\tau=1}^{i} w_\tau^t(R_\tau^a - R(a)).$$

Observe that $M_t^1(a), .., M_t^t(a)$ is a martingale [since $\mathbb{E}_{R_i^a}\left[M_t^i(a)\right] = M_t^{i-1}(a)$] and $M_t^t(a) = \left(\sum_{\tau=1}^t w_\tau^t\right)(\hat{R}_t^{w^t}(a) - R(a))$. Note that $(M_t^i - M_t^{i-1}) \in [-\frac{1}{\pi_i^{min}}, \frac{1}{\pi_i^{min}}]$ and $\mathbb{E}M_t^t = 0$. Hence, by Hoeffding-Azuma inequality (Lemma 8), for all a:

$$\mathbb{E}_{\mathcal{T}_t}\left[e^{\lambda_t\left(\sum_{\tau=1}^t w_\tau^t\right)(\hat{R}_t^{w^t}(a) - R(a))}\right] = \mathbb{E}_{\mathcal{T}_t}\left[e^{M_t^t(a)}\right] \leq e^{\frac{1}{2}\lambda_t^2\sum_{\tau=1}^t\left(\frac{w_\tau^t}{\pi_\tau^{min}}\right)^2}.$$

By going back to the proof of Theorem 2 and replacing $kl(\pi_t^{lmin}\hat{R}_t(a)\|\pi_t^{lmin}R(a))$ with $\hat{R}_t^{w^t}(a) - R(a)$ and substituting the bound on $\mathbb{E}_{\mathcal{T}_t}[e^{t \cdot kl(\pi_t^{lmin}\hat{R}_t(a)\|\pi_t^{lmin}R(a))}]$ with the bound on $\mathbb{E}_{\mathcal{T}_t}[e^{\lambda_t\left(\sum_{\tau=1}^t w_\tau^t\right)(\hat{R}_t^{w^t}(a) - R(a))}]$ we derived above we obtain that with probability greater than $1 - \frac{1}{2}\delta$ for all ρ_t

$$\hat{R}_t^{w^t}(\rho_t) - R(\rho_t) \leq \frac{KL(\rho_t\|\mu_t) + \frac{1}{2}\lambda_t^2\sum_{\tau=1}^t\left(\frac{w_\tau^t}{\pi_\tau^{min}}\right)^2 + 2\ln(t+1) + \ln\frac{2}{\delta}}{\lambda_t\sum_{\tau=1}^t w_\tau^t}$$

and, by a symmetric argument applied to $-M_t^1(a), .., -M_t^t(a)$,

$$R(\rho_t) - \hat{R}_t^{w^t}(\rho_t) \leq \frac{KL(\rho_t\|\mu_t) + \frac{1}{2}\lambda_t^2\sum_{\tau=1}^t\left(\frac{w_\tau^t}{\pi_\tau^{min}}\right)^2 + 2\ln(t+1) + \ln\frac{2}{\delta}}{\lambda_t\sum_{\tau=1}^t w_\tau^t}.$$

Hence, both hold simultaneously with probability greater than $1-\delta$ and yield (4). ■

6 Proof of Theorem 4 (The Regret Bound)

In this section we derive a regret bound based on Theorem 2. We then discuss some possible ways to tighten the regret bound.

The regret bound is derived for the special kind of posterior distribution $\tilde{\rho}_t^{exp}$ defined in (7) in Theorem 4, which is used as sampling distribution π_{t+1} for the next round of the game, as described in the theorem. Furthermore, we define a special kind of prior distribution μ_t^{exp} as:

$$\mu_t^{exp}(a) = \frac{1}{Z(\mu_t^{exp})} e^{\gamma_t R(a)}. \tag{24}$$

The prior μ_t^{exp} depends on the true expected rewards $R(a)$, but not on the sample and hence it is a legal prior.

Proof of Theorem 4: Let a^* be the action with the highest reward. The expected regret of the prediction strategy $\tilde{\rho}_t^{exp}$ at step $t+1$ can be written as follows:

$$R(a^*) - R(\tilde{\rho}_t^{exp}) = [R(a^*) - \hat{R}_t(a^*)] + [\hat{R}_t(a^*) - \hat{R}_t(\rho_t^{exp})] + [\hat{R}_t(\rho_t^{exp}) - R(\rho_t^{exp})] + [R(\rho_t^{exp}) - R(\tilde{\rho}_t^{exp})]. \tag{25}$$

We bound the terms in (25) one-by-one.

