PAC-Bayesian Inequalities for Martingales

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, Peter Auer

***Abstract*—We present a set of high-probability inequalities that control the concentration of weighted averages of multiple (possibly uncountably many) simultaneously evolving and interdependent martingales. Our results extend the PAC-Bayesian analysis in learning theory from the i.i.d. setting to martingales opening the way for its application to importance weighted sampling, reinforcement learning, and other interactive learning domains, as well as many other domains in probability theory and statistics, where martingales are encountered.**

We also present a comparison inequality that bounds the expectation of a convex function of a martingale difference sequence shifted to the $[0,1]$ interval by the expectation of the same function of independent Bernoulli random variables. This inequality is applied to derive a tighter analog of Hoeffding-Azuma's inequality.

***Index Terms*—Martingales, Hoeffding-Azuma's inequality, Bernstein's inequality, PAC-Bayesian bounds.**

I. Introduction

MARTINGALES are one of the fundamental tools in probability theory and statistics for modeling and studying sequences of random variables. Some of the most well-known and widely used concentration inequalities for individual martingales are Hoeffding-Azuma's and Bernstein's inequalities [1], [2], [3]. We present a comparison inequality that bounds the expectation of a convex function of a martingale difference sequence shifted to the $[0,1]$ interval by the expectation of the same function of independent Bernoulli random variables. We apply this inequality in order to derive a tighter analog of Hoeffding-Azuma's inequality for martingales.

More importantly, we present a set of inequalities that make it possible to control weighted averages of multiple simultaneously evolving and interdependent martingales (see Fig. 1 for an illustration). The inequalities are especially interesting when the number of martingales is uncountably infinite and the standard union bound over the individual martingales cannot be applied. The inequalities hold with high probability simultaneously for a large class of averaging laws ρ. In particular, ρ can depend on the sample.

One possible application of our inequalities is an analysis of importance-weighted sampling. Importance-weighted sampling is a general and widely used technique for estimating properties of a distribution by drawing samples from a different distribution. Via proper reweighting of the samples, the expectation of the desired statistics based on the reweighted samples from the controlled distribution can be made identical to the expectation of the same statistics based on unweighted samples from the desired distribution. Thus, the difference between the observed statistics and its expected value forms a martingale difference sequence. Our inequalities can be applied in order to control the deviation of the observed statistics from its expected value. Furthermore, since the averaging law ρ can depend on the sample, the controlled distribution can be adapted based on its outcomes from the preceding rounds, for example, for denser sampling in the data-dependent regions of interest. See [4] for an example of an application of this technique in reinforcement learning.

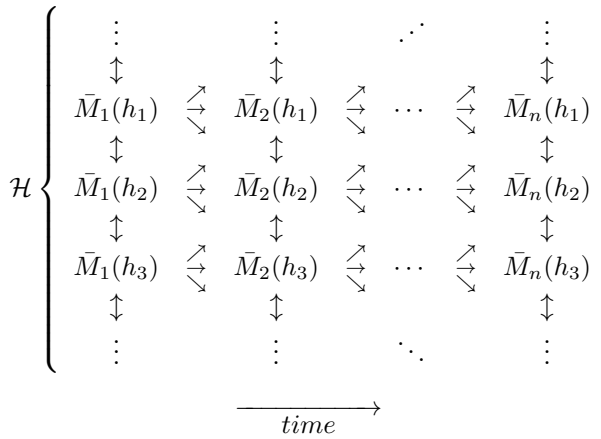

Fig. 1. Illustration of an infinite set of simultaneously evolving and interdependent martingales. $\mathcal{H}$ is a space that indexes the individual martingales. For a fixed point $h \in \mathcal{H}$, the sequence $\bar{M}_1(h), \bar{M}_2(h), \ldots, \bar{M}_n(h)$ is a single martingale. The arrows represent the dependencies between the values of the martingales: the value of a martingale h at time i, denoted by $\bar{M}_i(h)$, depends on $\bar{M}_j(h')$ for all $j \leq i$ and $h' \in \mathcal{H}$ (everything that is "before" and "concurrent" with $\bar{M}_i(h)$ in time; some of the arrows are omitted for clarity). A mean value of the martingales with respect to a probability distribution ρ over $\mathcal{H}$ is given by $\langle \bar{M}_n, \rho \rangle$. Our high-probability inequalities bound $|\langle \bar{M}_n, \rho \rangle|$ simultaneously for a large class of ρ.

Our concentration inequalities for weighted averages of martingales are based on a combination of Donsker-Varadhan's variational formula for relative entropy [5], [6], [7] with bounds on certain moment generating functions of martingales, including Hoeffding-Azuma's and Bernstein's inequalities, as well as the new inequality derived in this paper.

In a nutshell, the Donsker-Varadhan's variational formula implies that for a probability space $(\mathcal{H}, \mathcal{B})$, a bounded real-valued random variable Φ and any two probability distributions π and ρ over $\mathcal{H}$ (or, if $\mathcal{H}$ is uncountably infinite, two probability density functions), the expected value $\mathbb{E}_\rho[\Phi]$ is

Yevgeny Seldin is with Max Planck Institute for Intelligent Systems, Tübingen, Germany, and University College London, London, UK. E-mail: seldin@tuebingen.mpg.de

François Laviolette is with Université Laval, Québec, Canada. E-mail: francois.laviolette@ift.ulaval.ca

Nicolò Cesa-Bianchi is with Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy. E-mail: nicolo.cesa-bianchi@unimi.it

John Shawe-Taylor is with University College London, London, UK. E-mail: jst@cs.ucl.ac.uk

Peter Auer is with Chair for Information Technology, Montanuniversität Leoben, Leoben, Austria. E-mail: auer@unileoben.ac.at

bounded as:

$$\mathbb{E}_\rho[\Phi] \leq \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_\pi[e^\Phi], \tag{1}$$

where $\mathrm{KL}(\rho\|\pi)$ is the KL-divergence (relative entropy) between two distributions [8]. We can also think of Φ as $\Phi = \phi(h)$, where $\phi(h)$ is a measurable function $\phi : \mathcal{H} \to \mathbb{R}$. Inequality (1) can then be written using the dot-product notation

$$\langle \phi, \rho \rangle \leq KL(\rho\|\pi) + \ln\left(\langle e^\phi, \pi \rangle\right) \tag{2}$$

and $\mathbb{E}_\rho[\phi] = \langle \phi, \rho \rangle$ can be thought of as a weighted average of ϕ with respect to ρ (for countable $\mathcal{H}$ it is defined as $\langle \phi, \rho \rangle = \sum_{h\in\mathcal{H}} \phi(h)\rho(h)$ and for uncountable $\mathcal{H}$ it is defined as $\langle \phi, \rho \rangle = \int_{\mathcal{H}} \phi(h)\rho(h)dh$).[1]

The weighted averages $\langle \phi, \rho \rangle$ on the left hand side of (2) are the quantities of interest and the inequality allows us to relate all possible averaging laws ρ to a single "reference" distribution π. (Sometimes, π is also called a "prior" distribution, since it has to be selected before observing the sample.) We emphasize that inequality (2) is a deterministic relation. Thus, by a single application of Markov's inequality to $\langle e^\phi, \pi \rangle$ we obtain a statement that holds with high probability for all ρ simultaneously. The quantity $\ln\langle e^\phi, \pi \rangle$, known as the cumulant-generating function of ϕ, is closely related to the moment-generating function of ϕ. The bound on $\ln\langle e^\phi, \pi \rangle$, after some manipulations, is achieved via the bounds on moment-generating functions, which are identical to those used in the proofs of Hoeffding-Azuma's, Bernstein's, or our new inequality, depending on the choice of ϕ.

