

PAC-Bayesian-Empirical-Bernstein Inequality

Ilya Tolstikhin

Russian Academy of Sciences

GRAAL, Laval University

December 2013

Problem setting

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

- ▶ multiple sequences of r.v. $\{X_1^h, \dots, X_n^h\}$ 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. $\{X_1^h, \dots, X_n^h\}$ are **i.i.d.**
- ▶ **Goal:** obtain bound on $\mathbb{E}_{h \sim \rho} [\mathbb{E}[X_1^h]]$ for ρ over \mathcal{H} .

Example: Statistical Learning Theory

- ▶ $\mathcal{H} = \{h: \mathcal{X} \rightarrow \mathcal{Y}\}$ — hypothesis class of predictors;
- ▶ $X_i^h = \ell(y_i, h(x_i))$, where
- ▶ $\{(x_i, y_i)\}_{i=1}^n$ — training sample $\stackrel{iid}{\sim} \mathcal{D}^n$;
- ▶ \mathcal{D} — unknown distribution over input \times label space $\mathcal{X} \times \mathcal{Y}$;
- ▶ $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ — loss function.
- ▶ **Goal:** obtain bound on $L(G_\rho) = \mathbb{E}_{h \sim \rho} [\mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(h(x), y)]]$.

Short summary

Let $\{X_1^h, \dots, X_n^h\}$ for $h \in \mathcal{H}$ be i.i.d. with $\mathbb{E}[X_1^h] = \mu^h$ and $\mathbb{V}[X_1^h] = \mathbb{V}^h$.
Let $|X_i| \leq 1$ for all i and h .

PAC-Bayes-Bernstein inequality:

$$\mathbb{E}_{h \sim \rho} [\mu^h] \leq \mathbb{E}_{h \sim \rho} \left[\frac{1}{n} \sum_{i=1}^n X_i^h \right] + C \sqrt{\frac{\mathbb{E}_{h \sim \rho} [\mathbb{V}^h] \left(\text{KL}(\rho \parallel \pi) + \ln \frac{o(\dots)}{\delta} \right)}{n}}$$

- ▶ May be much tighter than Azuma-Hoeffding's and kl-inequalities.
- ▶ But the **average variance** $\mathbb{E}_{h \sim \rho} [\mathbb{V}^h]$ is unknown.
- ▶ Inequality still holds if we replace $\mathbb{E}_{h \sim \rho} [\mathbb{V}^h]$ with an upper bound $V_n(\rho)$ that holds **simultaneously for all** ρ :

$$\mathbb{E}_{h \sim \rho} [\mathbb{V}^h] \leq V_n(\rho) \leq \frac{1}{4}.$$

PAC-Bayes-Empirical-Bernstein inequality

1. Construct a PAC-Bayes bound for the variance.
2. Use it in a union bound with PAC-Bayes-Bernstein inequality.

Outline

Part 0: **Reminder**: Obtaining PAC-Bayesian inequalities

Part I: **Reminder**: PAC-Bayes-Bernstein inequality (i.i.d. case)

Part II: Upper bounding the variance

- ▶ Using McDiarmid's (bounded differences) inequality
- ▶ Using **entropy method** (Maurer and Pontil, 2009)
- ▶ Third approach by Andreas Maurer (or Gilles Blanchard?)

Part III: PAC-Bayes-Empirical-Bernstein inequality

(Tolstikhin and Seldin, 2013)

- ▶ Derivation of PAC-Bayesian inequality for the variance
- ▶ PAC-Bayes-Empirical-Bernstein inequality
- ▶ Comparison with PAC-Bayes-kl inequality
- ▶ **Application**: Linear regression with absolute loss

Reminder: Obtaining PAC-Bayesian Inequalities

Lemma (Donsker, Varadhan, 1975)

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

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

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

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

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

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

$$f_n: \mathcal{H} \times S \rightarrow \mathbb{R}.$$

Reminder: Obtaining PAC-Bayesian Inequalities

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

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

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} :

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

simultaneously for all ρ .

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

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

Theorem (PAC-Bayes-Bernstein inequality)

For $h \in \mathcal{H}$ let $\{X_1^h, \dots, X_n^h\}$ be i.i.d. with $\mathbb{E}[X_1^h] = \mu^h$ and $\mathbb{V}[X_1^h] = \mathbb{V}^h$ and $X_i^h \in [0, 1]$. For any fixed π over \mathcal{H} , $\delta \in (0, 1)$, and $c > 1$, with probability greater than $1 - \delta$ (over the random draw of $\{X_1^n\}$ for all $h \in \mathcal{H}$):

