Wang et al. *Journal of Inequalities and Applications* (2015) 2015:244
DOI 10.1186/s13660-015-0768-3

Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

PAC-Bayesian inequalities of some random variables sequences

Zhen Wang[1], Luming Shen[2], Yu Miao[1*], Shanshan Chen[1] and Wenfei Xu[1]

*Correspondence:
yumiao728@126.com
[1]Henan Engineering Laboratory for Big Data Statistical Analysis and Optimal Control, College of Mathematics and Information Science, Henan Normal University, Xinxiang, Henan 453007, China
Full list of author information is available at the end of the article

Abstract

In this paper, we establish some PAC-Bayesian inequalities for a sequence of random variables, which include conditionally symmetric random variables and locally square integrable martingales.

MSC: 60G42; 60E15; 60G50

Keywords: PAC-Bayesian inequality; conditionally symmetric random variables; martingale

1 Introduction

In this paper, we establish several PAC-Bayesian inequalities. The PAC-Bayesian analysis is an abbreviation for the Probably Approximately Correct learning model and has been introduced a decade ago (Shawe-Taylor and Williamson [1]; Shawe-Taylor *et al.* [2]) and has made a significant contribution to the analysis and development of supervised learning methods, the random multiarmed bandits problem and so on. 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. It supplies generalization guarantees for many influential machine learning algorithms.

Shawe-Taylor and Williamson [1] established the PAC-Bayesian learning theorems. They showed that if one can find a ball of sufficient volume in a parameterized concept space, then the center of that ball has a low error rate. For non-i.i.d. data, application of PAC-Bayesian analysis was partially addressed only recently by Ralaivola *et al.* [3] and Lever *et al.* [4]. For a martingale, taking advantage of the martingale's properties, Seldin *et al.* [5] obtained the PAC-Bayesian-Bernstein inequality and applied it to multiarmed bandits. Seldin *et al.* [6] presented a generalization of the PAC-Bayesian analysis. Their generalization makes it possible to consider model order selection simultaneously with the exploration-exploitation trade-off.

In the present paper, we continue to study the PAC-Bayesian inequalities for some random variable sequence. We concentrate on the conditionally symmetric random variables and the locally square integrable martingale. For the first case, we only assume the conditional symmetry of the random variable sequence, without any other dependent conditions and moment conditions. For the second case, the bounded condition in Seldin *et al.* [6] is weakened. The paper is divided as follows: In Section 2 we state the main results and make some remarks. Proofs of the main results are provided in Section 3.

2 Main results

In this section, we discuss the PAC-Bayesian inequalities for conditionally symmetric random variables and martingales. In order to present our main theorems, we give a few definitions. Let $\mathbb{H}$ be an index (or a hypothesis) space, possibly uncountably infinite. Let $\{X_1(h), X_2(h), \ldots : h \in \mathbb{H}\}$ be a sequence of variables adapted to an increasing sequence of σ-fields $\{\mathcal{F}_n\}$, where $\mathcal{F}_n = \{X_k(h) : 1 \le k \le n \text{ and } h \in \mathbb{H}\}$ is a set of random sequences observed up to time n (the history).

Seldin *et al.* [6] obtained a PAC-Bayes-Bernstein inequality for martingale with finite jumps.

Theorem 2.1 (Seldin *et al.* [6]) *Let $\{X_1(h), X_2(h), \ldots : h \in \mathbb{H}\}$ be a set of martingale difference sequences adapted to an increasing sequence of σ-fields $\mathcal{F}_n = \{X_k(h) : 1 \le k \le n \text{ and } h \in \mathbb{H}\}$. Furthermore, let $M_n(h) = \sum_{k=1}^{n} X_k(h)$ be martingales corresponding to the martingale difference sequences and let $V_n(h) = \sum_{k=1}^{n} \mathbb{E}[X_k(h)^2 | \mathcal{F}_{k-1}]$ be cumulative variances of the martingales. For a distribution ρ over $\mathbb{H}$, define $M_n(\rho) = \mathbb{E}_{\rho(h)}[M_n(h)]$ and $V_n(\rho) = \mathbb{E}_{\rho(h)}[V_n(h)]$. Let $\{C_1, C_2, \ldots\}$ be an increasing sequence set in advance, such that $|X_k(h)| \le C_k$ for all h with probability* 1. *Let $\{\mu_1, \mu_2, \ldots\}$ be a sequence of 'reference' ('prior') distributions over $\mathbb{H}$, such that μ_n is independent of $\mathcal{F}_n$ (but can depend on n). Let $\{\lambda_1, \lambda_2, \ldots\}$ be a sequence of positive numbers set in advance that satisfy $\lambda_k \le C_k^{-1}$. Then for all possible distributions ρ_n over $\mathbb{H}$ given n and for all n simultaneously with probability greater than* $1 - \delta$,