$R(a^*)$ and $\hat{R}_t(a^*)$ are the expected and the empirical rewards of a prediction strategy, which is a delta distribution on a^*. Hence, by Theorem 2:

$$\begin{aligned} R(a^*) - \hat{R}_t(a^*) &\leq \frac{1}{\varepsilon_t} \sqrt{\frac{-\ln \mu_t^{exp}(a^*) + 3\ln(t+1) - \ln\delta}{2t}} \\ &= \frac{1}{\varepsilon_t} \sqrt{\frac{\ln \frac{Z(\mu_t^{exp})}{e^{\gamma_t R(a^*)}} + 3\ln(t+1) - \ln\delta}{2t}} \\ &\leq \frac{1}{\varepsilon_t} \sqrt{\frac{\ln(K) + 3\ln(t+1) - \ln\delta}{2t}}, \end{aligned} \tag{26}$$

where in (26) we used the fact that $R(a^*) \geq R(a)$ for all a and hence $e^{\gamma_t R(a^*)} \geq \frac{1}{K}\sum_a e^{\gamma_t R(a)} = \frac{1}{K} Z(\mu_t^{exp})$.

For the second term in (25) we write:

$$\begin{aligned} \hat{R}_t(a^*) - \hat{R}_t(\rho_t^{exp}) &= \sum_a (\hat{R}_t(a^*) - \hat{R}_t(a)) \rho_t^{exp}(a) \\ &= \sum_a (\hat{R}_t(a^*) - \hat{R}_t(a)) \frac{e^{\gamma_t \hat{R}_t(a)}}{Z(\rho_t^{exp})} \\ &= \sum_a (\hat{R}_t(a^*) - \hat{R}_t(a)) \frac{e^{-\gamma_t(\hat{R}_t(a^*) - \hat{R}_t(a))}}{\sum_{a'} e^{-\gamma_t(\hat{R}_t(a^*) - \hat{R}_t(a'))}} \\ &\leq \frac{K}{\gamma_t}, \end{aligned} \tag{27}$$

where in (27) follows from the technical lemma below. The proof of the lemma is provided at the end of this section.

Lemma 10 *Let $x_1 = 0$ and $x_2, .., x_n$ be $n-1$ arbitrary numbers. For any $\alpha > 0$ and $n \geq 2$:*

$$\frac{\sum_{i=1}^n x_i e^{-\alpha x_i}}{\sum_{j=1}^n e^{-\alpha x_j}} \leq \frac{n}{\alpha}.$$

The third term in (25) is bounded by the following lemma adapted from Lever et al. (2010). The proof of this lemma is also provided at the end of this section.

Lemma 11 *For μ_t^{exp} and ρ_t^{exp} defined by* (24) *and* (8) *under the conditions of Theorem 2 the following holds simultaneously with the assertion of Theorem 2:*

$$\left| \hat{R}_t(\rho_t^{exp}) - R(\rho_t^{exp}) \right| \leq \frac{1}{\varepsilon_t \sqrt{2t}} \left(\frac{\gamma_t}{\varepsilon_t \sqrt{2t}} + \sqrt{3\ln(t+1) - \ln\delta} \right). \tag{28}$$

Finally, for the last term in (25):

$$\begin{aligned} R(\rho_t^{exp}) - R(\tilde{\rho}_t^{exp}) &= \sum_a (\rho_t^{exp}(a) - \tilde{\rho}_t^{exp}(a)) R(a) \\ &\leq \frac{1}{2} \|\rho_t^{exp} - \tilde{\rho}_t^{exp}\|_1 \qquad (29) \\ &= \frac{1}{2} \sum_a |\rho_t^{exp}(a) - (1 - K\varepsilon_{t+1})\rho_t^{exp}(a) - \varepsilon_{t+1}| \\ &= \frac{1}{2} \sum_a |K\varepsilon_{t+1}\rho_t^{exp}(a) - \varepsilon_{t+1}| \\ &\leq \frac{1}{2} K\varepsilon_{t+1} \sum_a \rho_t^{exp}(a) + \frac{1}{2} K\varepsilon_{t+1} \\ &= K\varepsilon_{t+1}. \end{aligned}$$

In (29) we used the fact that $R(a)$ is bounded by 1 and ρ_t^{exp} and $\tilde{\rho}_t^{exp}$ are probability distributions.