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

simultaneously for all ρ over \mathcal{H} such that

$$\sqrt{\frac{\text{KL}(\rho \|\pi) + \ln(\nu/\delta)}{(e-2)\mathbb{E}_{h \sim \rho}[\mathbb{V}^h]}} \leq \sqrt{n}$$

where $\nu = \left\lceil \frac{1}{\ln c} \ln \left(\sqrt{\frac{(e-2)n}{4 \ln(1/\delta)}} \right) \right\rceil + 1$ and for all other ρ we have:

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

PAC-Bayes-Bernstein Inequality: Proof Sketch

1. **Bernstein's upper bound on moment generating function.**

For $\lambda \in [0, 1]$:

$$\mathbb{E} \left[\exp \left(\lambda \sum_{i=1}^n X_i - (e-2)\lambda^2 n \mathbb{V}[X_1] \right) \right] \leq 1$$

2. **Combine with Donsker-Varadhan's inequality.**

For fixed π over \mathcal{H} with prob. greater than $1 - \delta$ for any $\lambda \in [0, 1]$:

$$\mathbb{E}_{h \sim \rho}[\mu^h] \leq \mathbb{E}_{h \sim \rho} \left[\frac{1}{n} \sum_{i=1}^n X_i^h \right] + \frac{\text{KL}(\rho \parallel \pi) + \ln \frac{1}{\delta}}{n\lambda} + (e-2)\lambda \mathbb{E}_{h \sim \rho}[\mathbb{V}^h]$$

simultaneously for all ρ .

3. **Optimize w.r.t. $\lambda \in [0, 1]$ using union bound over the grid $\{\lambda_t\}$.**

We can replace $\mathbb{E}_{h \sim \rho}[\mathbb{V}^h]$ with any upper bound $\bar{\mathbb{V}}_n(\rho)$ as long as $\mathbb{E}_{h \sim \rho}[\mathbb{V}^h] \leq \bar{\mathbb{V}}_n(\rho) \leq \frac{1}{4}$ simultaneously for all ρ !

Outline

Part 0: **Reminder**: Obtaining PAC-Bayesian inequalities

Part I: **Reminder**: PAC-Bayes-Bernstein inequality (i.i.d. case)

Part II: **Upper bounding the variance**

- ▶ Using McDiarmid's (bounded differences) inequality
- ▶ Using **entropy method** (Maurer and Pontil, 2009)
- ▶ Third approach by Andreas Maurer (or Gilles Blanchard?)

Part III: **PAC-Bayes-Empirical-Bernstein inequality**

(Tolstikhin and Seldin, 2013)

- ▶ Derivation of PAC-Bayesian inequality for the variance
- ▶ PAC-Bayes-Empirical-Bernstein inequality
- ▶ Comparison with PAC-Bayes-kl inequality
- ▶ **Application**: Linear regression with absolute loss

Upper bounding the variance

For i.i.d. sample $\{X_1, \dots, X_n\}$ an unbiased **empirical variance** is:

$$\mathbb{V}_n = \mathbb{V}_n(X_1, \dots, X_n) = \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \frac{1}{n} \sum_{j=1}^n X_j \right)^2$$

meaning

$$\mathbb{E}[\mathbb{V}_n(X_1, \dots, X_n)] = \mathbb{V}[X_1]$$

Can we upper bound $\mathbb{V}[X_1]$ using \mathbb{V}_n ?

Note: we know a lot of ways to deal with $f(X_1, \dots, X_m) = \sum_{i=1}^n X_i$:
Chernoff's method \rightarrow Hoeffding's and Bernstein's inequalities

But: what about other functions f ? For example $f = \mathbb{V}_n$?

$$\mathbb{P}\{\mathbb{V}[X_1] - \mathbb{V}_n \geq t\} \leq \frac{\mathbb{E}[e^{\lambda(\mathbb{V}[X_1] - \mathbb{V}_n)}]}{e^{\lambda t}} \leq \dots?$$

Concentration inequalities: references

(Boucheron et al., 2013) : overview of **many** approaches for upper bounding m.g.f.

Basic inequalities

- ▶ Markov's and Chebyshev's inequalities
- ▶ Chernoff's bounding method

Inequalities for sums of independent r.v. (or martingales)

- ▶ Hoeffding's, Bennett's, and Bernstein's inequalities

Inequalities for general functions of independent r.v.

- ▶ **Martingale method** and McDiarmid's inequality
- ▶ M. Talagrand's **inductive method** (1995, 1996)
- ▶ M. Ledoux's **entropy method** (1996)
developed by S. Boucheron, O. Bousquet, G. Lugosi, P. Massart

Outline

Part 0: **Reminder**: Obtaining PAC-Bayesian inequalities

Part I: **Reminder**: PAC-Bayes-Bernstein inequality (i.i.d. case)

Part II: Upper bounding the variance

- ▶ Using McDiarmid's (bounded differences) inequality
- ▶ Using **entropy method** (Maurer and Pontil, 2009)
- ▶ Third approach by Andreas Maurer (or Gilles Blanchard?)

