PAC-Bayesian Inequalities for Martingales

Ilya Tolstikhin
Russian Academy of Sciences

GRAAL, Laval University
December 2013

Short summary

Classic PAC-Bayesian analysis deals with i.i.d. case:

- multiple sequences of r.v. $\{X_1^h, \ldots, X_n^h\}$ indexed by $h \in \mathcal{H}$;
- the set $\mathcal{H}$ is possibly uncountable;
- sequences are *interdependent* for different h;
- ... but for fixed $h \in \mathcal{H}$ r.v. $\{X_1^h, \ldots, X_n^h\}$ are i.i.d.
- **Goal**: obtain bound on $\mathbb{E}_{h\sim\rho}\big[\mathbb{E}[X_1^h]\big]$ for ρ over $\mathcal{H}$.

Example: Statistical Learning Theory

- $\mathcal{H} = \{h\colon \mathcal{X} \to \mathcal{Y}\}$ — hypothesis class of predictors;
- $X_i^h = \ell\big(y_i, h(x_i)\big)$, where
- $\{(x_i, y_i)\}_{i=1}^n$ — training sample $\overset{iid}{\sim} \mathcal{D}^n$;
- $\mathcal{D}$ — unknown distribution over input×label space $\mathcal{X} \times \mathcal{Y}$;
- $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ — loss function.

Question: Can we do the same for non-i.i.d. sequences $\{X_i^h\}$?

Motivation: online learning, bandits, any other sequential games,...

Outline

Part 0: Obtaining PAC-Bayesian inequalities

Part I: Basic martingale definitions

Part II: Concentration inequalities for individual martingales

- Hoeffding-Azuma inequality (Azuma, 1967; Hoeffding, 1963)
- Bernstein inequality (Bernstein, 1946; Freedman, 1975)
- kl-inequality (Seldin et al., 2012)
- Comparison.

Part III: PAC-Bayesian inequalities for martingales (Seldin et al., 2012)

- PAC-Bayes-kl inequality
- PAC-Bayes-Hoeffding-Azuma inequality
- PAC-Bayes-Bernstein inequality

Outline

Part 0: Obtaining PAC-Bayesian inequalities

Part I: Basic martingale definitions

Part II: Concentration inequalities for individual martingales

- Hoeffding-Azuma inequality (Azuma, 1967; Hoeffding, 1963)
- Bernstein inequality (Bernstein, 1946; Freedman, 1975)
- kl-inequality (Seldin et al., 2012)
- Comparison.

Part III: PAC-Bayesian inequalities for martingales (Seldin et al., 2012)

- PAC-Bayes-kl inequality
- PAC-Bayes-Hoeffding-Azuma inequality
- PAC-Bayes-Bernstein inequality

Obtaining PAC-Bayesian inequalities

Lemma (Donsker, Varadhan, 1975)

For two distributions ρ and π over $\mathcal{H}$ we have:

$$\mathrm{KL}(\rho\|\pi) = \mathbb{E}_{h\sim\rho}\left[\ln\frac{\rho(h)}{\pi(h)}\right] = \sup_f\left(\mathbb{E}_{h\sim\rho}[f(h)] - \ln\mathbb{E}_{h\sim\pi}\left[e^{f(h)}\right]\right)$$

where supremum is taken over all measurable functions $f\colon \mathcal{H}\to\mathbb{R}$.

- Thus for any f following holds simultaneously for all pairs (π,ρ):

$$\mathbb{E}_{h\sim\rho}[f(h)] \le \mathrm{KL}(\rho\|\pi) + \ln\mathbb{E}_{h\sim\pi}\left[e^{f(h)}\right].$$

- f could also depend on any sample $S = \{X_1,\ldots,X_n\}\sim\mathcal{D}$:

$$f_n\colon \mathcal{H}\times S\to\mathbb{R}.$$

Obtaining PAC-Bayesian inequalities

For any $f_n : \mathcal{H} \times S \to \mathbb{R}$ where $S = \{X_1, \ldots, X_n\} \sim \mathcal{D}$ following holds simultaneously for all π, ρ, and S:

$$\mathbb{E}_{h\sim\rho}[f_n(h,S)] \leq \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{h\sim\pi}\left[e^{f_n(h,S)}\right].$$

Consider that π does not depend on S. Using Markov's inequality we obtain with prob. greater than $1-\delta$ over random draw of S from $\mathcal{D}$:

$$\begin{aligned}
\mathbb{E}_{h\sim\rho}[f_n(h,S)] &\leq \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{h\sim\pi}\left[e^{f_n(h,S)}\right] \\
&\leq \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathbb{E}_{S\sim\mathcal{D}}\left[\mathbb{E}_{h\sim\pi}\left[e^{f_n(h,S)}\right]\right]\right) \\
&= \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathbb{E}_{h\sim\pi}\left[\mathbb{E}_{S\sim\mathcal{D}}\left[e^{f_n(h,S)}\right]\right]\right)
\end{aligned}$$

simultaneously for all ρ.

- Choose proper function f_n
- Upper bound the m.g.d. $\mathbb{E}_{S\sim\mathcal{D}}\left[e^{f_n(h,S)}\right]$.

Outline

Part 0: Obtaining PAC-Bayesian inequalities

Part I: Basic martingale definitions

Part II: Concentration inequalities for individual martingales

- Hoeffding-Azuma inequality (Azuma, 1967; Hoeffding, 1963)
- Bernstein inequality (Bernstein, 1946; Freedman, 1975)
- kl-inequality (Seldin et al., 2012)
- Comparison.