Donsker-Varadhan's variational formula for relative entropy laid the basis for PAC-Bayesian analysis in statistical learning theory [9], [10], [11], [12], where PAC is an abbreviation for the Probably Approximately Correct learning model introduced by Valiant [13]. PAC-Bayesian analysis provides high probability bounds on the deviation of weighted averages of empirical means of sets of independent random variables from their expectations. In the learning theory setting, the space $\mathcal{H}$ usually corresponds to a hypothesis space; the function $\phi(h)$ is related to the difference between the expected and empirical error of a hypothesis h; the distribution π is a prior distribution over the hypothesis space; and the distribution ρ defines a randomized classifier. The randomized classifier draws a hypothesis h from $\mathcal{H}$ according to ρ at each round of the game and applies it to make the prediction on the next sample. PAC-Bayesian analysis supplied generalization guarantees for many influential machine learning algorithms, including support vector machines [14], [15], linear classifiers [16], and clustering-based models [17], to name just a few of them.

We show that PAC-Bayesian analysis can be extended to martingales. A combination of PAC-Bayesian analysis with Hoeffding-Azuma's inequality was applied by Lever et. al [18] in the analysis of U-statistics. The results presented here are both tighter and more general, and make it possible to apply PAC-Bayesian analysis in new domains, such as, for example, reinforcement learning [4].

[1]The complete statement of Donsker-Varadhan's variational formula for relative entropy states that under appropriate conditions $\mathrm{KL}(\rho\|\pi) = \sup_\phi\left(\langle \phi, \rho \rangle - \ln\langle e^\phi, \pi \rangle\right)$, where the supremum is achieved by $\phi(h) = \ln\frac{\rho(h)}{\pi(h)}$. However, in our case the choice of ϕ is directly related to the values of the martingales of interest and the free parameters in the inequality are the choices of ρ and π. Therefore, we are looking at the inequality in the form of equation (1) and a more appropriate name for it is "change of measure inequality".

II. Main Results

We first present our new inequalities for individual martingales, and then present the inequalities for weighted averages of martingales. All the proofs are provided in the appendix.

A. Inequalities for Individual Martingales

Our first lemma is a comparison inequality that bounds expectations of convex functions of martingale difference sequences shifted to the $[0, 1]$ interval by expectations of the same functions of independent Bernoulli random variables. The lemma generalizes a previous result by Maurer for independent random variables [19]. The lemma uses the following notation: for a sequence of random variables $X_1, \ldots, X_n$ we use $X_1^i := X_1, \ldots, X_i$ to denote the first i elements of the sequence.

Lemma 1: Let $X_1, \ldots, X_n$ be a sequence of random variables, such that $X_i \in [0, 1]$ with probability 1 and $\mathbb{E}[X_i|X_1^{i-1}] = b_i$ for $i = 1, \ldots, n$. Let $Y_1, \ldots, Y_n$ be independent Bernoulli random variables, such that $\mathbb{E}[Y_i] = b_i$. Then for any convex function $f : [0, 1]^n \to \mathbb{R}$:

$$\mathbb{E}\left[f(X_1, \ldots, X_n)\right] \leq \mathbb{E}\left[f(Y_1, \ldots, Y_n)\right].$$

Let $\mathrm{kl}(p\|q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$ be an abbreviation for $\mathrm{KL}\left([p, 1-p]\|[q, 1-q]\right)$, where $[p, 1-p]$ and $[q, 1-q]$ are Bernoulli distributions with biases p and q, respectively. By Pinsker's inequality [8],

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

which means that a bound on $\mathrm{kl}(p\|q)$ implies a bound on the absolute difference between the biases of the Bernoulli distributions.

We apply Lemma 1 in order to derive the following inequality, which is an interesting generalization of an analogous result for i.i.d. variables. The result is based on the method of types in information theory [8].

Lemma 2: Let $X_1, \ldots, X_n$ be a sequence of random variables, such that $X_i \in [0, 1]$ with probability 1 and $\mathbb{E}[X_i|X_1^{i-1}] = b$. Let $S_n := \sum_{i=1}^n X_i$. Then:

$$\mathbb{E}\left[e^{n\,\mathrm{kl}\left(\frac{1}{n}S_n\|b\right)}\right] \leq n + 1. \tag{3}$$

Note that in Lemma 2 the conditional expectation $\mathbb{E}[X_i|X_1^{i-1}]$ is identical for all i, whereas in Lemma 1 there is no such restriction. Combination of Lemma 2 with Markov's inequality leads to the following analog of Hoeffding-Azuma inequality.

Corollary 3: Let $X_1, \ldots, X_n$ be as in Lemma 2. Then, for any $\delta \in (0, 1)$, with probability greater than $1 - \delta$:

$$\mathrm{kl}\left(\frac{1}{n}S_n\middle\|b\right) \leq \frac{1}{n}\ln\frac{n+1}{\delta}. \tag{4}$$

S_n is a terminal point of a random walk with bias b after n steps. By combining Corollary 3 with Pinsker's inequality we

can obtain a more explicit bound on the deviation of the terminal point from its expected value, $|S_n - bn| \leq \sqrt{\frac{n}{2}\ln\frac{n+1}{\delta}}$, which is similar to the result we can obtain by applying Hoeffding-Azuma's inequality. However, in certain situations the less explicit bound in the form of kl is significantly tighter than Hoeffding-Azuma's inequality and it can also be tighter than Bernstein's inequality. A detailed comparison is provided in Section III.

B. PAC-Bayesian Inequalities for Weighted Averages of Martingales

Next, we present several inequalities that control the concentration of weighted averages of multiple simultaneously evolving and interdependent martingales. The first result shows that the classical PAC-Bayesian theorem for independent random variables [12] holds in the same form for martingales. The result is based on combination of Donsker-Varadhan's variational formula for relative entropy with Lemma 2. In order to state the theorem we need a few definitions.

Let $(\mathcal{H}, \mathcal{B})$ be a probability space. Let $\bar{X}_1, \ldots, \bar{X}_n$ be a sequence of random functions, such that $\bar{X}_i : \mathcal{H} \to [0,1]$. Assume that $\mathbb{E}[\bar{X}_i | \bar{X}_1, \ldots, \bar{X}_{i-1}] = \bar{b}$, where $\bar{b} : \mathcal{H} \to [0,1]$ is a deterministic function (possibly unknown). This means that $\mathbb{E}[\bar{X}_i(h) | \bar{X}_1, \ldots, \bar{X}_{i-1}] = \bar{b}(h)$ for each i and h. Note that for each $h \in \mathcal{H}$ the sequence $\bar{X}_1(h), \ldots, \bar{X}_n(h)$ satisfies the condition of Lemma 2.