Part III: PAC-Bayes-Empirical-Bernstein inequality

(Tolstikhin and Seldin, 2013)

- ▶ Derivation of PAC-Bayesian inequality for the variance
- ▶ PAC-Bayes-Empirical-Bernstein inequality
- ▶ Comparison with PAC-Bayes-kl inequality
- ▶ **Application**: Linear regression with absolute loss

McDiarmid's inequality (McDiarmid, 1989)

A function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ has a **bounded differences property** if for some nonnegative c_1, \dots, c_n :

$$\sup_{\substack{x_1, \dots, x_n, \\ x'_i \in \mathbb{R}}} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq c_i, \quad 1 \leq i \leq n.$$

Example: $f(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ if $x_i \in [0, 1]$.

Bounded difference property is enough to bound the m.g.f.!

Theorem

Let $\{X_1, \dots, X_n\}$ be i.i.d. and f have a bounded differences with c_1, \dots, c_n . Denote $\xi = f(X_1, \dots, X_n)$. Then for $\lambda \geq 0$:

$$\mathbb{E}[e^{\lambda(\xi - \mathbb{E}[\xi])}] \leq \exp\left(\frac{\lambda^2}{8} \sum_{i=1}^n c_i^2\right)$$

Note that the same holds true for $\xi = -f(X_1, \dots, X_n)$.

McDiarmid's inequality (McDiarmid, 1989)

Combine with Chernoff's bounding method and obtain:

Theorem (McDiarmid's inequality)

Let $\{X_1, \dots, X_n\}$ be i.i.d. and f have a bounded differences with c_1, \dots, c_n . Denote $\xi = f(X_1, \dots, X_n)$. Then

$$\mathbb{P}\{\xi - \mathbb{E}[\xi] \geq t\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

and

$$\mathbb{P}\{\mathbb{E}[\xi] - \xi \geq t\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

- ▶ If we use McDiarmid's inequality with $f(X_1, \dots, X_n) = \sum_{i=1}^n X_i$ for $X_i \in [0, 1]$ then we **exactly recover** Hoeffding's inequality!
- ▶ Can we use McDiarmid's inequality for $f = \mathbb{V}_n$?

Bounding the variance with McDiarmid's inequality

Question: Does \mathbb{V}_n have bounded differences property for $X_i \in [0, 1]$?

$$\left| \mathbb{V}_n(X_1, \dots, X_n) - \mathbb{V}_n(X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n) \right| = \dots?$$

It can be shown that:

$$\begin{aligned} \mathbb{V}_n(X_1, \dots, X_n) &= \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \frac{1}{n} \sum_{j=1}^n X_j \right)^2 \\ &= \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} (X_i - X_j)^2 \end{aligned}$$

Thus

$$\dots = \left| \frac{1}{n(n-1)} \sum_{j \neq i} \left\{ (X_j - X_i)^2 - (X_j - X'_i)^2 \right\} \right| \leq \frac{1}{n}.$$

Bounding the variance with McDiarmid's inequality

Using McDiarmid's inequality with $c_1 = \dots = c_n = \frac{1}{n}$ and denoting $\xi = V_n(X_1, \dots, X_n)$ we get:

$$\mathbb{E} \left[e^{\lambda(\mathbb{E}[\xi] - \xi)} \right] = \mathbb{E} \left[e^{\lambda(\mathbb{V}[X_1] - \xi)} \right] \leq \exp \left(\frac{\lambda^2}{8n} \right)$$

Combining with Chernoff's bounding method we get:

Theorem (McDiarmid-style upper bound on the variance)

Let $\{X_1, \dots, X_n\}$ be i.i.d. with $X_i \in [0, 1]$. For any $\delta \in (0, 1)$ with probability greater than $1 - \delta$ (over the random draw of X_1, \dots, X_n)

$$\mathbb{V}[X_1] \leq \mathbb{V}_n(X_1, \dots, X_n) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$$

and also

$$\mathbb{V}_n(X_1, \dots, X_n) \leq \mathbb{V}[X_1] + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$$

Bounding the variance with McDiarmid's inequality

- ▶ Choose $f_n = \lambda(\mathbb{V}[X_1] - \mathbb{V}_n(X_1, \dots, X_n))$
- ▶ Combine with Donsker-Varadhan's inequality:

For fixed π over \mathcal{H} with prob. greater than $1 - \delta$ for any $\lambda \geq 0$:

$$\mathbb{E}_{h \sim \rho}[\mathbb{V}^h] \leq \mathbb{E}_{h \sim \rho}[\mathbb{V}_n(X_1^h, \dots, X_n^h)] + \frac{\text{KL}(\rho \parallel \pi) + \ln \frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}$$

simultaneously for all ρ .