Part III: PAC-Bayesian inequalities for martingales (Seldin et al., 2012)

- PAC-Bayes-kl inequality
- PAC-Bayes-Hoeffding-Azuma inequality
- PAC-Bayes-Bernstein inequality

Martingales (basics)

- A sequence of r. v. $\{M_1, \ldots, M_n\}$ is a **martingale** if:

$$\mathbb{E}[|M_i|] < \infty, \text{ for } 1 \leq i \leq n;$$
$$\mathbb{E}[M_{i+1}|M_1, \ldots, M_i] = M_i, \text{ for } 1 \leq i \leq n-1.$$

Example: *If G_t is a gain of a gambler in a fair game at time t then $\{G_t\}$ is (supposed to be) a martingale.*

- $\{Z_1, \ldots, Z_n\}$ is a **martingale difference sequence (m.d.s.)** if:

$$\mathbb{E}[|Z_i|] < \infty, \text{ for } 1 \leq i \leq n;$$
$$\mathbb{E}[Z_{i+1}|Z_1, \ldots, Z_i] = 0, \text{ for } 0 \leq i \leq n-1.$$

Properties:

- If $\{M_i\}_{i=1}^n$ is a martingale then $\mathbb{E}[M_i] = \mathbb{E}[M_{i+1}]$ for $1 \leq i \leq n-1$.
- $\{M_i\}_{i=1}^n$ is a martingale iff $M_i = c + \sum_{i=1}^i Z_i$ for some m.d.s. $\{Z_i\}$.
- If $\{X_i\}$ are independent then $M_i = \sum_{j=1}^i (X_j - \mathbb{E}[X_j])$ is a martingale.

Outline

Part 0: Obtaining PAC-Bayesian inequalities

Part I: Basic martingale definitions

Part II: Concentration inequalities for individual martingales

- Hoeffding-Azuma inequality (Azuma, 1967; Hoeffding, 1963)
- Bernstein inequality (Bernstein, 1946; Freedman, 1975)
- kl-inequality (Seldin et al., 2012)
- Comparison.

Part III: PAC-Bayesian inequalities for martingales (Seldin et al., 2012)

- PAC-Bayes-kl inequality
- PAC-Bayes-Hoeffding-Azuma inequality
- PAC-Bayes-Bernstein inequality

Concentration inequalities: Chernoff's methog

How to upper bound $\mathbb{P}\{\xi \geq t\}$ for a random variable ξ?

- For any $\lambda \geq 0$:
$$\mathbb{P}\{\xi \geq t\} = \mathbb{P}\left\{e^{\lambda\xi} \geq e^{\lambda t}\right\}.$$
- For nonnegative r.v. $e^{\lambda\xi}$ apply Markov's inequality:
$$\mathbb{P}\left\{e^{\lambda\xi} \geq e^{\lambda t}\right\} \leq \frac{\mathbb{E}\left[e^{\lambda\xi}\right]}{e^{\lambda t}}.$$
- Upper bound on the moment generating function:
$$\mathbb{E}\left[e^{\lambda\xi}\right] \leq F(\lambda).$$
- Optimize the bound w.r.t. $\lambda \geq 0$:
$$\mathbb{P}\{\xi \geq t\} \leq \min_{\lambda \geq 0} \frac{F(\lambda, n)}{e^{\lambda t}}.$$

Outline

Part 0: Obtaining PAC-Bayesian inequalities

Part I: Basic martingale definitions

Part II: Concentration inequalities for individual martingales

- Hoeffding-Azuma inequality (Azuma, 1967; Hoeffding, 1963)
- Bernstein inequality (Bernstein, 1946; Freedman, 1975)
- kl-inequality (Seldin et al., 2012)
- Comparison.

Part III: PAC-Bayesian inequalities for martingales (Seldin et al., 2012)

- PAC-Bayes-kl inequality
- PAC-Bayes-Hoeffding-Azuma inequality
- PAC-Bayes-Bernstein inequality

Hoeffding's inequality for i.i.d. case

Lemma (Hoeffding, 1963)

Let ξ be a bounded random variable $a \le \xi \le b$ with $\mathbb{E}[\xi] = 0$. Then for all $\lambda \ge 0$:

$$\mathbb{E}\left[e^{\lambda\xi}\right] \le e^{\lambda^2(b-a)^2/8}.$$

Hoeffding's inequality :

Let $\{X_1, \ldots, X_n\}$ be independent bounded r.v. with $\mathbb{E}[X_i] = 0$, and $a_i \le X_i \le b_i$. Then for all $t > 0$ and any $\lambda \ge 0$:

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} X_i\right)\right] &= \prod_{i=1}^{n} \mathbb{E}\left[\exp \lambda X_i\right] \\ &\le \exp\left(\frac{1}{8}\lambda^2 \sum_{i=1}^{n} (b_i - a_i)^2\right).\end{aligned}$$

Azuma-Hoeffding inequality for martingales

Let $\{Z_1, \ldots, Z_n\}$ be a m.d.s. such that $a_i \leq Z_i \leq b_i$.
Denote $Z_1^i = \{Z_1, \ldots, Z_i\}$ for $1 \leq i \leq n$. For any $\lambda \geq 0$:

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} Z_i\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} Z_i\right) \middle| Z_1^{n-1}\right]\right]$$
$$= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} Z_i\right) \mathbb{E}\left[e^{\lambda Z_n} \middle| Z_1^{n-1}\right]\right].$$

But $\mathbb{E}[Z_n | Z_1^{n-1}] = 0$. Apply Hoeffding's lemma and get:

$$\cdots \leq \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} Z_i\right) e^{\lambda^2 (b_n - a_n)^2 / 8}\right] \leq \ldots$$

and repeat:

$$\cdots \leq \exp\left(\frac{1}{8}\lambda^2 \sum_{i=1}^{n} (b_i - a_i)^2\right).$$

Azuma-Hoeffding's inequality for martingales

Combine with Chernoff's method and get for any $\lambda \geq 0$:

$$\mathbb{P}\left\{\sum_{i=1}^{n} Z_i \geq t\right\} \leq \frac{\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} Z_i\right)\right]}{e^{\lambda t}}$$

$$\leq \exp\left(\frac{1}{8}\lambda^2 \sum_{i=1}^{n}(b_i - a_i)^2 - \lambda t\right) \to \min_{\lambda \geq 0}.$$

Choose $\lambda^* = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}$, denote r.h.s. with δ, solve for t, and get:

Theorem (Azuma-Hoeffgind's inequality)