$$\left|M_n(\rho_n)\right| \le (e-2)\lambda_n V_n(\rho_n) + \frac{\mathrm{KL}(\rho_n \parallel \mu_n) + 2\ln(n+1) + \ln\frac{2}{\delta}}{\lambda_n},$$

where $\mathrm{KL}(\mu \parallel \nu)$ *is the* KL*-divergence* (*relative entropy*) *between two distributions μ and ν.*

Seldin *et al.* [6] (see Theorem 2.1) considered the bounded martingale difference sequence and the parameters (λ_k) depend on the bounds $(1/C_k)$. Since C_k is an increasing sequence, λ_k is a decreasing sequence. Furthermore, Seldin *et al.* [6] studied the deviation properties between the martingale and the conditional variance of the martingale. Based on the above works, in the present paper, we want to consider the conditionally symmetric random variables and the locally square integrable martingale. For the conditionally symmetric random variables, we can establish the PAC-Bayesian inequality without any dependent assumptions (for example, independence or being a martingale) and any moment assumptions for the sequences. For the locally square integrable martingale, we can remove the bounded restriction.

2.1 Conditionally symmetric random variables

Assume that $\{X_k(h) : k \ge 1, h \in \mathbb{H}\}$ are conditionally symmetric with respect to $(\mathcal{F}_n)$ (*i.e.*, $\mathcal{L}(X_i(h)|\mathcal{F}_i) = \mathcal{L}(-X_i(h)|\mathcal{F}_i)$). Let

$$M_n(h) = \sum_{k=1}^{n} X_k(h) \quad \text{and} \quad V_n(h) = \sum_{k=1}^{n} \left[X_k(h)^2\right].$$

For a distribution ρ over $\mathbb{H}$, define $M_n(\rho) = \mathbb{E}_{\rho(h)}[M_n(h)]$ and $V_n(\rho) = \mathbb{E}_{\rho(h)}[V_n(h)]$. We establish the PAC-Bayesian inequality between the partial sums and the total quadratic variation of the partial sums.

Theorem 2.2 *Let* $\{\mu_1, \mu_2, \ldots\}$ *be a sequence of* '*reference*' ('*prior*') *distributions over* $\mathbb{H}$, *such that* μ_n *is independent of* $\mathcal{F}_n$ (*but can depend on n*). *Then for all possible distributions* ρ_n *over* $\mathbb{H}$ *given n and for all n simultaneously with probability greater than* $1-\delta$,

$$\left|M_n(\rho_n)\right| \le \frac{\lambda}{2} V_n(\rho_n) + \frac{\mathrm{KL}(\rho_n \parallel \mu_n) + 2\ln(n+1) + \ln\frac{2}{\delta}}{\lambda}, \quad \lambda > 0.$$

The following theorem gives an inequality in the sense of a self-normalized sequence.

Theorem 2.3 *Under the assumptions in Theorem* 2.2, *then for all* $y > 0$ *and n simultaneously with probability greater than* $1-\frac{\delta}{2}$,

$$\begin{aligned}&\mathbb{E}_{\rho_n(h)}\left(\ln\frac{y}{\sqrt{V_n(h)+y^2}} + \frac{M_n(h)^2}{2(V_n(h)+y^2)}\right)\\ &\quad\le \mathrm{KL}(\rho_n \parallel \mu_n) + 2\ln(n+1) + \ln\frac{2}{\delta}.\end{aligned}$$