Let us “cheat” and optimise directly w.r.t. λ .

We end up with **McAllester-style bound**:

$$\mathbb{E}_{h \sim \rho}[\mathbb{V}^h] \leq \mathbb{E}_{h \sim \rho}[\mathbb{V}_n(X_1^h, \dots, X_n^h)] + \sqrt{\frac{\text{KL}(\rho \parallel \pi) + \ln \frac{1}{\delta}}{2n}}$$

Question: Can we do better?

Outline

Part 0: **Reminder**: Obtaining PAC-Bayesian inequalities

Part I: **Reminder**: PAC-Bayes-Bernstein inequality (i.i.d. case)

Part II: Upper bounding the variance

- ▶ Using McDiarmid's (bounded differences) inequality
- ▶ Using **entropy method** (Maurer and Pontil, 2009)
- ▶ Third approach by Andreas Maurer (or Gilles Blanchard?)

Part III: PAC-Bayes-Empirical-Bernstein inequality

(Tolstikhin and Seldin, 2013)

- ▶ Derivation of PAC-Bayesian inequality for the variance
- ▶ PAC-Bayes-Empirical-Bernstein inequality
- ▶ Comparison with PAC-Bayes-kl inequality
- ▶ **Application**: Linear regression with absolute loss

Entropy method (Maurer, 2006)

Let $\mathbf{X} = \{X_1, \dots, X_n\}$ be independent r.v. taking values in \mathcal{X} .

For $1 \leq k \leq n$ and $x \in \mathcal{X}$ denote $\mathbf{X}_{k,x} = \{X_1, \dots, X_{k-1}, x, X_{k+1}, X_n\}$.

A function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is a **self-bounding function** if for some $a \geq 1$:

$$\forall k: \quad f(\mathbf{X}) - \inf_{x \in \mathbb{R}} f(\mathbf{X}_{k,x}) \leq 1$$

$$\sum_{k=1}^n \left(f(\mathbf{X}) - \inf_{x \in \mathcal{X}} f(\mathbf{X}_{k,x}) \right)^2 \leq a f(\mathbf{X})$$

Example: $f(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ if $x_i \in [0, 1]$.

For self-bounding functions m.g.f. can be also bounded:

Theorem (Maurer's Inequality)

Let f be a self-bounding function with constant a . Denote $\xi = f(X_1, \dots, X_n)$. Then for any $\lambda \geq 0$:

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

Empirical Bernstein inequality

(Maurer and Pontil, 2009; Audibert et al., 2009)

Question: is \mathbb{V}_n a self-bounding function for $X_i \in [0, 1]$?

Lemma ((Maurer and Pontil, 2009))

Function $f(X_1, \dots, X_n) = n \cdot \mathbb{V}_n(X_1, \dots, X_n)$ is a self-bounding function with $a = \frac{n}{n-1}$.

Proof sketch:

$$\begin{aligned}n \cdot \mathbb{V}_n(\mathbf{X}) - n \cdot \mathbb{V}_n(\mathbf{X}_{k,x}) &= \frac{1}{n-1} \sum_j ((X_k - X_j)^2 - (x - X_j)^2) \\ &\leq \frac{1}{n-1} \sum_j (X_k - X_j)^2 \leq 1.\end{aligned}$$

Second property is more complicated. Look in (Maurer and Pontil, 2009).

Empirical Bernstein inequality

(Maurer and Pontil, 2009; Audibert et al., 2009)

Applying Maurer's Inequality for $f(X_1, \dots, X_n) = n \cdot \mathbb{V}_n(X_1, \dots, X_n)$ and denoting $\xi = \mathbb{V}_n(X_1, \dots, X_n)$ we get:

Theorem

For any $\lambda \geq 0$:

$$\mathbb{E}[e^{\lambda(n\mathbb{E}[\xi] - n\xi)}] \leq \exp\left(\frac{\lambda^2}{2} \frac{n^2}{n-1} \mathbb{E}[\xi]\right)$$

Combine with Chernoff's bounding method and obtain:

Theorem

Let $\{X_1, \dots, X_n\}$ be i.i.d. with $X_i \in [0, 1]$. For any $\delta \in (0, 1)$ with probability greater than $1 - \delta$ (over the random draw of X_1, \dots, X_n)

$$\mathbb{V}[X_1] \leq \mathbb{V}_n(X_1, \dots, X_n) + \sqrt{\frac{2\mathbb{V}[X_1] \ln \frac{1}{\delta}}{n-1}}$$

Empirical Bernstein inequality

Theorem (Entropy-method upper bound on the variance)

Let $\{X_1, \dots, X_n\}$ be i.i.d. with $X_i \in [0, 1]$. For any $\delta \in (0, 1)$ with probability greater than $1 - \delta$ (over the random draw of X_1, \dots, X_n)

$$\mathbb{V}[X_1] \leq \mathbb{V}_n(X_1, \dots, X_n) + \sqrt{\frac{2\mathbb{V}[X_1] \ln \frac{1}{\delta}}{n-1}}$$