Let $\bar{S}_n := \sum_{i=1}^n \bar{X}_i$. In the following theorem we are bounding the mean of $\bar{S}_n$ with respect to any probability measure ρ over $\mathcal{H}$.

Theorem 4 (PAC-Bayes-kl Inequality): Fix a reference distribution π over $\mathcal{H}$. Then, for any $\delta \in (0,1)$, with probability greater than $1-\delta$ over $\bar{X}_1, \ldots, \bar{X}_n$, for all distributions ρ over $\mathcal{H}$ simultaneously:

$$\mathrm{kl}\left(\left\langle \frac{1}{n}\bar{S}_n, \rho \right\rangle \middle\| \langle \bar{b}, \rho \rangle\right) \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n}. \tag{5}$$

By Pinsker's inequality, Theorem 4 implies that

$$\left|\left\langle \frac{1}{n}\bar{S}_n, \rho \right\rangle - \langle \bar{b}, \rho \rangle\right| = \left|\left\langle \left(\frac{1}{n}\bar{S}_n - \bar{b}\right), \rho \right\rangle\right| \leq \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{2n}}, \tag{6}$$

however, if $\left\langle \frac{1}{n}\bar{S}_n, \rho \right\rangle$ is close to zero or one, inequality (5) is significantly tighter than (6).

The next result is based on combination of Donsker-Varadhan's variational formula for relative entropy with Hoeffding-Azuma's inequality. This time let $\bar{Z}_1, \ldots, \bar{Z}_n$ be a sequence of random functions, such that $\bar{Z}_i : \mathcal{H} \to \mathbb{R}$. Let $\bar{Z}_1^i$ be an abbreviation for a subsequence of the first i random functions in the sequence. We assume that $\mathbb{E}[\bar{Z}_i | \bar{Z}_1^i] = \bar{0}$. In other words, for each $h \in \mathcal{H}$ the sequence $Z_1(h), \ldots, Z_n(h)$ is a martingale difference sequence.

Let $\bar{M}_i := \sum_{j=1}^i \bar{Z}_j$. Then, for each $h \in \mathcal{H}$ the sequence $\bar{M}_1(h), \ldots, \bar{M}_n(h)$ is a martingale. In the following theorems we bound the mean of $\bar{M}_n$ with respect to any probability measure ρ on $\mathcal{H}$.

Theorem 5: Assume that $\bar{Z}_i : \mathcal{H} \to [\alpha_i, \beta_i]$. Fix a reference distribution π over $\mathcal{H}$ and $\lambda > 0$. Then, for any $\delta \in (0,1)$, with probability greater than $1-\delta$ over $\bar{Z}_1^n$, for all distributions ρ over $\mathcal{H}$ simultaneously:

$$|\langle \bar{M}_n, \rho \rangle| \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2}{\delta}}{\lambda} + \frac{\lambda}{8}\sum_{i=1}^n (\beta_i - \alpha_i)^2. \tag{7}$$

We note that we cannot minimize inequality (7) simultaneously for all ρ by a single value of λ. In the following theorem we take a grid of λ-s in a form of a geometric sequence and for each value of $\mathrm{KL}(\rho\|\pi)$ we pick a value of λ from the grid, which is the closest to the one that minimizes (7). The result is almost as good as what we could achieve if we would minimize the bound just for a single value of ρ.

Theorem 6 (PAC-Bayes-Hoeffding-Azuma Inequality): Assume that $\bar{Z}_1^n$ is as in Theorem 5. Fix a reference distribution π over $\mathcal{H}$. Take an arbitrary number $c > 1$. Then, for any $\delta \in (0,1)$, with probability greater than $1-\delta$ over $\bar{Z}_1^n$, for all distributions ρ over $\mathcal{H}$ simultaneously:

$$|\langle \bar{M}_n, \rho \rangle| \leq \frac{1+c}{2\sqrt{2}}\sqrt{\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{2}{\delta} + \epsilon(\rho)\right)\sum_{i=1}^n (\beta_i - \alpha_i)^2}, \tag{8}$$

where

$$\epsilon(\rho) = \frac{\ln 2}{2\ln c}\left(1 + \ln\left(\frac{\mathrm{KL}(\rho\|\pi)}{\ln\frac{2}{\delta}}\right)\right).$$

Our last result is based on a combination of Donsker-Varadhan's variational formula with a Bernstein-type inequality for martingales. Let $\bar{V}_i : \mathcal{H} \to \mathbb{R}$ be such that $\bar{V}_i(h) := \sum_{j=1}^i \mathbb{E}\left[\bar{Z}_j(h)^2 \middle| \bar{Z}_1^{j-1}\right]$. In other words, $\bar{V}_i(h)$ is the variance of the martingale $\bar{M}_i(h)$ defined earlier. Let $\|\bar{Z}_i\|_\infty = \sup_{h\in\mathcal{H}} \bar{Z}_i(h)$ be the L_∞ norm of $\bar{Z}_i$.

Theorem 7: Assume that $\|\bar{Z}_i\|_\infty \leq K$ for all i with probability 1 and pick λ, such that $\lambda \leq 1/K$. Fix a reference distribution π over $\mathcal{H}$. Then, for any $\delta \in (0,1)$, with probability greater than $1-\delta$ over $\bar{Z}_1^n$, for all distributions ρ over $\mathcal{H}$ simultaneously:

$$|\langle \bar{M}_n, \rho \rangle| \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2}{\delta}}{\lambda} + (e-2)\lambda\langle \bar{V}_n, \rho \rangle. \tag{9}$$

As in the previous case, the right hand side of (9) cannot be minimized for all ρ simultaneously by a single value of λ. Furthermore, $\bar{V}_n$ is a random function. In the following theorem we take a similar grid of λ-s, as we did in Theorem 6, and a union bound over the grid. Picking a value of λ from the grid closest to the value of λ that minimizes the right hand side of (9) yields almost as good result as we would get if we would minimize (9) for a single choice of ρ. In this approach the variance $\bar{V}_n$ can be replaced by a sample-dependent upper bound. For example, in importance-weighted sampling such an upper bound is derived from the reciprocal of the sampling distribution at each round [4].