Let $\{Z_1, \ldots, Z_n\}$ be m.d.s. with $a_i \leq X_i \leq b_i$. Then for any $\delta \in (0, 1)$ with probability greater than $1 - \delta$ (over Z_1^n):

$$\sum_{i=1}^{n} Z_i \leq \sqrt{\frac{1}{2} \ln \frac{1}{\delta} \sum_{i=1}^{n}(b_i - a_i)^2}.$$

Azuma-Hoeffding's inequality for martingales

Combine with Chernoff's method and get for any $\lambda \geq 0$:

$$\mathbb{P}\left\{\sum_{i=1}^{n} Z_i \geq t\right\} \leq \frac{\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} Z_i\right)\right]}{e^{\lambda t}}$$

$$\leq \exp\left(\frac{1}{8}\lambda^2 \sum_{i=1}^{n}(b_i - a_i)^2 - \lambda t\right) \to \min_{\lambda \geq 0}.$$

Choose $\lambda^* = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}$, denote r.h.s. with δ, solve for t, and get:

Theorem (Azuma-Hoeffgind's inequality)

Let $\{Z_1, \ldots, Z_n\}$ be m.d.s. with $a_i \leq X_i \leq b_i$. Then for any $\delta \in (0, 1)$ with probability greater than $1 - \delta$ (over Z_1^n):

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

Outline

Part 0: Obtaining PAC-Bayesian inequalities

Part I: Basic martingale definitions

Part II: Concentration inequalities for individual martingales

- Hoeffding-Azuma inequality (Azuma, 1967; Hoeffding, 1963)
- Bernstein inequality (Bernstein, 1946; Freedman, 1975)
- kl-inequality (Seldin et al., 2012)
- Comparison.

Part III: PAC-Bayesian inequalities for martingales (Seldin et al., 2012)

- PAC-Bayes-kl inequality
- PAC-Bayes-Hoeffding-Azuma inequality
- PAC-Bayes-Bernstein inequality

Bernstein-type inequality for i.i.d. case

Lemma (Bennett, 1962)

Let ξ be a bounded random variable $\xi \leq 1$ with $\mathbb{E}[\xi] = 0$ and denote $\psi(s) = e^s - s - 1$. Then for all $\lambda \geq 0$:

$$\mathbb{E}\left[e^{\lambda\xi}\right] \leq \exp\left(\psi(\lambda)\mathbb{E}[\xi^2]\right).$$

Noting that $\psi(s) \leq (e-2)s^2$ for $s \in [0,1]$, we also have for $\lambda \in [0,1]$:

$$\mathbb{E}\left[e^{\lambda\xi}\right] \leq \exp\left((e-2)\lambda^2\mathbb{E}[\xi^2]\right).$$

Bernstein-type inequality:

Let $\{X_1, \ldots, X_n\}$ be independent bounded r.v. with $\mathbb{E}[X_i] = 0$, and $|X_i| \leq 1$. Then for all $t > 0$ and any $0 \leq \lambda \leq 1$:

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} X_i\right)\right] = \prod_{i=1}^{n} \mathbb{E}\left[\exp \lambda X_i\right]$$

$$\leq \exp\left((e-2)\lambda^2 \mathbb{V}\left[\sum_{i=1}^{n} X_i\right]\right).$$

Bernstein-type inequality for martingales

Let $\{Z_1, \ldots, Z_n\}$ be a m.d.s. such that $|Z_i| \leq 1$.

Denote $Z_1^i = \{Z_1, \ldots, Z_i\}$ and $V_i = \sum_{j=1}^{i} \mathbb{V}\left[Z_j | Z_1^{j-1}\right]$. For any $\lambda \geq 0$:

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} Z_i - (e-2)\lambda^2 V_n\right)\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} Z_i - (e-2)\lambda^2 V_n\right) \middle| Z_1^{n-1}\right]\right]$$

$$= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} Z_i - (e-2)\lambda^2 V_{n-1}\right) \mathbb{E}\left[e^{\lambda Z_n} \middle| Z_1^{n-1}\right] e^{-(e-2)\lambda^2 \mathbb{V}\left[Z_n | Z_1^{n-1}\right]}\right].$$

But $\mathbb{E}\left[Z_n | Z_1^{n-1}\right] = 0$. Apply Bennett's lemma and get for any $\lambda \in [0, 1]$:

$$\cdots \leq \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} Z_i - (e-2)\lambda^2 V_{n-1}\right)\right] \leq \ldots$$

and repeat:

$$\cdots \leq 1.$$

Bernstein-type inequality for martingales

Denote $\eta_\lambda = \lambda \sum_{i=1}^n Z_i - (e-2)\lambda^2 V_n$. Combine with Chernoff's method and get for any $\lambda \in [0,1]$:

$$\mathbb{P}\{\eta_\lambda \geq t\} \leq \frac{\mathbb{E}\left[e^{\eta_\lambda}\right]}{e^t} \leq e^{-t}.$$

Denoting r.h.s. by δ and solving for t:

Lemma (Bernstein-type inequality)

Let $\{Z_1, \ldots, Z_n\}$ be m.d.s. with $Z_i \leq 1$. Then for any $\delta \in (0,1)$ with probability greater than $1-\delta$ (over Z_1^n) and for any $\lambda \in [0,1]$:

$$\sum_{i=1}^n Z_i \leq \frac{1}{\lambda} \ln \frac{1}{\delta} + (e-2)\lambda V_n.$$

Problem: optimal $\lambda^* = \sqrt{\frac{\ln \frac{1}{\delta}}{(e-2)V_n}}$ depends on V_n and can not be fixed.

Idea: Introduce a grid of λ-s and use the union bound over the grid.