We can “solve” this inequality w.r.t. $\mathbb{V}[X_1]$ using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$:

$$\left(\sqrt{\mathbb{V}[X_1]} - \frac{1}{2} \sqrt{\frac{2 \ln \frac{1}{\delta}}{n-1}} \right)^2 \leq \mathbb{V}_n(X_1, \dots, X_n) + \frac{1}{4} \left(\frac{2 \ln \frac{1}{\delta}}{n-1} \right)$$

and obtain a better bound:

$$\mathbb{V}[X_1] \leq \mathbb{V}_n(X_1, \dots, X_n) + 2\sqrt{\frac{\mathbb{V}_n(X_1, \dots, X_n) \ln \frac{1}{\delta}}{n-1}} + \frac{2 \ln \frac{1}{\delta}}{n-1}$$

Outline

Part 0: **Reminder**: Obtaining PAC-Bayesian inequalities

Part I: **Reminder**: PAC-Bayes-Bernstein inequality (i.i.d. case)

Part II: Upper bounding the variance

- ▶ Using McDiarmid's (bounded differences) inequality
- ▶ Using **entropy method** (Maurer and Pontil, 2009)
- ▶ **Third approach by Andreas Maurer (or Gilles Blanchard?)**

Part III: PAC-Bayes-Empirical-Bernstein inequality

(Tolstikhin and Seldin, 2013)

- ▶ Derivation of PAC-Bayesian inequality for the variance
- ▶ PAC-Bayes-Empirical-Bernstein inequality
- ▶ Comparison with PAC-Bayes-kl inequality
- ▶ **Application**: Linear regression with absolute loss

The third way: due to Maurer (or Blanchard?)

Note: if X_1 and X_2 are i.i.d. then

$$\begin{aligned}\mathbb{E}[(X_1 - X_2)^2] &= \mathbb{E}[(X_1 - \mathbb{E}[X_1] + \mathbb{E}[X_2] - X_2)^2] \\ &= \mathbb{E}[(X_1 - \mathbb{E}[X_1])^2] + \mathbb{E}[(\mathbb{E}[X_2] - X_2)^2] \\ &= 2\mathbb{V}[X_1].\end{aligned}$$

If we have $\{X_1, \dots, X_n\}$ i.i.d., where $n = 2k$, then

$$\hat{V}_n = \frac{1}{k} \sum_{i=1}^k \frac{1}{2} (X_{2i} - X_{2i-1})^2$$

is an **unbiased sample estimate** of the variance: $\mathbb{E}[\hat{V}_n] = \mathbb{V}[X_1]$.

Important: \hat{V}_n is a sum of i.i.d. random variables!

Theorem (?Unpublished?)

$$\text{kl} \left(\hat{V}_n \left\| \mathbb{V}[X_1] \right. \right) \leq \frac{2}{n} \ln \frac{2\sqrt{\frac{n}{2}}}{\delta} = \frac{2}{n} \ln \frac{\sqrt{2n}}{\delta}$$

Outline

Part 0: **Reminder**: Obtaining PAC-Bayesian inequalities

Part I: **Reminder**: PAC-Bayes-Bernstein inequality (i.i.d. case)

Part II: Upper bounding the variance

- ▶ Using McDiarmid's (bounded differences) inequality
- ▶ Using **entropy method** (Maurer and Pontil, 2009)
- ▶ Third approach by Andreas Maurer (or Gilles Blanchard?)

Part III: PAC-Bayes-Empirical-Bernstein inequality

(Tolstikhin and Seldin, 2013)

- ▶ **Derivation of PAC-Bayesian inequality for the variance**
- ▶ PAC-Bayes-Empirical-Bernstein inequality
- ▶ Comparison with PAC-Bayes-kl inequality
- ▶ **Application**: Linear regression with absolute loss

PAC-Bayes inequality for the variance: Derivation

For any fixed $h \in \mathcal{H}$ let $\mathbf{X}^h = \{X_1^h, \dots, X_n^h\}$ be i.i.d. with $X_i^h \in [0, 1]$.