Theorem 8 (PAC-Bayes-Bernstein Inequality): Assume that $\|\bar{Z}_i\|_\infty \leq K$ for all i with probability 1. Fix a reference distribution π over $\mathcal{H}$. Pick an arbitrary number $c > 1$. Then,

for any $\delta \in (0,1)$, with probability greater than $1-\delta$ over $\bar{Z}_1^n$, simultaneously for all distributions ρ over $\mathcal{H}$ that satisfy

$$\sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2\nu}{\delta}}{(e-2)\langle \bar{V}_n, \rho\rangle}} \le \frac{1}{K} \tag{10}$$

we have

$$|\langle \bar{M}_n, \rho\rangle| \le (1+c)\sqrt{(e-2)\langle \bar{V}_n, \rho\rangle \left(\mathrm{KL}(\rho\|\pi) + \ln\frac{2\nu}{\delta}\right)}, \tag{11}$$

where

$$\nu = \left\lceil \frac{\ln\left(\sqrt{\frac{(e-2)n}{\ln\frac{2}{\delta}}}\right)}{\ln c} \right\rceil + 1, \tag{12}$$

and for all other ρ

$$|\langle \bar{M}_n, \rho\rangle| \le 2K\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{2\nu}{\delta}\right). \tag{13}$$

($\lceil x \rceil$ is the smallest integer value that is larger than x.)

III. Comparison of the Inequalities

In this section we remind the reader of Hoeffding-Azuma's and Bernstein's inequalities for individual martingales and compare them with our new kl-form inequality. Then, we compare inequalities for weighted averages of martingales with inequalities for individual martingales.

A. Background

We first recall Hoeffding-Azuma's inequality [1], [2]. For a sequence of random variables $Z_1, \ldots, Z_n$ we use $Z_1^i := Z_1, \ldots, Z_i$ to denote the first i elements of the sequence.

Lemma 9 (Hoeffding-Azuma's Inequality): Let $Z_1, \ldots, Z_n$ be a martingale difference sequence, such that $Z_i \in [\alpha_i, \beta_i]$ with probability 1 and $\mathbb{E}[Z_i|Z_1^{i-1}] = 0$. Let $M_i = \sum_{j=1}^i Z_j$ be the corresponding martingale. Then for any $\lambda \in \mathbb{R}$:

$$\mathbb{E}[e^{\lambda M_n}] \le e^{(\lambda^2/8)\sum_{i=1}^n (\beta_i - \alpha_i)^2}.$$

By combining Hoeffding-Azuma's inequality with Markov's inequality and taking $\lambda = \sqrt{\frac{8\ln\frac{2}{\delta}}{\sum_{i=1}^n(\beta_i-\alpha_i)^2}}$ it is easy to obtain the following corollary.

Corollary 10: For M_n defined in Lemma 9 and $\delta \in (0,1)$, with probability greater than $1-\delta$:

$$|M_n| \le \sqrt{\frac{1}{2}\ln\left(\frac{2}{\delta}\right)\sum_{i=1}^n (\beta_i - \alpha_i)^2}.$$

The next lemma is a Bernstein-type inequality [3], [20]. We provide the proof of this inequality in Appendix C, the proof is a part of the proof of [21, Theorem 1].

Lemma 11 (Bernstein's Inequality): Let $Z_1, \ldots, Z_n$ be a martingale difference sequence, such that $|Z_i| \le K$ with probability 1 and $\mathbb{E}[Z_i|Z_1^{i-1}] = 0$. Let $M_i := \sum_{j=1}^i Z_j$ and let $V_i := \sum_{j=1}^i \mathbb{E}[(Z_j)^2|Z_1^{j-1}]$. Then for any $\lambda \in [0, \frac{1}{K}]$:

$$\mathbb{E}\left[e^{\lambda M_n - (e-2)\lambda^2 V_n}\right] \le 1.$$

By combining Lemma 11 with Markov's inequality we obtain that for any $\lambda \in [0, \frac{1}{K}]$ and $\delta \in (0,1)$, with probability greater than $1-\delta$:

$$|M_n| \le \frac{1}{\lambda}\ln\frac{2}{\delta} + \lambda(e-2)V_n. \tag{14}$$

V_n is a random variable and can be replaced by an upper bound. Inequality (14) is minimized by $\lambda^* = \sqrt{\frac{\ln\frac{2}{\delta}}{(e-2)V_n}}$. Note that λ^* depends on V_n and is not accessible until we observe the entire sample. We can bypass this problem by constructing the same grid of λ-s, as the one used in the proof of Theorem 8, and taking a union bound over it. Picking a value of λ closest to λ^* from the grid leads to the following corollary. In this bounding technique the upper bound on V_n can be sample-dependent, since the bound holds simultaneously for all λ-s in the grid. Despite being a relatively simple consequence of Lemma 11, we have not seen this result in the literature. The corollary is tighter than an analogous result by Beygelzimer et. al. [21, Theorem 1].

Corollary 12: For M_n and V_n as defined in Lemma 11, $c > 1$ and $\delta \in (0,1)$, with probability greater than $1-\delta$, if

$$\sqrt{\frac{\ln\frac{2\nu}{\delta}}{(e-2)V_n}} \le \frac{1}{K} \tag{15}$$

then

$$|M_n| \le (1+c)\sqrt{(e-2)V_n \ln\frac{2\nu}{\delta}},$$

where ν is defined in (12), and otherwise

$$|M_n| \le 2K\ln\frac{2\nu}{\delta}.$$

The technical condition (15) follows from the requirement of Lemma 11 that $\lambda \in [0, \frac{1}{K}]$.

B. Comparison

We first compare inequalities for individual martingales in Corollaries 3, 10, and 12.

Comparison of Inequalities for Individual Martingales: The comparison between Corollaries 10 and 12 is relatively straightforward. We note that the assumption $\mathbb{E}[Z_i|Z_1^{i-1}] = 0$ implies that $\alpha_i \le 0$ and that $V_n \le \sum_{i=1}^n \max\{\alpha_i^2, \beta_i^2\} \le \sum_{i=1}^n (\beta_i - \alpha_i)^2$. Hence, Corollary 12 (derived from Bernstein's inequality) matches Corollary 10 (derived from Hoeffding-Azuma's inequality) up to minor constants and logarithmic factors in the general case, and can be much tighter when the variance is small.

The comparison with the kl inequality in Corollary 3 is a bit more involved. As we mentioned after Corollary 3, its combination with Pinsker's inequality implies that $|S_n - bn| \le \sqrt{\frac{n}{2}\ln\frac{n+1}{\delta}}$, where $S_n - bn$ is a martingale corresponding to the martingale difference sequence $Z_i = X_i - b$. Thus, Corollary 3 is at least as tight as Hoeffding-Azuma's inequality in Corollary 10, up to a factor of $\sqrt{\ln\frac{n+1}{2}}$. This is also true if $X_i \in [\alpha_i, \beta_i]$ (rather than $[0,1]$), as long as we can simultaneously project all X_i-s to the $[0,1]$ interval without losing too much.