Bernstein-type inequality for martingales

Theorem (Bernstein-type inequality for martingales)

Denote $V_n = \sum_{i=1}^{n} \mathbb{V}\left[Z_i | Z_1^{i-1}\right]$ *for m.d.s.* $\{Z_1, \ldots, Z_n\}$, *such that* $|Z_i| \leq 1$. *For any* $c > 1$ *and* $\delta \in (0,1)$ *with prob. greater than* $1-\delta$ *if*

$$\sqrt{\frac{\ln \frac{T}{\delta}}{(e-2)V_n}} \leq 1$$

then

$$\sum_{i=1}^{n} Z_n \leq (1+c)\sqrt{(e-2)V_n \ln \frac{T}{\delta}},$$

where $T = \left\lceil \frac{1}{\ln c} \ln\left(\sqrt{\frac{(e-2)n}{\ln \frac{1}{\delta}}}\right)\right\rceil$. *Otherwise:*

$$\sum_{i=1}^{n} Z_n \leq 2 \ln \frac{T}{\delta}.$$

Bernstein-type inequality for martingales

Theorem (Bernstein-type inequality for martingales)

Denote $V_n = \sum_{i=1}^{n} \mathbb{V}\left[Z_i | Z_1^{i-1}\right]$ *for m.d.s.* $\{Z_1, \ldots, Z_n\}$, *such that* $|Z_i| \leq 1$. *For any* $c > 1$ *and* $\delta \in (0,1)$ *with prob. greater than* $1-\delta$ *if*

$$\sqrt{\frac{\ln \frac{T}{\delta}}{(e-2)V_n}} \leq 1$$

then

$$\left| \sum_{i=1}^{n} Z_n \right| \leq (1+c)\sqrt{(e-2)V_n \ln \frac{2T}{\delta}},$$

where $T = \left\lceil \frac{1}{\ln c} \ln \left(\sqrt{\frac{(e-2)n}{\ln \frac{2}{\delta}}} \right) \right\rceil$. *Otherwise:*

$$\left| \sum_{i=1}^{n} Z_n \right| \leq 2 \ln \frac{2T}{\delta}.$$

Proof: Bernstein-type inequality for martingales

$$\sum_{i=1}^{n} Z_i \le \frac{1}{\lambda} \ln \frac{1}{\delta} + (e-2)\lambda V_n . \qquad (*)$$

Note that $\lambda^* = \sqrt{\frac{\ln \frac{1}{\delta}}{(e-2)V_n}} \ge \sqrt{\frac{\ln \frac{1}{\delta}}{(e-2)n}}$ since $V_n = \sum_{i=1}^n \mathbb{V}[Z_i | Z_1^{i-1}] = \sum_{i=1}^n \mathbb{E}[Z_i^2 | Z_1^{i-1}] \le n$.
Thus we are interested in interval

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

Fix some $c > 1$ and introduce a grid $\{\lambda_t\}_{i=1}^T$ of λ-s:

$$\lambda_t = c^t \sqrt{\frac{\ln \frac{1}{\delta}}{(e-2)n}}, \quad t = 0, \dots, T.$$

It is enough to take $T = \left\lceil \frac{1}{\ln c} \ln \left(\sqrt{\frac{(e-2)n}{4 \ln \frac{1}{\delta}}} \right) \right\rceil$ to cover the interval and $\lambda_{T-1} \le 1 \le \lambda_T$. We take $\lambda_T = 1$. Use (*) in a union bound over the grid $\{\lambda_t\}_{t=1}^T$ with $\delta = \frac{\delta}{T}$. Then the optimal value changes: $\lambda^*_{\text{new}} = \sqrt{\frac{\ln \frac{T}{\delta}}{(e-2)V_n}}$. If $\lambda^*_{\text{new}} > 1$ take $\lambda = \lambda_T = 1$ and note that $V_n < \frac{\ln \frac{T}{\delta}}{(e-2)}$. Else there is t such that

$$\lambda^*_{\text{new}} \le \lambda_t \le c\lambda^*_{\text{new}}.$$

Use λ_t in (*).

Outline

Part 0: Obtaining PAC-Bayesian inequalities

Part I: Basic martingale definitions

Part II: Concentration inequalities for individual martingales

- Hoeffding-Azuma inequality (Azuma, 1967; Hoeffding, 1963)
- Bernstein inequality (Bernstein, 1946; Freedman, 1975)
- kl-inequality (Seldin et al., 2012)
- Comparison.

Part III: PAC-Bayesian inequalities for martingales (Seldin et al., 2012)

- PAC-Bayes-kl inequality
- PAC-Bayes-Hoeffding-Azuma inequality
- PAC-Bayes-Bernstein inequality

kl-inequality for i.i.d. case (Seeger, 2003; Maurer, 2004)

Consider kl-function:

$$\mathrm{kl}(q\|p) = q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p}.$$

Nice properties:

- Pinsker's inequality: $\mathrm{kl}(q\|p) \geq 2(p-q)^2 \Rightarrow$ upper bound on $\mathrm{kl}(p\|q)$ implies upper bound on $|p-q|$.
- $\mathrm{kl}(q\|p)$ is convex in both arguments and easy to invert numerically.

Let $\{X_1, \ldots, X_n\}$ be i.i.d. r.v. with $\mathbb{E}[X_i] = \mu$, and $X_i \in [0,1]$. We want an upper bound for

$$n \cdot \mathrm{kl}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \middle\| \mu \right)$$

which will provide implicit bound for μ in terms of $\frac{1}{n} \sum_{i=1}^{n} X_i$.

kl-inequality for i.i.d. case (Seeger, 2003; Maurer, 2004)

Denoting $\xi = n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \middle\| \mu\right)$, how can we bound $\mathbb{E}\left[e^{\xi}\right]$?

3-step derivation :

Step 1: Comparison inequality (Maurer, 2004).

For convex f relate $\mathbb{E}[f(X_1, \ldots, X_n)]$ to $\mathbb{E}[f(X'_1, \ldots, X'_n)]$, where $\{X'_1, \ldots, X'_n\}$ are i.i.d. Bernoulli with $\mathbb{E}[X'_i] = \mathbb{E}[X_i] = \mu$.

Step 2: Convexity.

$f : \{x_1, \ldots, x_n\} \in [0,1]^n \to e^{n\,\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \| \mu\right)}$ is convex.