Gathering all the terms and substituting them back into (25) we obtain:

$$\begin{aligned} R(a^*) - R(\tilde{\rho}_t^{exp}) \leq \frac{1}{\varepsilon_t} \sqrt{\frac{\ln(K) + 3\ln(t+1) - \ln\delta}{2t}} + \frac{K}{\gamma_t} \\ + \frac{1}{\varepsilon_t \sqrt{2t}} \left(\frac{\gamma_t}{\varepsilon_t \sqrt{2t}} + \sqrt{3\ln(t+1) - \ln\delta} \right) + K\varepsilon_{t+1}. \end{aligned}$$

By choosing $\gamma_t = K^{1/4} t^{1/4}$ and $\varepsilon_t = K^{-1/4} t^{-1/4}$ we get:

$$R(a^*) - R(\tilde{\rho}_t^{exp}) \leq \frac{K^{3/4}}{(t+1)^{1/4}} \left(\sqrt{\frac{\ln(K) + 3\ln(t+1) - \ln\delta}{2K}} + 1 + \frac{1}{2} + \sqrt{\frac{3\ln(t+1) - \ln\delta}{2K}} + 1 \right).$$

By integration over t the total regret is bounded by $\tilde{O}(K^{3/4} t^{3/4})$, where $\tilde{O}$ hides logarithmic factors. ■

6.1 Proofs of Technical Lemmas for Section 6

We conclude this section with proofs of the two technical lemmas used in the proof of the regret bound.

Proof of Lemma 10: Since $x_1 = 0$ we have:

$$\begin{aligned} \frac{\sum_{i=1}^n x_i e^{-\alpha x_i}}{\sum_{j=1}^n e^{-\alpha x_j}} &= \frac{\sum_{i=1}^n x_i e^{-\alpha x_i}}{1 + \sum_{j=2}^n e^{-\alpha x_j}} \\ &\leq \sum_{i=1}^n x_i e^{-\alpha x_i} \\ &\leq \frac{n}{\alpha}, \end{aligned}$$

where the last inequality follows from the fact that $xe^{-\alpha x} \leq \frac{1}{\alpha}$. ■

We note that by numerical simulations it seems that a tighter bound $\frac{\sum_{i=1}^n x_i e^{-\alpha x_i}}{\sum_{j=1}^n e^{-\alpha x_j}} \leq \frac{\ln(K)}{\alpha}$ holds, but we were unable to prove it analytically.

The proof of Lemma 11 is adapted with minor modifications from Lever et al. (2010) and is based on the following two lemmas, which are also adapted from Lever et al. (2010) and are proved right after the proof of Lemma 11.

Lemma 12 *For μ_t^{exp} and ρ_t^{exp} defined by* (24) *and* (8):

$$KL(\rho_t^{exp} \| \mu_t^{exp}) \leq \gamma_t \left([\hat{R}_t(\rho_t^{exp}) - R(\rho_t^{exp})] + [R(\mu_t^{exp}) - \hat{R}_t(\mu_t^{exp})] \right). \qquad (30)$$

Lemma 13 *For μ_t^{exp} and ρ_t^{exp} defined by* (24) *and* (8) *under the conditions of Theorem 2 the following holds simultaneously with the assertion of Theorem 2*:

$$KL(\rho_t^{exp} \| \mu_t^{exp}) \leq \left(\frac{\gamma_t}{\varepsilon_t \sqrt{2t}} \right)^2 + 2 \left(\frac{\gamma_t}{\varepsilon_t \sqrt{2t}} \right) \sqrt{3\ln(t+1) - \ln\delta}. \qquad (31)$$

Proof of Lemma 11: Substitution of (31) into (3) yields (28). ■

Proof of Lemma 12:

$$\begin{aligned}
KL(\rho_t^{exp}\|\mu_t^{exp}) &= \sum_a \rho_t^{exp}(a)\ln\left(\frac{e^{\gamma_t\hat{R}_t(a)}Z(\mu_t^{exp})}{e^{\gamma_t R(a)}Z(\rho_t^{exp})}\right)\\
&= \sum_a \rho_t^{exp}(a)\gamma_t(\hat{R}_t(a)-R(a)) - \ln\left(\frac{\sum_a e^{\gamma_t\hat{R}_t(a)}}{Z(\mu_t^{exp})}\right)\\
&= \gamma_t[\hat{R}_t(\rho_t^{exp})-R(\rho_t^{exp})] - \ln\left(\sum_a \mu_t^{exp}(a)e^{\gamma_t(\hat{R}_t(a)-R(a))}\right) \qquad (32)\\
&\le \gamma_t\left([\hat{R}_t(\rho_t^{exp})-R(\rho_t^{exp})] + [R(\mu_t^{exp})-\hat{R}_t(\mu_t^{exp})]\right). \qquad (33)
\end{aligned}$$