Tighter upper bounds on the kl divergence show that in certain situations Corollary 3 is actually much tighter than Hoeffding-Azuma's inequality. One possible application of Corollary 3 is estimation of the value of the drift b of a random walk from empirical observation S_n. If S_n is close to zero, it is possible to use a tighter bound on kl, which states that for $p > q$ we have $p \leq q + \sqrt{2q\,\mathrm{kl}(q||p)} + 2\mathrm{kl}(q||p)$ [15]. From this inequality, we obtain that with probability greater than $1-\delta$:

$$b \leq \frac{1}{n} S_n + \sqrt{\frac{\frac{2}{n} S_n \ln \frac{n+1}{\delta}}{n}} + \frac{2 \ln \frac{n+1}{\delta}}{n}.$$

The above inequality is tighter than Hoeffding-Azuma inequality whenever $\frac{1}{n} S_n < 1/8$. Since kl is convex in each of its parameters, it is actually easy to invert it numerically, and thus avoid the need to resort to approximations in practice. In a similar manner, tighter bounds can be obtained when S_n is close to n.

The comparison of kl inequality in Corollary 3 with Bernstein's inequality in Corollary 12 is not as equivocal as the comparison with Hoeffding-Azuma's inequality. If there is a bound on V_n that is significantly tighter than n, Bernstein's inequality can be significantly tighter than the kl inequality, but otherwise it can also be the opposite case. In the example of estimating a drift of a random walk without prior knowledge on its variance, if the empirical drift is close to zero or to n the kl inequality is tighter. In this case the kl inequality is comparable with empirical Bernstein's bounds [22], [23], [24].

Comparison of Inequalities for Individual Martingales with PAC-Bayesian Inequalities for Weighted Averages of Martingales: The "price" that is paid for considering weighted averages of multiple martingales is the KL-divergence $\mathrm{KL}(\rho\|\pi)$ between the desired mixture weights ρ and the reference mixture weights π. (In the case of PAC-Bayes-Hoeffding-Azuma inequality, Theorem 6, there is also an additional minor term originating from the union bound over the grid of λ-s.) Note that for $\rho = \pi$ the KL term vanishes.

IV. Discussion

We presented a comparison inequality that bounds expectation of a convex function of martingale difference type variables by expectation of the same function of independent Bernoulli random variables. This inequality enables to reduce a problem of studying continuous dependent random variables on a bounded interval to a much simpler problem of studying independent Bernoulli random variables.

As an example of an application of the inequality we derived an analog of Hoeffding-Azuma's inequality for martingales. Our result is always comparable to Hoeffding-Azuma's inequality up to a logarithmic factor and in cases, where the empirical drift of a corresponding random walk is close to the region boundaries it is tighter than Hoeffding-Azuma's inequality by an order of magnitude. It can also be tighter than Bernstein's inequality for martingales, unless there is a tight bound on the martingale variance.

Finally, but most importantly, we presented a set of inequalities on concentration of weighted averages of multiple simultaneously evolving and interdependent martingales. These inequalities are especially useful for controlling uncountably many martingales, where standard union bounds cannot be applied. Martingales are one of the most basic and important tools for studying time-evolving processes and we believe that our results will be useful for multiple domains. One such application in analysis of importance weighted sampling in reinforcement learning was already presented in [4].

Appendix A
Proofs of the Results for Individual Martingales

Proof of Lemma 1: The proof follows the lines of the proof of Maurer [19, Lemma 3]. Any point $\bar{x} = (x_1, \ldots, x_n) \in [0,1]^n$ can be written as a convex combination of the extreme points $\bar{\eta} = (\eta_1, \ldots, \eta_n) \in \{0,1\}^n$ in the following way:

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

Convexity of f therefore implies

$$f(\bar{x}) \leq \sum_{\bar{\eta} \in \{0,1\}^n} \left(\prod_{i=1}^{n} [(1-x_i)(1-\eta_i) + x_i \eta_i] \right) f(\bar{\eta}) \quad (16)$$

with equality if $\bar{x} \in \{0,1\}^n$. Let $X_1^i := X_1, \ldots, X_i$ be the first i elements of the sequence $X_1, \ldots, X_n$. Let $W_i(\eta_i) = (1-X_i)(1-\eta_i) + X_i \eta_i$ and let $w_i(\eta_i) = (1-b_i)(1-\eta_i) + b_i \eta_i$. Note that by the assumption of the lemma:

$$\begin{aligned} \mathbb{E}[W_i(\eta_i) | X_1^{i-1}] &= \mathbb{E}[(1-X_i)(1-\eta_i) + X_i \eta_i | X_1^{i-1}] \\ &= (1-b_i)(1-\eta_i) + b_i \eta_i = w_i(\eta_i). \end{aligned}$$

By taking expectation of both sides of (16) we obtain:

$$\begin{aligned} &\mathbb{E}_{X_1^n}[f(X_1^n)] \leq \mathbb{E}_{X_1^n} \left[\sum_{\bar{\eta} \in \{0,1\}^n} \left(\prod_{i=1}^{n} W_i(\eta_i) \right) f(\bar{\eta}) \right] \\ &= \sum_{\bar{\eta} \in \{0,1\}^n} \mathbb{E}_{X_1^n} \left[\prod_{i=1}^{n} W_i(\eta_i) \right] f(\bar{\eta}) \\ &= \sum_{\bar{\eta} \in \{0,1\}^n} \mathbb{E}_{X_1^{n-1}} \left[\mathbb{E}_{X_n} \left[\prod_{i=1}^{n} W_i(\eta_i) \middle| X_1^{n-1} \right] \right] f(\bar{\eta}) \\ &= \sum_{\bar{\eta} \in \{0,1\}^n} \mathbb{E}_{X_1^{n-1}} \left[\prod_{i=1}^{n-1} W_i(\eta_i) \mathbb{E}_{X_n} \left[W_n(\eta_n) | X_1^{n-1} \right] \right] f(\bar{\eta}) \\ &= \sum_{\bar{\eta} \in \{0,1\}^n} \mathbb{E}_{X_1^{n-1}} \left[\prod_{i=1}^{n-1} W_i(\eta_i) \right] w_n(\eta_n) f(\bar{\eta}) \\ &= \ldots \qquad (17) \\ &= \sum_{\bar{\eta} \in \{0,1\}^n} \left(\prod_{i=1}^{n} w_i(\eta_i) \right) f(\bar{\eta}) \\ &= \sum_{\bar{\eta} \in \{0,1\}^n} \left(\prod_{i=1}^{n} [(1-b_i)(1-\eta_i) + b_i \eta_i] \right) f(\bar{\eta}) \\ &= \mathbb{E}_{Y_1^n}[f(Y_1^n)]. \end{aligned}$$

In (17) we apply induction in order to replace X_i by b_i, one-by-one from the last to the first, same way we did it for X_n. ∎

Lemma 2 follows from the following concentration result for independent Bernoulli random variables that is based on the method of types in information theory [8]. Its proof can be found in [25], [17].

Lemma 13: Let $Y_1, \ldots, Y_n$ be i.i.d. Bernoulli random variables, such that $\mathbb{E}[Y_i] = b$. Then:

$$\mathbb{E}\left[e^{n\,\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n Y_i \middle\| b\right)}\right] \le n+1. \tag{18}$$

For $n \ge 8$ it is possible to prove even stronger result $\sqrt{n} \le \mathbb{E}[e^{n\,\mathrm{kl}(\frac{1}{n}\sum_{i=1}^n Y_i\|b)}] \le 2\sqrt{n}$ using Stirling's approximation of the factorial [19]. For the sake of simplicity we restrict ourselves to the slightly weaker bound (18), although all results that are based on Lemma 2 can be slightly improved by using the tighter bound.