Step 3: Method of types (Seeger, 2003).

Upper bound $\mathbb{E}\left[e^{n\,\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} X'_i \| \mu\right)}\right]$ which is now simpler.

kl-inequality for i.i.d. case (Seeger, 2003; Maurer, 2004)

Step 1: Comparison inequality

Lemma (Maurer, 2004)

Let $\{X_1, \ldots, X_n\}$ be i.i.d. r.v. such that $X_i \in [0,1]$ and $\mathbb{E}[X_i] = \mu$. Denote by $\{X'_1, \ldots, X'_n\}$ i.i.d. Bernoulli r.v. with $\mathbb{E}[X'_i] = \mu$. Then for any convex $f : [0,1]^n \to \mathbb{R}$ we have:

$$\mathbb{E}[f(X_1, \ldots, X_n)] \leq \mathbb{E}[f(X'_1, \ldots, X'_n)].$$

kl-inequality for i.i.d. case (Seeger, 2003; Maurer, 2004)

Step 1: ProofComparison inequality.

We can write any $v \in [0,1]^n$ as a convex combination of $\eta \in \{0,1\}^n$:

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

Indeed for any $z_1, \ldots, z_n$:

$$1 = \Big(z_1 + (1 - z_1) \Big) \Big(z_2 + (1 - z_2) \Big) \cdots \Big(z_n + (1 - z_n) \Big).$$

Convexity of f implies:

$$f(v) \leq \sum_{\eta \in \{0,1\}^n} \left(\prod_{i=1}^{n} \left[(1 - v_i)(1 - \eta_i) + v_i \eta_i \right] \right) f(\eta).$$

Choose $v = X_1^n$, take expectations and use independence of X_i. End up with $\mathbb{E}[f(X_1', \ldots, X_n')]$ in the r.h.s.

kl-inequality for i.i.d. case (Seeger, 2003; Maurer, 2004)

Step 3: Method of types

Lemma (Seeger, 2003; Maurer, 2004)

Let $\{Y_1, \ldots, Y_n\}$ be i.i.d. Bernoulli r.v. with $\mathbb{E}[Y_i] = \mu$. Then:

$$\mathbb{E}\left[e^{n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n Y_i \| \mu\right)}\right] \leq n+1 \qquad \text{(Seeger)}$$

$$\leq 2\sqrt{n}. \qquad \text{(Maurer)}$$

Proof sketch:

$$\begin{aligned}
\mathbb{E}\left[e^{n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n Y_i \| \mu\right)}\right] &= \sum_{k=0}^n \binom{n}{k} \mu^k (1-\mu)^{n-k} e^{n \cdot \mathrm{kl}\left(\frac{k}{n} \| \mu\right)} \\
&= \sum_{k=0}^n \binom{n}{k} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \\
&\leq \sum_{k=0}^n e^{-k \ln \frac{k}{n} - (n-k) \ln \frac{n-k}{n}} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \\
&= \sum_{k=0}^n 1 = n+1,
\end{aligned}$$

where bound on $\binom{n}{k}$ is due to (Cover and Thomas, 1991).

kl-inequality for i.i.d. case (Seeger, 2003; Maurer, 2004)

Theorem (Seeger, 2003; Maurer, 2004)

Let $\{X_1, \ldots, X_n\}$ *be i.i.d. r.v. such that* $X_i \in [0,1]$ *and* $\mathbb{E}[X_i] = \mu$. *Then:*

$$\mathbb{E}\left[e^{n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \| \mu\right)}\right] \leq n+1.$$

Combining with Chernoff's method we get:

Theorem (Seeger, 2003; Maurer, 2004)

For any $\delta \in (0,1)$ *with probability greater than* $1-\delta$*:*

$$\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \middle\| \mu\right) \leq \frac{1}{n} \ln \frac{n+1}{\delta}.$$

kl-inequality for martingales (Seldin et al., 2012)

Now consider $\{Z_1, \ldots, Z_n\}$ such that $Z_i \in [0,1]$ and $\mathbb{E}[Z_i | Z_1^{i-1}] = b$, meaning $\{Z_i - b\}_{i=1}^n$ is a bounded m.d.s.

3-step derivation :

Step 1: Comparison inequality. Problem: $\{Z_1, \ldots, Z_n\}$ are not i.i.d.

Step 2: Convexity. OK!

Step 3: Method of types. OK!

Generalization of Comparison inequality:

Lemma (Seldin et al., 2012)

Let $\{Z_1, \ldots, Z_n\}$ be such that $Z_i \in [0,1]$ and $\mathbb{E}[Z_i | Z_1^{i-1}] = \mu_i$. Denote by $\{Z'_1, \ldots, Z'_n\}$ i.i.d. Bernoulli r.v. with $\mathbb{E}[Z'_i] = \mu_i$. Then for convex f:

$$\mathbb{E}[f(Z_1, \ldots, Z_n)] \leq \mathbb{E}[f(Z'_1, \ldots, Z'_n)].$$

kl-inequality for martingales (Seldin et al., 2012)

Proof sketch :
We can write any $v \in [0,1]^n$ as a convex combination of $\eta \in \{0,1\}^n$:

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

Convexity of f implies:

$$f(v) \leq \sum_{\eta \in \{0,1\}^n} \left(\prod_{i=1}^{n} \left[(1-v_i)(1-\eta_i) + v_i \eta_i \right] \right) f(\eta).$$

kl-inequality for martingales (Seldin et al., 2012)

Proof sketch :