- ▶ Let us choose for $\lambda \geq 0$:

$$f_n = \lambda (n \cdot \mathbb{V}[X_1^h] - n \cdot \mathbb{V}_n(\mathbf{X}^h)) - \frac{\lambda^2}{2} \frac{n^2}{n-1} \mathbb{V}[X_1^h]$$

- ▶ Combine it with the Donsker-Varadhan's inequality:

$$\left(1 - \frac{\lambda n}{2(n-1)}\right) \mathbb{E}_{h \sim \rho} [\mathbb{V}[X_1^h]] \leq \mathbb{E}_{h \sim \rho} [\mathbb{V}_n(\mathbf{X}^h)] + \frac{\text{KL}(\rho \parallel \pi) + \ln \frac{1}{\delta}}{\lambda n}$$

- ▶ For $0 \leq \lambda \leq \frac{2(n-1)}{n}$ divide both sides by $1 - \frac{\lambda n}{2(n-1)}$:

$$\mathbb{E}_{h \sim \rho} [\mathbb{V}[X_1^h]] \leq \frac{\mathbb{E}_{h \sim \rho} [\mathbb{V}_n(\mathbf{X}^h)]}{\left(1 - \frac{\lambda n}{2(n-1)}\right)} + \frac{\text{KL}(\rho \parallel \pi) + \ln \frac{1}{\delta}}{\lambda n \left(1 - \frac{\lambda n}{2(n-1)}\right)}$$

- ▶ Optimize w.r.t. λ

PAC-Bayes inequality for the variance: Derivation

For any fixed $h \in \mathcal{H}$ let $\mathbf{X}^h = \{X_1^h, \dots, X_n^h\}$ be i.i.d. with $X_i^h \in [0, 1]$.

For $0 \leq \lambda \leq \frac{2(n-1)}{n}$:

$$\mathbb{E}_{h \sim \rho} [\mathbb{V}[X_1^h]] \leq \frac{\mathbb{E}_{h \sim \rho} [\mathbb{V}_n(\mathbf{X}^h)]}{\left(1 - \frac{\lambda n}{2(n-1)}\right)} + \frac{\text{KL}(\rho \parallel \pi) + \ln \frac{1}{\delta}}{\lambda n \left(1 - \frac{\lambda n}{2(n-1)}\right)} \quad (1)$$

Notations:

$$t = \frac{\lambda n}{2(n-1)}, \quad a = \mathbb{E}_{h \sim \rho} [\mathbb{V}_n(\mathbf{X})], \quad b = \frac{\text{KL}(\rho \parallel \pi) + \ln \frac{1}{\delta}}{2(n-1)}$$

Then (1) can be written as:

$$\mathbb{E}_{h \sim \rho} [\mathbb{V}[X_1^h]] \leq F(t) = \frac{a}{1-t} + \frac{b}{t(1-t)},$$

where $a, b \geq 0$ and $0 < t \leq 1$. $F(t)$ is **convex** on $(0, 1]$ and achieves its minimum at the **positive solution** of

$$at^2 + 2bt - b = 0$$

PAC-Bayes inequality for the variance: Derivation

For optimal point t we have:

$$\arg \min_{t \in (0,1]} F(t) = t^* = \frac{\sqrt{b^2 + ab} - b}{a} = \frac{1}{\sqrt{a/b + 1} + 1} \leq \frac{1}{2}$$

Also by definitions of a and b :

$$t^* = \left(\sqrt{\frac{2(n-1)\mathbb{E}_{h \sim \rho}[\mathbb{V}_n(\mathbf{X})]}{\text{KL}(\rho \parallel \pi) + \ln 1/\delta} + 1 + 1} \right)^{-1}$$

Noting that $\mathbb{V}_n(\mathbf{X}) \leq \frac{1}{n}$ and $\text{KL}(\rho \parallel \pi) \geq 0$ we have:

$$t^* \geq \left(\sqrt{\frac{n-1}{\ln 1/\delta} + 1 + 1} \right)^{-1}$$

We are seeking in the range

$$\left[\left(\sqrt{\frac{n-1}{\ln 1/\delta} + 1 + 1} \right)^{-1}, \frac{1}{2} \right] \triangleq \left[\Delta(n, \delta), \frac{1}{2} \right]$$

PAC-Bayes inequality for the variance: Derivation

We introduce geometrically spaced grid of t . Fix $c > 1$ and let

$$t_i = c^i \Delta(n, \delta), \quad i = 0, \dots, M - 1$$

In order to cover the interval it suffices to take

$$M = \left\lceil \frac{1}{\ln c} \ln \frac{1}{2\Delta(n, \delta)} \right\rceil$$

For any t^* (which depends on ρ) we can find $i^* \in \{0, \dots, M - 1\}$ such that:

$$t_{i^*} \leq t^* \leq ct_{i^*} \tag{1}$$

Substituting $t = t_{i^*}$ into $F(t)$ and using (1) we get:

$$F(t_{i^*}) \leq a + (1 + c)\sqrt{ab} + 4cb. \tag{2}$$

Finally combining (2) in the union bound over $i^* \in \{0, \dots, M - 1\}$ completes the derivation.