Proof of Lemma 2: Since KL-divergence is a convex function [8] and the exponent function is convex and non-decreasing, $e^{n\,\mathrm{kl}(p\|q)}$ is also a convex function. Therefore, Lemma 2 follows from Lemma 13 by Lemma 1. ∎

Corollary 3 follows from Lemma 2 by Markov's inequality.

Lemma 14 (Markov's inequality): For $\delta \in (0,1)$ and a random variable $X \ge 0$, with probability greater than $1-\delta$:

$$X \le \frac{1}{\delta}\mathbb{E}[X]. \tag{19}$$

Proof of Corollary 3: By Markov's inequality and Lemma 2, with probability greater than $1-\delta$:

$$e^{n\,\mathrm{kl}\left(\frac{1}{n}S_n\|b\right)} \le \frac{1}{\delta}\mathbb{E}\left[e^{n\,\mathrm{kl}\left(\frac{1}{n}S_n\|b\right)}\right] \le \frac{n+1}{\delta}.$$

Taking logarithm of both sides of the inequality and normalizing by n completes the proof. ∎

Appendix B
Proofs of PAC-Bayesian Theorems for Martingales

In this appendix we provide the proofs of Theorems 4, 7, and 8. The proof of Theorem 5 is very similar to the proof of Theorem 7 and, therefore, omitted. The proof of Theorem 6 is very similar to the proof of Theorem 8, so we only provide the way of how to choose the grid of λ-s in this theorem.

The proofs of all PAC-Bayesian theorems are based on the following lemma, which is obtained by changing sides in Donsker-Varadhan's variational definition of relative entropy. The lemma takes roots back in information theory and statistical physics [5], [6], [7]. The lemma provides a deterministic relation between averages of ϕ with respect to all possible distributions ρ and the cumulant generating function $\ln\langle e^\phi, \pi\rangle$ with respect to a single reference distribution π. A single application of Markov's inequality combined with the bounds on moment generating functions in Lemmas 2, 9, and 11 is then used in order to bound the last term in (20) in the proofs of Theorems 4, 5, and 7, respectively.

Lemma 15 (Change of Measure Inequality): For any probability space $(\mathcal{H}, \mathcal{B})$, a measurable function $\phi : \mathcal{H} \to \mathbb{R}$, and any distributions π and ρ over $\mathcal{H}$, we have:

$$\langle \phi, \rho\rangle \le \mathrm{KL}(\rho\|\pi) + \ln\langle e^\phi, \pi\rangle. \tag{20}$$

Since the KL-divergence is infinite when the support of ρ exceeds the support of π, inequality (20) is interesting when $\pi \gg \rho$. For a similar reason, it is interesting only when $\langle e^\phi, \pi\rangle$ is finite. We note that the inequality is tight in the same sense as Jensen's inequality is tight: for $\phi(h) = \ln\frac{\rho(h)}{\pi(h)}$ it becomes an equality.

Proof of Theorem 4: Take $\phi(h) := n\,\mathrm{kl}\left(\frac{1}{n}\bar{S}_n(h)\middle\|\bar{b}(h)\right)$. More compactly, denote $\phi = \mathrm{kl}\left(\frac{1}{n}\bar{S}_n\middle\|\bar{b}\right) : \mathcal{H} \to \mathbb{R}$. Then with probability greater than $1-\delta$ for all ρ:

$$\begin{aligned}
&n\,\mathrm{kl}\left(\left\langle\frac{1}{n}\bar{S}_n, \rho\right\rangle \middle\| \langle\bar{b}, \rho\rangle\right)\\
&\quad\le n\left\langle \mathrm{kl}\left(\frac{1}{n}\bar{S}_n \middle\| \bar{b}\right), \rho\right\rangle &(21)\\
&\quad\le \mathrm{KL}(\rho\|\pi) + \ln\left\langle e^{n\,\mathrm{kl}(\frac{1}{n}\bar{S}_n\|\bar{b})}, \pi\right\rangle &(22)\\
&\quad\le \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathbb{E}_{\bar{X}_1^n}\left[\left\langle e^{n\,\mathrm{kl}(\frac{1}{n}\bar{S}_n\|\bar{b})}, \pi\right\rangle\right]\right) &(23)\\
&\quad= \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\left\langle \mathbb{E}_{\bar{X}_1^n}\left[e^{n\,\mathrm{kl}(\frac{1}{n}\bar{S}_n\|\bar{b})}\right], \pi\right\rangle\right) &(24)\\
&\quad\le \mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}, &(25)
\end{aligned}$$

where (21) is by convexity of the kl divergence [8]; (22) is by change of measure inequality (Lemma 15); (23) holds with probability greater than $1-\delta$ by Markov's inequality; in (24) we can take the expectation inside the dot product due to linearity of both operations and since π is deterministic; and (25) is by Lemma 2.[2] Normalization by n completes the proof of the theorem. ∎

Proof of Theorem 7: For the proof of Theorem 7 we take $\phi(h) := \lambda\bar{M}_n(h) - (e-2)\lambda^2\bar{V}_n(h)$. Or, more compactly, $\phi = \lambda\bar{M}_n - (e-2)\lambda^2\bar{V}_n$. Then with probability greater than $1-\frac{\delta}{2}$ for all ρ:

$$\begin{aligned}
\lambda\langle\bar{M}_n,\rho\rangle - (e-2)\lambda^2\langle\bar{V}_n,\rho\rangle &= \langle\lambda\bar{M}_n - (e-2)\lambda^2\bar{V}_n, \rho\rangle\\
&\le \mathrm{KL}(\rho\|\pi) + \ln\left\langle e^{\lambda\bar{M}_n-(e-2)\lambda^2\bar{V}_n}, \pi\right\rangle\\
&\le \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{2}{\delta}\mathbb{E}_{\bar{Z}_1^n}\left[\left\langle e^{\lambda\bar{M}_n-(e-2)\lambda^2\bar{V}_n}, \pi\right\rangle\right]\right) &(26)\\
&= \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{2}{\delta}\left\langle \mathbb{E}_{\bar{Z}_1^n}\left[e^{\lambda\bar{M}_n-(e-2)\lambda^2\bar{V}_n}\right], \pi\right\rangle\right)\\
&\le \mathrm{KL}(\rho\|\pi) + \ln\frac{2}{\delta}, &(27)
\end{aligned}$$

where (27) is by Lemma 11 and other steps are justified in the same way as in the previous proof.