2.2 Martingales

Let $(M_n(h))$ be a locally square integrable real martingale adapted to the filtration $(\mathcal{F}_n)$ with $M_0 = 0$. The predictable quadratic variation and the total quadratic variation of $(M_n(h))$ are, respectively, given by

$$\langle M\rangle_n(h) = \sum_{k=1}^{n} \mathbb{E}\left[\left(\Delta M_k(h)\right)^2 | \mathcal{F}_{k-1}\right] \quad \text{and} \quad [M]_n(h) = \sum_{k=1}^{n} \left(\Delta M_k(h)\right)^2,$$

where $\Delta M_n(h) = M_n(h) - M_{n-1}(h)$. For a distribution ρ over $\mathbb{H}$, define

$$M_n(\rho) = \mathbb{E}_{\rho(h)}\left[M_n(h)\right].$$

Theorem 2.4 *Let* $(M_n(h))$ *be a locally square integrable real martingale adapted to a filtration* $\mathbb{F} = (\mathcal{F}_n)$. *Let* $\{\mu_1, \mu_2, \ldots\}$ *be a sequence of* '*reference*' ('*prior*') *distributions over* $\mathbb{H}$, *such that* μ_n *is independent of* $\mathcal{F}_n$ (*but it can depend on n*). *Then for all possible distributions* ρ_n *over* $\mathbb{H}$ *given n and for all n simultaneously with probability greater than* $1-\delta$,

$$\left|M_n(\rho_n)\right| \le \frac{\lambda}{2}\left(\frac{[M]_n(\rho_n)}{3} + \frac{2\langle M\rangle_n(\rho_n)}{3}\right) + \frac{\mathrm{KL}(\rho_n \parallel \mu_n) + 2\ln(n+1) + \ln\frac{2}{\delta}}{\lambda}, \quad \lambda > 0.$$

Theorem 2.5 *Under the assumptions in Theorem* 2.4, *for all* $y > 0$ *and n simultaneously with probability greater than* $1-\frac{\delta}{2}$,

$$\begin{aligned}&\mathbb{E}_{\rho_n(h)}\left(\ln\frac{y}{\sqrt{(\frac{[M]_n(h)}{3} + \frac{2\langle M\rangle_n(h)}{3}) + y^2}} + \frac{M_n(h)^2}{2((\frac{[M]_n}{3} + \frac{2\langle M\rangle_n}{3}) + y^2)}\right)\\ &\quad\le \mathrm{KL}(\rho_n \parallel \mu_n) + 2\ln(n+1) + \ln\frac{2}{\delta}.\end{aligned}$$

3 The proofs of main results

Before giving the proofs of our results, we state the following basic lemmas.

Lemma 3.1 [7] *Let $\{d_i\}$ be a sequence of variables adapted to an increasing sequence of σ-fields $\{\mathcal{F}_i\}$. Assume that the $\{d_i\}$'s are conditionally symmetric. Then*

$$\exp\left(\lambda\sum_{i=1}^{n} d_i - \frac{\lambda^2}{2}\sum_{i=1}^{n} d_i^2\right), \quad n \geq 1,$$

is a supermartingale with mean ≤ 1, for all $\lambda \in \mathbb{R}$.

Lemma 3.2 *Under the assumptions of Lemma* 3.1, *for any $y > 0$, we have*

$$\mathbb{E}\frac{y}{\sqrt{\sum_{i=1}^{n} d_i^2 + y^2}} \exp\left(\frac{(\sum_{i=1}^{n} d_i)^2}{2(\sum_{i=1}^{n} d_i^2 + y^2)}\right) \leq 1.$$

Remark 3.1 Hitczenko [8] proved the above inequality for conditionally symmetric martingale difference sequences, and De la Peña [7] obtained the above inequality without the martingale difference assumption, hence without any integrability assumptions. Note that any sequence of real valued random variables X_i can be 'symmetrized' to produce an exponential supermartingale by introducing random variables X_i' such that

$$\mathcal{L}\big(X_i'|X_1, X_1', \ldots, X_{n-1}, X_{n-1}', X_n\big) = \mathcal{L}\big(X_n'|X_1, \ldots, X_{n-1}\big)$$

and we set $d_n = X_n - X_n'$.