In (32) we used the fact that $\frac{1}{Z(\mu_t^{exp})} = \mu_t^{exp}(a)e^{-\gamma_t R(a)}$ (for any a) and in (33) we used the concavity of ln. ■

Proof of Lemma 13: By Theorem 2 and simultaneously with it we have:

$$\begin{aligned}
\hat{R}_t(\rho_t^{exp}) - R(\rho_t^{exp}) &\le \frac{1}{\varepsilon_t}\sqrt{\frac{KL(\rho_t^{exp}\|\mu_t^{exp}) + 3\ln(t+1) - \ln\delta}{2t}}\\
R(\mu_t^{exp}) - \hat{R}_t(\mu_t^{exp}) &\le \frac{1}{\varepsilon_t}\sqrt{\frac{3\ln(t+1)-\ln\delta}{2t}}.
\end{aligned}$$

By substituting this into (30) we have:

$$KL(\rho_t^{exp}\|\mu_t^{exp}) \le \frac{\gamma_t}{\varepsilon_t\sqrt{2t}}\sqrt{KL(\rho_t^{exp}\|\mu_t^{exp}) + 3\ln(t+1) - \ln\delta} + \frac{\gamma_t}{\varepsilon_t\sqrt{2t}}\sqrt{3\ln(t+1)-\ln\delta}.$$

If $KL(\rho_t^{exp}\|\mu_t^{exp}) \le \frac{\gamma_t}{\varepsilon_t\sqrt{2t}}$ we are done. Otherwise, by rearranging the terms we obtain:

$$\begin{aligned}
(KL(\rho_t^{exp}\|\mu_t^{exp}))^2 - 2KL(\rho_t^{exp}\|\mu_t^{exp})\frac{\gamma_t}{\varepsilon_t\sqrt{2t}}\sqrt{3\ln(t+1)-\ln\delta} + \left(\frac{\gamma_t}{\varepsilon_t\sqrt{2t}}\right)^2(3\ln(t+1)-\ln\delta)\\
\le \left(\frac{\gamma_t}{\varepsilon_t\sqrt{2t}}\right)^2 KL(\rho_t^{exp}\|\mu_t^{exp}) + \left(\frac{\gamma_t}{\varepsilon_t\sqrt{2t}}\right)^2(3\ln(t+1)-\ln\delta),
\end{aligned}$$

which together with the fact that $KL(\rho_t^{exp}\|\mu_t^{exp}) \ge 0$ implies the result. ■

7 Discussion

We presented a lemma that allows to bound expectations of convex functions of certain sequentially dependent variables by expectations of the same functions of i.i.d. Bernoulli variables. We showed that this lemma can be used to derive an alternative to Hoeffding-Azuma inequality for convergence of martingale values.

We presented two different approaches to PAC-Bayesian analysis of martingale-type sequentially dependent random variables, which was an important challenge for PAC-Bayesian analysis for a long time. Our contribution opens the possibility to apply PAC-Bayesian analysis in multiple domains, where sequentially dependent variables are encountered. For example, Theorems 2 and 3 can be used to bound convergence of uncountable number of parallel martingale sequences, where simple union bound does not apply.

We answered positively an important open question whether PAC-Bayesian analysis can be applied under limited feedback and used to study the exploration-exploitation trade-off. Although our regret bound for the multiarmed bandit problem is far from state-of-the-art yet, we believe that this gap can be closed in future work.

Multiarmed bandits are just the first tier in a whole hierarchy of reinforcement learning problems with increasing structural complexity, including continuum-armed bandits, contextual bandits, and reinforcement learning in discrete and continuous spaces. In many of these domains Bayesian approaches and incorporation of prior knowledge have already proved beneficial in practice, but their rigorous analysis remains difficult to carry out. We believe that PAC-Bayesian approach will prove to be as useful for this purpose as it already proved itself in the domain of supervised learning.

Acknowledgements

We thank John Langford for helpful discussions at the early stages of this work. We are also grateful to anonymous reviewers for their insightful comments 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. This publication only reflects the authors' views.

References

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.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tôhoku Mathematical Journal*, 19(3), 1967.

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

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

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

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

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 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.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

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.

Andreas Maurer. A note on the PAC-Bayesian theorem. www.arxiv.org, 2004.

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.

Matthias Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalization Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, 2003.

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

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.

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.