PAC-Bayes-Empirical-Bernstein Inequality

(Tolstikhin and Seldin, 2013)

Theorem (PAC-Bayesian inequality for the variance)

For any fixed $h \in \mathcal{H}$ let $\mathbf{X}^h = \{X_1^h, \dots, X_n^h\}$ be i.i.d. with $X_i^h \in [0, 1]$.

For any fixed π over \mathcal{H} , $c > 1$, and $\delta \in (0, 1)$ with probability greater than $1 - \delta$ for all ρ over \mathcal{H} simultaneously:

$$\mathbb{E}_{h \sim \rho} [\mathbb{V}[X_1^h]] \leq \mathbb{E}_{h \sim \rho} [\mathbb{V}_n(\mathbf{X})] + (1 + c) \sqrt{\frac{\mathbb{E}_{h \sim \rho} [\mathbb{V}_n(\mathbf{X})] (\text{KL}(\rho \parallel \pi) + \ln \frac{T}{\delta})}{2(n-1)}} \\ + \frac{2c (\text{KL}(\rho \parallel \pi) + \ln \frac{T}{\delta})}{n-1}$$

where

$$T = \left\lceil \frac{1}{\ln c} \ln \left(\frac{1}{2} \sqrt{\frac{n-1}{\ln(1/\delta)} + 1} + \frac{1}{2} \right) \right\rceil$$

PAC-Bayes-Empirical-Bernstein inequality: use upper bound $\bar{\mathbb{V}}_n(\rho)$ in the union bound with PAC-Bayes-Bernstein inequality.

Outline

Part 0: **Reminder**: Obtaining PAC-Bayesian inequalities

Part I: **Reminder**: PAC-Bayes-Bernstein inequality (i.i.d. case)

Part II: Upper bounding the variance

- ▶ Using McDiarmid's (bounded differences) inequality
- ▶ Using **entropy method** (Maurer and Pontil, 2009)
- ▶ Third approach by Andreas Maurer (or Gilles Blanchard?)

Part III: PAC-Bayes-Empirical-Bernstein inequality

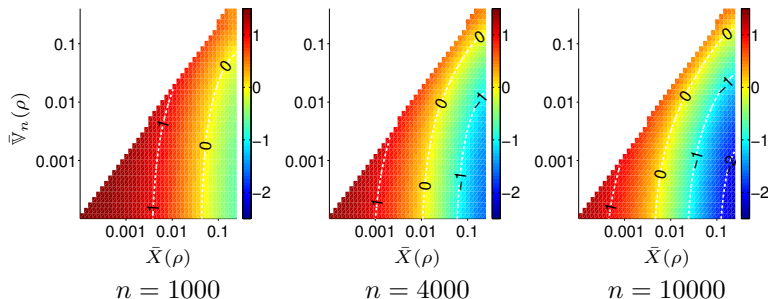
(Tolstikhin and Seldin, 2013)

- ▶ Derivation of PAC-Bayesian inequality for the variance
- ▶ PAC-Bayes-Empirical-Bernstein inequality
- ▶ **Comparison with PAC-Bayes-kl inequality**
- ▶ **Application**: Linear regression with absolute loss

Comparison with PAC-Bayes-kl inequality

- ▶ Fix $\text{KL}(\rho \parallel \pi) = 18$, $c_1 = c_2 = 1.15$, and $\delta = 0.05$
- ▶ Denote $\bar{X}(\rho) = \mathbb{E}_{h \sim \rho} \left[\frac{1}{n} \sum_{i=1}^n X_i^h \right]$
- ▶ Compare the complexity terms of the bounds:

$$\log_2 \left(\frac{\text{PAC-Bayes-Empirical-Bernstein} - \bar{X}(\rho)}{\text{PAC-Bayes-kl} - \bar{X}(\rho)} \right)$$



Note: PB-EB is 4 times tighter on the right plot!

Outline

Part 0: **Reminder**: Obtaining PAC-Bayesian inequalities

Part I: **Reminder**: PAC-Bayes-Bernstein inequality (i.i.d. case)

Part II: Upper bounding the variance

- ▶ Using McDiarmid's (bounded differences) inequality
- ▶ Using **entropy method** (Maurer and Pontil, 2009)
- ▶ Third approach by Andreas Maurer (or Gilles Blanchard?)