Proof By using Fubini's theorem, we have

$$\begin{aligned}
&\mathbb{E}\frac{y}{\sqrt{\sum_{i=1}^{n} d_i^2 + y^2}} \exp\left(\frac{(\sum_{i=1}^{n} d_i)^2}{2(\sum_{i=1}^{n} d_i^2 + y^2)}\right) \\
&\quad = \mathbb{E}\left[\frac{y}{\sqrt{\sum_{i=1}^{n} d_i^2 + y^2}} \exp\left(\frac{(\sum_{i=1}^{n} d_i)^2}{2(\sum_{i=1}^{n} d_i^2 + y^2)}\right)\right. \\
&\qquad \left. \times \frac{\sqrt{\sum_{i=1}^{n} d_i^2 + y^2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{\frac{\sum_{i=1}^{n} d_i^2 + y^2}{2}\left(\lambda - \frac{\sum_{i=1}^{n} d_i^2}{\sum_{i=1}^{n} d_i^2 + y^2}\right)^2\right\} d\lambda\right] \\
&\quad = \int_{-\infty}^{\infty} \mathbb{E}\left[\frac{y}{\sqrt{2\pi}} \exp\left(\lambda\sum_{i=1}^{n} d_i - \frac{\lambda^2}{2}\sum_{i=1}^{n} d_i^2\right) \exp\left(-\frac{\lambda^2 y^2}{2}\right)\right] d\lambda \\
&\quad \leq \int_{-\infty}^{\infty}\left[\frac{y}{\sqrt{2\pi}} \exp\left(-\frac{\lambda^2 y^2}{2}\right)\right] d\lambda = 1,
\end{aligned}$$

where we used the fact

$$\frac{\sqrt{\sum_{i=1}^{n} d_i^2 + y^2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{\frac{\sum_{i=1}^{n} d_i^2 + y^2}{2}\left(\lambda - \frac{\sum_{i=1}^{n} d_i^2}{\sum_{i=1}^{n} d_i^2 + y^2}\right)^2\right\} d\lambda = 1. \qquad \square$$

The following inequality is about a transformation of the measure inequality [9].

Lemma 3.3 *For any measurable function $\phi(h)$ on $\mathbb{H}$ and any distributions $\mu(h)$ and $\rho(h)$ on $\mathbb{H}$, we have*

$$\mathbb{E}_{\rho(h)}\big(\phi(h)\big) \leq \mathrm{KL}(\rho \parallel \mu) + \ln \mathbb{E}_{\mu(h)}\big(e^{\phi(h)}\big).$$

Proof of Theorem 2.2 Taking

$$\phi(h) = \lambda M_n(h) - \frac{\lambda^2}{2} V_n(h), \quad \lambda > 0,$$

then from Lemma 3.1 and Lemma 3.3, for all ρ_n and n simultaneously with probability greater than $1 - \frac{\delta}{2}$,

$$\begin{aligned}
&\lambda M_n(\rho_n) - \frac{\lambda^2}{2} V_n(\rho_n) \\
&\quad = \mathbb{E}_{\rho_n(h)}\left(\lambda M_n(h) - \frac{\lambda^2}{2} V_n(h)\right) \\
&\quad \le \mathrm{KL}(\rho_n \parallel \mu_n) + \ln \mathbb{E}_{\mu_n(h)}\left(e^{\lambda M_n(h) - \frac{\lambda^2}{2} V_n(h)}\right) \\
&\quad \le \mathrm{KL}(\rho_n \parallel \mu_n) + \ln \mathbb{E}_{\mathcal{F}_n}\left(\mathbb{E}_{\mu_n(h)} e^{\lambda M_n(h) - \frac{\lambda^2}{2} V_n(h)}\right) + 2\ln(n+1) + \ln\frac{2}{\delta} \\
&\quad = \mathrm{KL}(\rho_n \parallel \mu_n) + \ln \mathbb{E}_{\mu_n(h)}\left(\mathbb{E}_{\mathcal{F}_n} e^{\lambda M_n(h) - \frac{\lambda^2}{2} V_n(h)}\right) + 2\ln(n+1) + \ln\frac{2}{\delta} \\
&\quad \le \mathrm{KL}(\rho_n \parallel \mu_n) + 2\ln(n+1) + \ln\frac{2}{\delta},
\end{aligned}$$

where the second equality is due to the fact that μ_n is independent of $\mathcal{F}_n$. By applying the same argument to martingales $-M_n(h)$, we obtain the result that, with probability greater than $1 - \delta$,

$$\left|M_n(\rho_n)\right| \le \frac{\lambda}{2} V_n(\rho_n) + \frac{\mathrm{KL}(\rho_n \parallel \mu_n) + 2\ln(n+1) + \ln\frac{2}{\delta}}{\lambda}, \quad \lambda > 0. \qquad \square$$