By applying the same argument to $-\bar{M}_n$, taking a union bound over the two results, taking $(e-2)\lambda^2\langle\bar{V}_n,\rho\rangle$ to the other side of the inequality, and normalizing by λ, we obtain the statement of the theorem. ∎

Proof of Theorem 8: The value of λ that minimizes (9) depends on ρ, whereas we would like to have a result that holds for all possible distributions ρ simultaneously. This requires considering multiple values of λ simultaneously and we have

[2] By Lemma 2, for each $h \in \mathcal{H}$ we have $\mathbb{E}_{\bar{X}_1^n}\left[e^{n\,\mathrm{kl}(\frac{1}{n}\bar{S}_n(h)\|\bar{b}(h))}\right] \le n+1$ and, therefore, $\left\langle \mathbb{E}_{\bar{X}_1^n}\left[e^{n\,\mathrm{kl}(\frac{1}{n}\bar{S}_n\|\bar{b})}\right], \pi\right\rangle \le n+1$.

to take a union bound over λ-s in step (26) of the proof of Theorem 7. We cannot take all possible values of λ, since there are uncountably many possibilities. Instead we determine the relevant range of λ and take a union bound over a grid of λ-s that forms a geometric sequence over this range. Since the range is finite, the grid is also finite.

The upper bound on the relevant range of λ is determined by the constraint that $\lambda \leq \frac{1}{K}$. For the lower bound we note that since $\mathrm{KL}(\rho\|\pi) \geq 0$, the value of λ that minimizes (9) is lower bounded by $\sqrt{\frac{\ln\frac{2}{\delta}}{(e-2)\langle \bar{V}_n,\rho\rangle}}$. We also note that $\langle \bar{V}_n,\rho\rangle \leq K^2 n$, since $|Z_i(h)| \leq K$ for all h and i. Hence, $\lambda \geq \frac{1}{K}\sqrt{\frac{\ln\frac{2}{\delta}}{(e-2)n}}$ and the range of λ we are interested in is

$$\lambda \in \left[\frac{1}{K}\sqrt{\frac{\ln\frac{2}{\delta}}{(e-2)n}}, \frac{1}{K}\right].$$

We cover the above range with a grid of λ_i-s, such that $\lambda_i := c^i \frac{1}{K}\sqrt{\frac{\ln\frac{2}{\delta}}{(e-2)n}}$ for $i = 0, \ldots, m-1$. It is easy to see that in order to cover the interval of relevant λ we need

$$m = \left\lceil \frac{1}{\ln c} \ln\left(\sqrt{\frac{(e-2)n}{\ln\frac{2}{\delta}}}\right)\right\rceil.$$

(λ_{m-1} is the last value that is strictly less than $1/K$ and we take $\lambda_m := 1/K$ for the case when the technical condition (10) is not satisfied). This defines the value of ν in (12).

Finally, we note that (9) has the form $g(\lambda) = \frac{U}{\lambda} + \lambda V$. For the relevant range of λ, there is λ_{i^*} that satisfies $\sqrt{U/V} \leq \lambda_{i^*} < c\sqrt{U/V}$. For this value of λ we have $g(\lambda_{i^*}) \leq (1+c)\sqrt{UV}$.

Therefore, whenever (10) is satisfied we pick the highest value of λ_i that does not exceed the left hand side of (10), substitute it into (9), and obtain (11), where the $\ln\nu$ factor comes from the union bound over λ_i-s. If (10) is not satisfied, we know that $\langle \bar{V}_n,\rho\rangle < K^2\left(KL(\rho\|\pi) + \ln\frac{2\nu}{\delta}\right)/(e-2)$ and by taking $\lambda = 1/K$ and substituting into (9) we obtain (13). ∎

Proof of Theorem 6: Theorem 6 follows from Theorem 5 in the same way as Theorem 8 follows from Theorem 7. The only difference is that the relevant range of λ is unlimited from above. If $\mathrm{KL}(\rho\|\pi) = 0$ the bound is minimized by

$$\lambda = \sqrt{\frac{8\ln\frac{2}{\delta}}{\sum_{i=1}^n(\beta_i-\alpha_i)^2}},$$

hence, we are interested in λ that is larger or equal to this value. We take a grid of λ_i-s of the form

$$\lambda_i := c^i\sqrt{\frac{8\ln\frac{2}{\delta}}{\sum_{i=1}^n(\beta_i-\alpha_i)^2}}$$

for $i \geq 0$. Then for a given value of $\mathrm{KL}(\rho\|\pi)$ we have to pick λ_i, such that

$$i = \left\lfloor \frac{\ln\left(\frac{\mathrm{KL}(\rho\|\pi)}{\ln\frac{2}{\delta}} + 1\right)}{2\ln c}\right\rfloor,$$

where $\lfloor x\rfloor$ is the largest integer value that is smaller than x. Taking a weighted union bound over λ_i-s with weights $2^{-(i+1)}$ completes the proof. (In the weighted union bound we take $\delta_i = \delta 2^{-(i+1)}$. Then by substitution of δ with δ_i, (7) holds with probability greater than $1-\delta_i$ for each λ_i individually, and with probability greater than $1 - \sum_{i=0}^{\infty}\delta_i = 1-\delta$ for all λ_i simultaneously.) ∎

APPENDIX C
BACKGROUND

In this section we provide a proof of Lemma 11. The proof reproduces an intermediate step in the proof of [21, Theorem 1].

Proof of Lemma 11: First, we have:

$$\mathbb{E}_{Z_i}\left[e^{\lambda Z_i}\middle|Z_1^{i-1}\right] \leq \mathbb{E}_{Z_i}\left[1+\lambda Z_i + (e-2)\lambda^2(Z_i)^2\middle|Z_1^{i-1}\right] \tag{28}$$

$$= 1 + (e-2)\lambda^2\mathbb{E}_{Z_i}\left[(Z_i)^2\middle|Z_1^{i-1}\right] \tag{29}$$

$$\leq e^{(e-2)\lambda^2\mathbb{E}_{Z_i}\left[(Z_i)^2\middle|Z_1^{i-1}\right]}, \tag{30}$$

where (28) uses the fact that $e^x \leq 1 + x + (e-2)x^2$ for $x \leq 1$ (this restricts the choice of λ to $\lambda \leq \frac{1}{K}$, which leads to technical conditions (10) and (15) in Theorem 8 and Corollary 12, respectively); (29) uses the martingale property $\mathbb{E}_{Z_i}[Z_i|Z_1^{i-1}] = 0$; and (30) uses the fact that $1 + x \leq e^x$ for all x.