Choose $v = Z_1^n$, denote $W_i(\eta_i) = (1 - Z_i)(1 - \eta_i) + Z_i\eta_i$ and take expectations:

$$\mathbb{E}[f(X_1^n)] \leq \mathbb{E}\left[\sum_{\eta\in\{0,1\}^n}\left(\prod_{i=1}^n W_i(\eta_i)\right)\right] f(\eta)$$

$$= \sum_{\eta\in\{0,1\}^n} \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^n W_i(\eta_i)\middle| Z_1^{n-1}\right]\right] f(\eta)$$

$$= \sum_{\eta\in\{0,1\}^n} \mathbb{E}\left[\prod_{i=1}^{n-1} W_i(\eta_i)\mathbb{E}\left[W_n\middle| Z_1^{n-1}\right]\right] f(\eta)$$

$$= \sum_{\eta\in\{0,1\}^n} \mathbb{E}\left[\prod_{i=1}^{n-1} W_i(\eta_i)\right]\Big((1-\mu_i)(1-\eta_i) + \mu_i\eta_i\Big) f(\eta) = \dots$$

$$= \sum_{\eta\in\{0,1\}^n} \left(\prod_{i=1}^{n}\Big((1-\mu_i)(1-\eta_i) + \mu_i\eta_i\Big)\right) f(\eta) = \mathbb{E}[f(Z_1', \dots, Z_n')].$$

kl-inequality for martingales (Seldin et al., 2012)

Theorem (Seldin et al., 2012)

Let $\{Z_1, \ldots, Z_n\}$ *be such that* $Z_i \in [0,1]$ *and* $\mathbb{E}[Z_i | Z_1^{i-1}] = \mu$. *Then*

$$\mathbb{E}\left[e^{n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n Z_i \| \mu\right)}\right] \leq n+1.$$

Combine with Chernoff's method and get:

Theorem (kl-inequality for martingales)

For any $\delta \in (0,1)$ *with probability greater than* $1-\delta$*:*

$$\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n Z_i \middle\| \mu\right) \leq \frac{1}{n} \ln \frac{n+1}{\delta}.$$

Outline

Part 0: Obtaining PAC-Bayesian inequalities

Part I: Basic martingale definitions

Part II: Concentration inequalities for individual martingales

- Hoeffding-Azuma inequality (Azuma, 1967; Hoeffding, 1963)
- Bernstein inequality (Bernstein, 1946; Freedman, 1975)
- kl-inequality (Seldin et al., 2012)
- Comparison.

Part III: PAC-Bayesian inequalities for martingales (Seldin et al., 2012)

- PAC-Bayes-kl inequality
- PAC-Bayes-Hoeffding-Azuma inequality
- PAC-Bayes-Bernstein inequality

Comparison of inequalities for martingales

Azuma-Hoeffding's :

$$\left|\sum_{i=1}^{n} Z_i\right| \leq \sqrt{\frac{1}{2}\ln\frac{2}{\delta}\sum_{i=1}^{n}(b_i - a_i)^2}$$

Bernstein's :

$$\left|\sum_{i=1}^{n} Z_n\right| \leq (1+c)\sqrt{(e-2)V_n \ln\frac{2T}{\delta}}$$

kl-inequality :

$$\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \middle\| \mu\right) \leq \frac{1}{n}\ln\frac{n+1}{\delta}$$

Comparison: Azuma-Hoeffding vs. Bernstein

$\{Z_1, \ldots, Z_n\}$ is m.d.s. with $|Z_i| \leq 1$. Thus $(b_i - a_i)^2 = 4$.

Azuma-Hoeffding's

$$\left|\sum_{i=1}^{n} Z_i\right| \leq \sqrt{4n \cdot \frac{1}{2} \cdot \ln \frac{2}{\delta}}$$

Bernstein's

$$\left|\sum_{i=1}^{n} Z_i\right| \leq (1+c)\sqrt{(e-2)V_n \ln \frac{2T}{\delta}}$$

Note that:

$$V_n = \sum_{i=1}^{n} \mathbb{V}\left[Z_i | Z_1^{i-1}\right] = \sum_{i=1}^{n} \mathbb{E}\left[Z_i^2 | Z_1^{i-1}\right] \leq n.$$

Bernstein's is at least as tight as Azuma-Hoeffdings's.
If V_n is small Bernstein's is **significantly tighter**.

Comparison: Azuma-Hoeffding vs. kl-inequality

$\{Z_1 - \mu, \ldots, Z_n - \mu\}$ is m.d.s. with $Z_i \in [0,1]$. Thus $(b_i - a_i)^2 = 1$.

Azuma-Hoeffding's :

$$\left|\sum_{i=1}^{n}(Z_i - \mu)\right| \leq \sqrt{n \cdot \frac{1}{2} \cdot \ln \frac{1}{\delta}}$$

kl-inequality :

$$\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \middle\| \mu\right) \leq \frac{1}{n} \ln \frac{n+1}{\delta}$$

Pinsker's inequality $\mathrm{kl}(q\|p) \geq 2(p-q)^2$ imply:

$$\left|\sum_{i=1}^{n}(Z_i - \mu)\right| \leq \sqrt{\frac{n}{2} \ln \frac{n+1}{\delta}}$$

There is a finer one: $p \leq q + \sqrt{2q\mathrm{kl}(q\|p)} + 2\mathrm{kl}(q\|p)$ for $p > q$. Thus:

$$\sum_{i=1}^{n}(\mu - Z_i) \leq \sqrt{2\sum_{i=1}^{n} Z_i \ln \frac{n+1}{\delta} + 2\ln \frac{n+1}{\delta}}$$

kl-inequality is at least as tight as Azuma-Hoeffdings's.
If $\sum_{i=1}^{n} Z_i$ is small kl-inequality is tighter.

Comparison: Bernstein vs. kl-inequality

$\{Z_1 - \mu, \ldots, Z_n - \mu\}$ is m.d.s. with $Z_i \in [0, 1]$.

Bernstein's :

$$\left|\sum_{i=1}^{n}(Z_i - \mu)\right| \leq (1+c)\sqrt{(e-2)V_n \ln \frac{2T}{\delta}}$$

kl-inequality :

$$\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \middle\| \mu\right) \leq \frac{1}{n} \ln \frac{n+1}{\delta}$$

$p \leq q + \sqrt{2q\mathrm{kl}(q\|p)} + 2\mathrm{kl}(q\|p)$ for $p > q$ implies:

$$\sum_{i=1}^{n}(\mu - Z_i) \leq \sqrt{2\sum_{i=1}^{n} Z_i \ln \frac{n+1}{\delta} + 2\ln \frac{n+1}{\delta}}$$

Comparison is not so trivial.
If there is a tight upper bound on V_n Bernstein's inequality can be much tighter than kl-inequality. Otherwise it can be the opposite. . .