Proof of Theorem 2.3 For all $y > 0$, taking

$$\phi(h) = \ln \frac{y}{\sqrt{V_n(h) + y^2}} + \frac{M_n(h)^2}{2(V_n(h) + y^2)},$$

then from Lemma 3.2 and Lemma 3.3, for all ρ_n and n simultaneously with probability greater than $1 - \frac{\delta}{2}$,

$$\begin{aligned}
&\mathbb{E}_{\rho_n(h)}\left(\ln \frac{y}{\sqrt{V_n(h) + y^2}} + \frac{M_n(h)^2}{2(V_n(h) + y^2)}\right) \\
&\quad \le \mathrm{KL}(\rho_n \parallel \mu_n) + \ln \mathbb{E}_{\mu_n(h)}\left(e^{\ln \frac{y}{\sqrt{V_n(h)+y^2}} + \frac{M_n(h)^2}{2(V_n(h)+y^2)}}\right) \\
&\quad \le \mathrm{KL}(\rho_n \parallel \mu_n) + \ln \mathbb{E}_{\mathcal{F}_n}\left(\mathbb{E}_{\mu_n(h)} e^{\ln \frac{y}{\sqrt{V_n(h)+y^2}} + \frac{M_n(h)^2}{2(V_n(h)+y^2)}}\right) + 2\ln(n+1) + \ln\frac{2}{\delta} \\
&\quad = \mathrm{KL}(\rho_n \parallel \mu_n) + \ln \mathbb{E}_{\mu_n(h)}\left(\mathbb{E}_{\mathcal{F}_n} e^{\ln \frac{y}{\sqrt{V_n(h)+y^2}} + \frac{M_n(h)^2}{2(V_n(h)+y^2)}}\right) + 2\ln(n+1) + \ln\frac{2}{\delta} \\
&\quad \le \mathrm{KL}(\rho_n \parallel \mu_n) + 2\ln(n+1) + \ln\frac{2}{\delta}.
\end{aligned}$$

$\square$

In order to prove Theorem 2.4, we need to introduce the following lemma.

Lemma 3.4 *Let (M_n) be a locally square integrable martingale. Putting*

$$\langle M\rangle_n = \sum_{k=1}^{n} \mathbb{E}\big[(\Delta M_k)^2|\mathcal{F}_{k-1}\big] \quad \textit{and} \quad [M]_n = \sum_{k=1}^{n} (\Delta M_k)^2,$$

for all $t \in \mathbb{R}$ and $n \geq 0$, denote

$$G_n(t) = \exp\left(tM_n - \frac{t^2}{6}[M]_n - \frac{t^2}{3}\langle M\rangle_n\right).$$

Then, for all $t \in \mathbb{R}$, $(G_n(t))$ is a positive supermartingale with $\mathbb{E}[G_n(t)] \leq 1$.

Proof Let X be a square integrable random variable with $\mathbb{E}X = 0$ and $0 < \sigma^2 := \mathbb{E}X^2 < \infty$. Because of the basic inequality

$$\exp\left(x - \frac{x^2}{6}\right) \leq 1 + x + \frac{x^2}{3}, \quad x \in \mathbb{R},$$

we know

$$\mathbb{E}\left[\exp\left(tX - \frac{t^2}{6}X^2\right)\right] \leq 1 + \frac{t^2}{3}\sigma^2. \tag{3.1}$$

Then, for all $t \in \mathbb{R}$ and $n \geq 0$, we obtain from (3.1)

$$G_n(t) = G_n(t-1)\exp\left(t\Delta M_n - \frac{t^2}{6}\Delta[M]_n - \frac{t^2}{3}\Delta\langle M\rangle_n\right).$$

Hence, we deduce that, for all $t \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\big[G_n(t)|\mathcal{F}_{n-1}\big] &\leq G_n(t-1)\exp\left(-\frac{t^2}{3}\Delta\langle M\rangle_n\right)\cdot\left(1 + \frac{t^2}{3}\Delta\langle M\rangle_n\right)\\ &= G_n(t-1).\end{aligned}$$