Part III: PAC-Bayes-Empirical-Bernstein inequality

(Tolstikhin and Seldin, 2013)

- ▶ Derivation of PAC-Bayesian inequality for the variance
- ▶ PAC-Bayes-Empirical-Bernstein inequality
- ▶ Comparison with PAC-Bayes-kl inequality
- ▶ **Application**: Linear regression with absolute loss

Application: linear regression with absolute loss

- ▶ $\mathcal{X} = \{x \in \mathbb{R}^d: \|x\|_2 \leq 1\}$
- ▶ $\mathcal{Y} = [-0.5, 0.5]$
- ▶ $\mathcal{H} = \{h_w(x) = \langle x, w \rangle: w \in \mathbb{R}^d, \|w\|_2 \leq 0.5\}$
- ▶ $\ell(y', y'') = |y' - y''|$
- ▶ Define $X_i^h = \ell(h(x_i), y_i)$ where $h \in \mathcal{H}$
- ▶ Thus $X_i^h = \ell(h(x_i), y_i) \in [0, 1]!$

For sample $\{(x_i, y_i)\}_{i=1}^n$ solve ERM problem:

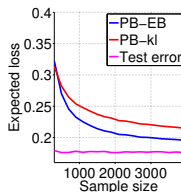
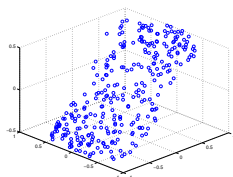
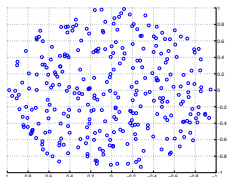
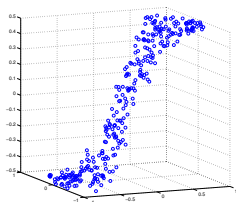
$$h_{\hat{w}} = \arg \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(h_w(x_i), y_i)$$

- ▶ Let π be uniform over \mathcal{H}
- ▶ Let $\rho_{\hat{w}}$ be uniform over $\{h_w \in \mathcal{H}: \|w - \hat{w}\|_2 \leq \epsilon\}$

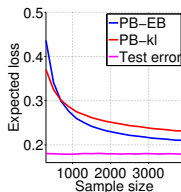
Application: linear regression with absolute loss

Synthetic datasets:

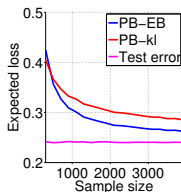
- ▶ Draw x_1, \dots, x_n i.i.d. uniformly over $\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$
- ▶ Set $y_i = \sigma(50 \cdot \langle w_0, x_i \rangle) + \eta_i$, where $\sigma(x) = 1/(1 + e^{-x}) - 0.5$ and η_i is a **truncated** independent random noise



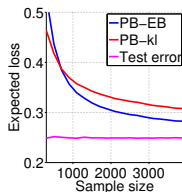
$d=2, \|w_0\|_0=2$



$d=5, \|w_0\|_0=2$



$d=3, \|w_0\|_0=3$



$d=6, \|w_0\|_0=3$

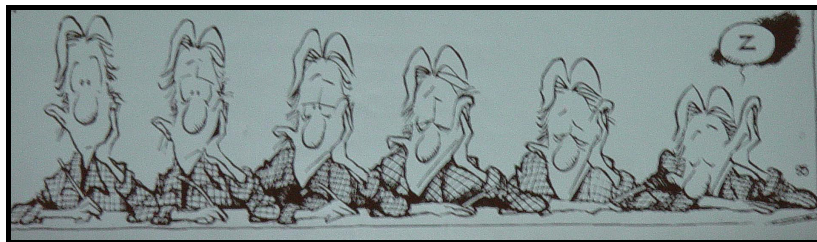
Application: linear regression with absolute loss

UCI data-sets:

Dataset	n	d	Test	PB-kl bound	PB-EB bound
winequality	6497	11	0.106 ± 0.0022	0.175 ± 0.0006	0.162 ± 0.0006
parkinsons	5875	16	0.188 ± 0.0055	0.266 ± 0.0013	0.250 ± 0.0012
concrete	1030	8	0.111 ± 0.0038	0.242 ± 0.0010	0.264 ± 0.0011

Sketches of experiments:

- ▶ $\text{KL}(\rho_{\hat{w}} \parallel \pi) = d \ln \frac{2}{\epsilon}$
- ▶ Let n_ϵ be number of points such that $|y_i - \langle \hat{w}, x_i \rangle| < \epsilon$
- ▶ $\mathbb{E}_{h \sim \rho} \left[\frac{1}{n} \sum_{i=1}^n \ell(h_{\hat{w}}(x_i), y_i) \right] \leq \frac{1}{n} \sum_{i=1}^n \ell(h_{\hat{w}}(x_i), y_i) + \epsilon \frac{n_\epsilon}{n}$
- ▶ And expression for $\mathbb{E}_{h \sim \rho} [\mathbb{V}_n(\dots)]$ won't fit even on 2 slides :(



Jean Yves Audibert, Rémi Munos, and Csaba Szepesvári.

Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 2009.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart.

Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press, 2013.

Andreas Maurer. Concentration inequalities for functions of independent variables. *Random Structures and Algorithms*, 29(2), 2006.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2009.

C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188, Cambridge, 1989. Cambridge University Press.

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.

Ilya Tolstikhin and Yevgeny Seldin. PAC-Bayes-Empirical-Bernstein inequality. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.