We apply inequality (30) in the following way:

$$\begin{aligned}
&\mathbb{E}_{Z_1^n}\left[e^{\lambda M_n - (e-2)\lambda^2 V_n}\right] \\
&= \mathbb{E}_{Z_1^n}\left[e^{\lambda M_{n-1} - (e-2)\lambda^2 V_{n-1} + \lambda Z_n - (e-2)\lambda^2\mathbb{E}\left[(Z_n)^2\middle|Z_1^{n-1}\right]}\right] \\
&= \mathbb{E}_{Z_1^{n-1}}\left[\begin{array}{l} e^{\lambda M_{n-1} - (e-2)\lambda^2 V_{n-1}} \\ \times\, \mathbb{E}_{Z_n}\left[e^{\lambda Z_n}\middle|Z_1^{n-1}\right] \times e^{-(e-2)\lambda^2\mathbb{E}\left[(Z_n)^2\middle|Z_1^{n-1}\right]}\end{array}\right] \\
&\leq \mathbb{E}_{Z_1^{n-1}}\left[e^{\lambda M_{n-1} - (e-2)\lambda^2 V_{n-1}}\right] \qquad (31) \\
&\leq \ldots \qquad (32) \\
&\leq 1.
\end{aligned}$$

Inequality (31) applies inequality (30) and inequality (32) recursively proceeds with $Z_{n-1}, \ldots, Z_1$ (in reverse order). ∎

Note that conditioning on additional variables in the proof of the lemma does not change the result. This fact is exploited in the proof of Theorem 7, when we allow interdependence between multiple martingales.

ACKNOWLEDGMENTS

The authors would like to thank Andreas Maurer for his comments on Lemma 1. We are also very grateful to the anonymous reviewers for their valuable comments that helped to improve the presentation of our work. This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886, and by the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement N^o270327. This publication only reflects the authors' views.

REFERENCES

[1] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

[2] K. Azuma, "Weighted sums of certain dependent random variables," *Tôhoku Mathematical Journal*, vol. 19, no. 3, 1967.

[3] S. N. Bernstein, *Probability Theory*, 4th ed., Moscow-Leningrad, 1946, in Russian.

[4] Y. Seldin, P. Auer, F. Laviolette, J. Shawe-Taylor, and R. Ortner, "PAC-Bayesian analysis of contextual bandits," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[5] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time." *Communications on Pure and Applied Mathematics*, vol. 28, 1975.

[6] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley-Interscience, 1997.

[7] R. M. Gray, *Entropy and Information Theory*, 2nd ed. Springer, 2011.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1991.

[9] J. Shawe-Taylor and R. C. Williamson, "A PAC analysis of a Bayesian estimator," in *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 1997.

[10] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony, "Structural risk minimization over data-dependent hierarchies," *IEEE Transactions on Information Theory*, vol. 44, no. 5, 1998.

[11] D. McAllester, "Some PAC-Bayesian theorems," in *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 1998.

[12] M. Seeger, "PAC-Bayesian generalization error bounds for Gaussian process classification," *Journal of Machine Learning Research*, 2002.

[13] L. G. Valiant, "A theory of the learnable," *Communications of the Association for Computing Machinery*, vol. 27, no. 11, 1984.

[14] J. Langford and J. Shawe-Taylor, "PAC-Bayes & margins," in *Advances in Neural Information Processing Systems (NIPS)*, 2002.

[15] D. McAllester, "PAC-Bayesian stochastic model selection," *Machine Learning*, vol. 51, no. 1, 2003.

[16] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand, "PAC-Bayesian learning of linear classifiers," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[17] Y. Seldin and N. Tishby, "PAC-Bayesian analysis of co-clustering and beyond," *Journal of Machine Learning Research*, vol. 11, 2010.

[18] G. Lever, F. Laviolette, and J. Shawe-Taylor, "Distribution-dependent PAC-Bayes priors," in *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2010.

[19] A. Maurer, "A note on the PAC-Bayesian theorem," www.arxiv.org, 2004.

[20] D. A. Freedman, "On tail probabilities for martingales," *The Annals of Probability*, vol. 3, no. 1, 1975.

[21] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire, "Contextual bandit algorithms with supervised learning guarantees," in *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[22] V. Mnih, C. Szepesvári, and J.-Y. Audibert, "Empirical Bernstein stopping," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[23] J. Y. Audibert, R. Munos, and C. Szepesvári, "Exploration-exploitation trade-off using variance estimates in multi-armed bandits," *Theoretical Computer Science*, 2009.

[24] A. Maurer and M. Pontil, "Empirical Bernstein bounds and sample variance penalization," in *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2009.

[25] M. Seeger, "Bayesian Gaussian process models: PAC-Bayesian generalization error bounds and sparse approximations," Ph.D. dissertation, University of Edinburgh, 2003.

Yevgeny Seldin received his Ph.D. in computer science from the Hebrew University of Jerusalem in 2010. Since 2009 he is a Research Scientist at the Max Planck Institute for Intelligent Systems in Tübingen and since 2011 he is also an Honorary Research Associate at the Department of Computer Science in University College London. His research interests include statistical learning theory, PAC-Bayesian analysis, and reinforcement learning. He has contributions in PAC-Bayesian analysis, reinforcement learning, clustering-based models in supervised and unsupervised learning, collaborative filtering, image processing, and bioinformatics.

François Laviolette received his Ph.D. in mathematics from Université de Montréal in 1997. His thesis solved a long-standing conjecture (60 years old) on graph theory and was among the seven finalists of the 1998 Council of Graduate Schools / University Microfilms International Distinguished Dissertation Award of Washington, in the category Mathematics-Physic-Engineering. He then moved to Université Laval, where he works on Probabilistic Verification of Systems, Bio-Informatics, and Machine Learning, with a particular interest in PAC-Bayesian analysis, for which he has already more than a dozen of scientific publications.

Nicolò Cesa-Bianchi is a faculty member of the Computer Science Department at the Università degli Studi di Milano, Italy. His main research interests include statistical learning theory, game-theoretic learning, and pattern analysis. He is co-author with Gàbor Lugosi of the monography "Prediction, Learning, and Games" (Cambridge University Press, 2006).

John Shawe-Taylor obtained a Ph.D. in Mathematics at Royal Holloway, University of London in 1986. He subsequently completed an M.Sc. in the Foundations of Advanced Information Technology at Imperial College. He was promoted to Professor of Computing Science in 1996. He has published over 200 research papers. In 2006 he was appointed Director of the Center for Computational Statistics and Machine Learning at University College London. He has pioneered the development of the well-founded approaches to Machine Learning inspired by statistical learning theory (including Support Vector Machine, Boosting and Kernel Principal Components Analysis) and has shown the viability of applying these techniques to document analysis and computer vision. He is co-author of an Introduction to Support Vector Machines, the first comprehensive account of this new generation of machine learning algorithms. A second book on Kernel Methods for Pattern Analysis was published in 2004.

Peter Auer received his Ph.D. in mathematics from the Vienna University of Technology in 1992, working on probability theory with Pal Revesz and on Symbolic Computation with Alexander Leitsch. He then moved to Graz University of Technology, working on Machine Learning with Wolfgang Maass, and was appointed associate professor in 1997. He has also been a research scholar at the University of California, Santa Cruz. In 2003 he accepted the position of a full professor for Information Technology at the Montanuniversität Leoben. He has authored scientific publications in the areas of probability theory, symbolic computation, and machine learning, he is a member of the editorial board of Machine Learning, and he has been principal investigator in several research projects funded by the European Union. His current research interests include Machine Learning focused on autonomous learning and exploration algorithms.