As a result, for all $t \in \mathbb{R}$, $G_n(t)$ is a positive supermartingale, *i.e.*, for all $n \geq 1$, $\mathbb{E}[G_n(t)] \leq \mathbb{E}[G_{n-1}(t)]$, which implies that $\mathbb{E}[G_n(t)] \leq \mathbb{E}[G_0(t)] = 1$. □

Proof of Theorem 2.4 Taking

$$\phi(h) = \lambda M_n(h) - \frac{\lambda^2}{2}\left(\frac{[M]_n(h)}{3} + \frac{2\langle M\rangle_n(h)}{3}\right), \quad \lambda > 0,$$

then from Lemma 3.3 and Lemma 3.4, for all ρ_n and n simultaneously with probability greater than $1 - \frac{\delta}{2}$,

$$\begin{aligned}&\lambda M_n(\rho_n) - \frac{\lambda^2}{2}\left(\frac{[M]_n(\rho_n)}{3} + \frac{2\langle M\rangle_n(\rho_n)}{3}\right)\\ &\quad = \mathbb{E}_{\rho_n(h)}\left[\lambda M_n(h) - \frac{\lambda^2}{2}\left(\frac{[M]_n(h)}{3} + \frac{2\langle M\rangle_n(h)}{3}\right)\right]\end{aligned}$$

$$
\begin{aligned}
&\le \mathrm{KL}(\rho_n \parallel \mu_n) + \ln \mathbb{E}_{\mu_n(h)}\left[e^{\lambda M_n(h) - \frac{\lambda^2}{2}\left(\frac{[M]_n(h)}{3} + \frac{2\langle M\rangle_n(h)}{3}\right)}\right] \\
&\le \mathrm{KL}(\rho_n \parallel \mu_n) + \ln \mathbb{E}_{\mathcal{F}_n}\left[\mathbb{E}_{\mu_n(h)} e^{\lambda M_n(h) - \frac{\lambda^2}{2}\left(\frac{[M]_n(h)}{3} + \frac{2\langle M\rangle_n(h)}{3}\right)}\right] + 2\ln(n+1) + \ln\frac{2}{\delta} \\
&= \mathrm{KL}(\rho_n \parallel \mu_n) + \ln \mathbb{E}_{\mu_n(h)}\left[\mathbb{E}_{\mathcal{F}_n} e^{\lambda M_n(h) - \frac{\lambda^2}{2}\left(\frac{[M]_n(h)}{3} + \frac{2\langle M\rangle_n(h)}{3}\right)}\right] + 2\ln(n+1) + \ln\frac{2}{\delta} \\
&\le \mathrm{KL}(\rho_n \parallel \mu_n) + 2\ln(n+1) + \ln\frac{2}{\delta},
\end{aligned}
$$

where the second equality is due to the fact that μ_n is independent of $\mathcal{F}_n$. By applying the same argument to martingales $-M_n(h)$, we obtain the result that, with probability greater than $1-\delta$,

$$
\left|M_n(\rho_n)\right| \le \frac{\lambda}{2}\left(\frac{[M]_n(\rho_n)}{3} + \frac{2\langle M\rangle_n(\rho_n)}{3}\right) + \frac{\mathrm{KL}(\rho_n\|\mu_n) + 2\ln(n+1) + \ln\frac{2}{\delta}}{\lambda}, \quad \lambda > 0. \ \square
$$

Proof of Theorem 2.5 For all $y > 0$, taking

$$
\phi(h) = \ln\frac{y}{\sqrt{\left(\frac{[M]_n(h)}{3} + \frac{2\langle M\rangle_n(h)}{3}\right) + y^2}} + \frac{M_n(h)^2}{2\left(\left(\frac{[M]_n(h)}{3} + \frac{2\langle M\rangle_n(h)}{3}\right) + y^2\right)},
$$

and $V_n(h) = [\frac{[M]_n(h)}{3} + \frac{2\langle M\rangle_n(h)}{3}]$, from Lemma 3.2 and Lemma 3.3, we can get for all ρ_n and n simultaneously with probability greater than $1 - \frac{\delta}{2}$,