Outline

Part 0: Obtaining PAC-Bayesian inequalities

Part I: Basic martingale definitions

Part II: Concentration inequalities for individual martingales

- Hoeffding-Azuma inequality (Azuma, 1967; Hoeffding, 1963)
- Bernstein inequality (Bernstein, 1946; Freedman, 1975)
- kl-inequality (Seldin et al., 2012)
- Comparison.

Part III: PAC-Bayesian inequalities for martingales (Seldin et al., 2012)

- PAC-Bayes-kl inequality
- PAC-Bayes-Hoeffding-Azuma inequality
- PAC-Bayes-Bernstein inequality

PAC-Bayesian inequalities

For any $f_n : \mathcal{H} \times S \to \mathbb{R}$ where $S = \{X_1, \ldots, X_n\} \sim \mathcal{D}$, for any π that does not depend on S, with probability greater than $1 - \delta$ over the draw of S, for all distributions ρ simultaneously:

$$\mathbb{E}_{h\sim\rho}[f_n(h,S)] \leq \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln \mathbb{E}_{h\sim\pi}\left[\mathbb{E}_{S\sim\mathcal{D}}\left[e^{f_n(h,S)}\right]\right]$$

PAC-Bayes-kl :

$$f_n = n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i^h \middle\| \mathbb{E}[Z_1^h]\right)$$

PAC-Bayes-Azuma-Hoeffding :

$$f_n = \lambda \sum_{i=1}^{n} Z_i^h - \frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2$$

PAC-Bayes-Bernstein :

$$f_n = \lambda \sum_{i=1}^{n} Z_i^h - (e-2)\lambda^2 \sum_{i=1}^{n} \mathbb{V}\left[(Z_i^h)^2 | Z_1^h, \ldots, Z_{i-1}^h\right]$$

Outline

Part 0: Obtaining PAC-Bayesian inequalities

Part I: Basic martingale definitions

Part II: Concentration inequalities for individual martingales

- Hoeffding-Azuma inequality (Azuma, 1967; Hoeffding, 1963)
- Bernstein inequality (Bernstein, 1946; Freedman, 1975)
- kl-inequality (Seldin et al., 2012)
- Comparison.

Part III: PAC-Bayesian inequalities for martingales (Seldin et al., 2012)

- PAC-Bayes-kl inequality
- PAC-Bayes-Hoeffding-Azuma inequality
- PAC-Bayes-Bernstein inequality

PAC-Bayes-kl inequality

For $h \in \mathcal{H}$ let $\{Z_1^h, \ldots, Z_n^h\}$ be such that $\mathbb{E}[Z_i^h | Z_1^h, \ldots, Z_{i-1}^h] = \mu^h$ and $Z_i^h \in [0,1]$. Thus $\{Z_1^h - \mu^h, \ldots, Z_n^h - \mu^h\}$ is a m.d.s. for any h. Set:

$$f_n = n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n Z_i^h \middle\| \mu^h\right).$$

Then:

$$\begin{aligned}\mathbb{E}_{h\sim\rho}[f_n] &\leq \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln \mathbb{E}_{h\sim\pi}\left[\mathbb{E}\left[e^{f_n}\right]\right] \\ &\leq \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln(n+1).\end{aligned}$$

Theorem (PAC-Bayes-kl, Seldin et al., 2012)

For any fixed π over $\mathcal{H}$ with probability greater than $1-\delta$ over random draw of $\{Z_1^h, \ldots, Z_n^h\}$ for all $h \in \mathcal{H}$, for all ρ simultaneously:

$$\mathrm{kl}\left(\mathbb{E}_{h\sim\rho}\left[\frac{1}{n}\sum_{i=1}^n Z_i^h\right] \middle\| \mathbb{E}_{h\sim\rho}\left[\mu^h\right]\right) \leq \frac{\mathbb{E}_{h\sim\rho}[f_n]}{n} \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n}.$$

Outline

Part 0: Obtaining PAC-Bayesian inequalities

Part I: Basic martingale definitions

Part II: Concentration inequalities for individual martingales

- Hoeffding-Azuma inequality (Azuma, 1967; Hoeffding, 1963)
- Bernstein inequality (Bernstein, 1946; Freedman, 1975)
- kl-inequality (Seldin et al., 2012)
- Comparison.

Part III: PAC-Bayesian inequalities for martingales (Seldin et al., 2012)

- PAC-Bayes-kl inequality
- PAC-Bayes-Hoeffding-Azuma inequality
- PAC-Bayes-Bernstein inequality

PAC-Bayes-Azuma-Hoeffding inequality

For $h \in \mathcal{H}$ let $\{Z_1^h, \ldots, Z_n^h\}$ be m.d.s. with $Z_i^h \in [a_i, b_i]$. Set for $\lambda \geq 0$:

$$f_n = \lambda \sum_{i=1}^{n} Z_i^h - \frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2.$$

Then:

$$\begin{aligned} \mathbb{E}_{h\sim\rho}[f_n] &\leq \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln \mathbb{E}_{h\sim\pi}\left[\mathbb{E}\left[e^{f_n}\right]\right] \\ &\leq \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln 1. \end{aligned}$$

Theorem (Seldin et al., 2012)

For any fixed π over $\mathcal{H}$ and $\lambda \geq 0$ with probability greater than $1-\delta$ over random draw of $\{Z_1^h, \ldots, Z_n^h\}$ for all $h \in \mathcal{H}$, for all ρ simultaneously:

$$\mathbb{E}_{h\sim\rho}\left[\sum_{i=1}^{n} Z_i^h\right] \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{\lambda} + \frac{\lambda}{8}\sum_{i=1}^{n}(b_i - a_i)^2.$$