$$
\begin{aligned}
&\mathbb{E}_{\rho_n(h)}\left(\ln\frac{y}{\sqrt{\left(\frac{[M]_n(h)}{3} + \frac{2\langle M\rangle_n(h)}{3}\right) + y^2}} + \frac{M_n(h)^2}{2\left(\left(\frac{[M]_n(h)}{3} + \frac{2\langle M\rangle_n(h)}{3}\right) + y^2\right)}\right) \\
&= \mathbb{E}_{\rho_n(h)}\left(\ln\frac{y}{\sqrt{V_n(h) + y^2}} + \frac{M_n(h)^2}{2(V_n(h) + y^2)}\right) \\
&\le \mathrm{KL}(\rho_n \parallel \mu_n) + \ln \mathbb{E}_{\mu_n(h)}\left(e^{\ln\frac{y}{\sqrt{V_n(h)+y^2}} + \frac{M_n(h)^2}{2(V_n(h)+y^2)}}\right) \\
&\le \mathrm{KL}(\rho_n \parallel \mu_n) + \ln \mathbb{E}_{\mathcal{F}_n}\left(\mathbb{E}_{\mu_n(h)} e^{\ln\frac{y}{\sqrt{V_n(h)+y^2}} + \frac{M_n(h)^2}{2(V_n(h)+y^2)}}\right) + 2\ln(n+1) + \ln\frac{2}{\delta} \\
&= \mathrm{KL}(\rho_n \parallel \mu_n) + \ln \mathbb{E}_{\mu_n(h)}\left(\mathbb{E}_{\mathcal{F}_n} e^{\ln\frac{y}{\sqrt{V_n(h)+y^2}} + \frac{M_n(h)^2}{2(V_n(h)+y^2)}}\right) + 2\ln(n+1) + \ln\frac{2}{\delta} \\
&\le \mathrm{KL}(\rho_n \parallel \mu_n) + 2\ln(n+1) + \ln\frac{2}{\delta}.
\end{aligned}
\qquad \square
$$

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors contributed equally to the manuscript, and they read and approved the final manuscript.

Author details
[1] Henan Engineering Laboratory for Big Data Statistical Analysis and Optimal Control, College of Mathematics and Information Science, Henan Normal University, Xinxiang, Henan 453007, China. [2] Science College, Hunan Agricultural University, Changsha, Hunan 410128, China.

Acknowledgements
This work is supported by IRTSTHN (14IRTSTHN023), NSFC (11471104), NCET (NCET-11-0945).

Received: 23 March 2015 Accepted: 24 July 2015 Published online: 07 August 2015

References

1. Shawe-Taylor, J, Williamson, RC: A PAC analysis of a Bayesian estimator. In: Proceedings of the International Conference on COLT, pp. 2-9 (1997)
2. Shawe-Taylor, J, Bartlett, PL, Williamson, RC, Anthony, M: Structural risk minimization over data-dependent hierarchies. IEEE Trans. Inf. Theory **44**(5), 1926-1940 (1998)
3. Ralaivola, L, Szafranski, M, Stempfel, G: Chromatic PAC-Bayes bounds for non-IID data: applications to ranking and stationary β-mixing processes. J. Mach. Learn. Res. **11**, 1927-1956 (2010)
4. Lever, G, Laviolette, F, Shawe-Taylor, J: Distribution-dependent PAC-Bayes priors. In: Algorithmic Learning Theory. Lecture Notes in Computer Science, vol. 6331, pp. 119-133 (2010)
5. Seldin, Y, Laviolette, F, Cesa-Bianchi, N, Shawe-Taylar, J, Auer, P: PAC-Bayesian inequalities for martingales. IEEE Trans. Inf. Theory **58**, 7086-7093 (2012)
6. Seldin, Y, Cesa-Bianchi, N, Auer, P, Laviolette, F, Shawe-Taylor, J: PAC-Bayes-Bernstein inequality for martingales and its application to multiarmed bandits. In: JMLR: Workshop and Conference Proceedings, vol. 26, pp. 98-111 (2012)
7. De la Peña, VH: A general class of exponential inequalities for martingales and ratios. Ann. Probab. **27**(1), 537-564 (1999)
8. Hitczenko, P: Upper bounds for the L_p-norms of martingales. Probab. Theory Relat. Fields **86**(2), 225-238 (1990)
9. Dupuis, P, Ellis, RS: A Weak Convergence Approach to the Theory of Large Deviations. Wiley Series in Probability and Statistics. Wiley, New York (2011)