PAC-Bayes-Azuma-Hoeffding inequality

Optimizing w.r.t. $\lambda > 0$ using the grid we obtain:

Theorem (PAC-Bayes-Azuma-Hoeffding, Seldin et al., 2012)

For any fixed π over $\mathcal{H}$, $c > 1$, and $\lambda \geq 0$ with probability greater than $1 - \delta$ over random draw of $\{Z_1^h, \ldots, Z_n^h\}$ for all $h \in \mathcal{H}$, for all ρ simultaneously:

$$\mathbb{E}_{h\sim\rho}\left[\sum_{i=1}^{n} Z_i^h\right] \leq \frac{1+c}{2}\sqrt{\frac{1}{2}\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \epsilon(\rho)\right)\sum_{i=1}^{n}(b_i - a_i)^2}$$

where

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

Outline

Part 0: Obtaining PAC-Bayesian inequalities

Part I: Basic martingale definitions

Part II: Concentration inequalities for individual martingales

- Hoeffding-Azuma inequality (Azuma, 1967; Hoeffding, 1963)
- Bernstein inequality (Bernstein, 1946; Freedman, 1975)
- kl-inequality (Seldin et al., 2012)
- Comparison.

Part III: PAC-Bayesian inequalities for martingales (Seldin et al., 2012)

- PAC-Bayes-kl inequality
- PAC-Bayes-Hoeffding-Azuma inequality
- PAC-Bayes-Bernstein inequality

PAC-Bayes-Bernstein inequality

For $h \in \mathcal{H}$ let $\{Z_1^h, \ldots, Z_n^h\}$ be m.d.s. with $|Z_i^h| \leq 1$. Set for $\lambda \in [0, 1]$:

$$f_n = \lambda \sum_{i=1}^{n} Z_i^h - (e-2)\lambda^2 \underbrace{\sum_{i=1}^{n} \mathbb{V}\left[(Z_i^h)^2 | Z_1^h, \ldots, Z_{i-1}^h\right]}_{V_n^h}.$$

Then:

$$\begin{aligned}\mathbb{E}_{h\sim\rho}[f_n] &\leq \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln \mathbb{E}_{h\sim\pi}\left[\mathbb{E}\left[e^{f_n}\right]\right] \\ &\leq \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln 1.\end{aligned}$$

Theorem (Seldin et al., 2012)

For any fixed π over $\mathcal{H}$ and $\lambda \in [0, 1]$ with probability greater than $1 - \delta$ over random draw of $\{Z_1^h, \ldots, Z_n^h\}$ for all $h \in \mathcal{H}$, for all ρ simultaneously:

$$\mathbb{E}_{h\sim\rho}\left[\sum_{i=1}^{n} Z_i^h\right] \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{\lambda} + (e-2)\lambda\, \mathbb{E}_{h\sim\rho}\left[V_n^h\right].$$

PAC-Bayes-Bernstein inequality

Optimizing w.r.t. $\lambda \in [0, 1]$ using the grid we obtain:

Theorem (PAC-Bayes-Bernstein, Seldin et al., 2012)

For any fixed π over $\mathcal{H}$, $c > 1$, and $\lambda \geq 0$ with probability greater than $1 - \delta$ over random draw of $\{Z_1^h, \ldots, Z_n^h\}$ for all $h \in \mathcal{H}$, simultaneously for all ρ that satisfy:

$$\sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln \frac{\lceil \nu \rceil + 1}{\delta}}{(e-2)\mathbb{E}_{h\sim\rho}[V_n^h]}} \leq 1$$

where $\nu = \frac{1}{\ln c} \ln \left(\sqrt{\frac{(e-2)n}{\ln \frac{1}{\delta}}} \right)$ we have

$$\mathbb{E}_{h\sim\rho}\left[\sum_{i=1}^{n} Z_i^h\right] \leq (1+c)\sqrt{(e-2)\mathbb{E}_{h\sim\rho}[V_n^h]\left(\mathrm{KL}(\rho\|\pi) + \ln \frac{\lceil \nu \rceil + 1}{\delta}\right)}$$

and for all other ρ:

$$\mathbb{E}_{h\sim\rho}\left[\sum_{i=1}^{n} Z_i^h\right] \leq 2\left(\mathrm{KL}(\rho\|\pi) + \ln \frac{\lceil \nu \rceil + 1}{\delta}\right).$$

Comparison of all inequalities for martingales

Azuma-Hoeffding's :

$$\sqrt{\frac{1}{2}\ln\frac{1}{\delta}\sum_{i=1}^{n}(b_i-a_i)^2}$$

Bernstein's :

$$(1+c)\sqrt{(e-2)V_n^h\ln\frac{T}{\delta}}$$

kl-inequality :

$$\mathrm{kl}(\cdot\|\cdot)\leq\frac{1}{n}\ln\frac{n+1}{\delta}$$

PAC-Bayes-Azuma-Hoeffding :

$$\frac{1+c}{2}\sqrt{\frac{1}{2}\left(\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}+\epsilon(\rho)\right)\sum_{i=1}^{n}(b_i-a_i)^2}$$

PAC-Bayes-Bernstein :

$$(1+c)\sqrt{(e-2)\mathbb{E}_{h\sim\rho}[V_n^h]\left(\mathrm{KL}(\rho\|\pi)+\ln\frac{\lceil\nu\rceil+1}{\delta}\right)}$$

PAC-Bayes-kl :

$$\mathrm{kl}\Big(\mathbb{E}_{h\sim\rho}[\cdot]\Big\|\mathbb{E}_{h\sim\rho}[\cdot]\Big)\leq\frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{n+1}{\delta}}{n}$$

We pay $\mathrm{KL}(\rho\|\pi)$ to go from bounds on individual martingales to PAC-Bayesian bounds on averages of multiple martingales.

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

Sergei N. Bernstein. *Probability Theory*. Moscow-Leningrad, 4^{th} edition, 1946. In Russian.

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

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1), 1975.

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

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

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

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for

martingales. *IEEE Transactions on Information Theory*, 58, 2012.