PAC-Bayes Un-Expected Bernstein Inequality

Zakaria Mhammedi Peter D. Grünwald Benjamin Guedj

Australian National University CWI Centrum Wiskunde & Informatica DATA 61 Inria inventors for the digital world UCL

Contribution

We derive a new **second-order** (PAC-Bayesian) generalization bound. The key tool behind the bound is **a new empirical Bernstein concentration inequality**.

Abstract

Standard PAC-Bayesian bounds contain a $\sqrt{L_n \cdot \mathrm{KL}/n}$ term which dominates unless L_n, the empirical error, vanishes. We managed to **replace** L_n by a term V_n which **vanishes** whenever the employed learning algorithm is **sufficiently stable**. The **key novelties** are:

Informed Priors: We split the data in two and learn a prior from each. The bound is small when the priors are close (**i.e. stable algorithm**).

Online Estimators: Our bound has a second order term which is in the form of a sum of (squared) errors incurred by online estimators.

Connection with Excess Risks: We connect our new PAC-Bayesian bound with **excess risks** under a Bernstein condition.

New Concentration Inequality: The key tool we use is a new **concentration inequality** which is like Bernstein's but with X^2 outside the $\mathbb{E}$.

Setting and Notation

We consider $Z_1, \dots, Z_n$ i.i.d. random variable in $\mathcal{Z}$, with $Z_1 \sim \mathbf{D}$. Let $\mathcal{H}$ be a hypothesis set and $\ell : \mathcal{H} \times \mathcal{Z} \to [0, b]$, $b > 0$, be a **loss** such that $\ell_h(Z) := \ell(h, Z)$. For $h \in \mathcal{H}$, we denote its **risk** by

$$L(h) := \mathbb{E}_{Z\sim\mathbf{D}}[\ell_h(Z)],$$

and its **empirical risk** by

$$L_n(h) := \frac{1}{n}\sum_{i=1}^{n} \ell_h(Z_i).$$

For a distribution P on $\mathcal{H}$, we write

$$L(P) := \mathbb{E}_{h\sim P}[L(h)] \ \ and \ \ L_n(P) := \mathbb{E}_{h\sim P}[L_n(h)].$$

For $m \in [n]$ and random variables $Z_1, \dots, Z_n$, we denote $Z_{\leq m} := (Z_1, \dots, Z_m)$ and $Z_{<m} := Z_{\leq m-1}$, with $Z_{\leq 0} = \emptyset$. Similarly, $Z_{\geq m} := (Z_m, \dots, Z_n)$ and $Z_{>m} := Z_{\geq m+1}$, with $Z_{\geq n+1} = \emptyset$.

A **learning algorithm** is a map $P : \bigcup_{i=1}^n \mathcal{Z}^i \to \mathcal{P}(\mathcal{H})$, and an **estimator** is a map $\hat{h} : \bigcup_{i=1}^n \mathcal{Z}^i \to \mathcal{H}$. We will abbreviate $P(Z_{\leq n}) \in \mathcal{P}(\mathcal{H})$ to P_n, and denote P_0 any prior distribution, with the convention $P(\emptyset) := P_0$.

With a slight abuse of notation, for $m \in [n]$ and estimator $\hat{h}$, we denote $\hat{h}_{\leq m} := \hat{h}(Z_{\leq m})$, $\hat{h}_{<m} := \hat{h}(Z_{<m})$, $\hat{h}_{\geq m} := \hat{h}(Z_{\geq m})$, and $\hat{h}_{>m} := \hat{h}(Z_{>m})$.

Standard PAC-Bayesian Bounds

Both existing **state-of-the-art** PAC-Bayesian bounds and **ours** essentially take the following form; there exists constants $\mathcal{P}, \mathcal{A}, \mathcal{C} \geq 0$, and a function $\epsilon_{\delta,n}$, logarithmic in $1/\delta$ and n, such that for all $\delta \in]0,1[$, with probability at least $1-\delta$ over $Z_{\leq n}$, we have,

$$L(P_n) - L_n(P_n) \leq \mathcal{P} \cdot \sqrt{\frac{R_n \cdot (\mathrm{COMP}_n + \varepsilon_{\delta,n})}{n}} + \mathcal{A} \cdot \frac{\mathrm{COMP}_n + \varepsilon_{\delta,n}}{n} + \mathcal{C} \cdot \sqrt{\frac{R'_n \cdot \varepsilon_{\delta,n}}{n}}, \tag{1}$$

For **most bounds**, $R_n = L_n(P_n)$, $\mathrm{COMP}_n = \mathrm{KL}(P_n \| P_0)$, and $R'_n = 0$. For the **Tolstikhin and Seldin**'s empirical Bernstein bound $R_n = 1/n \cdot \mathbb{E}_{h\sim P_n}[\sum_{i=1}^n (\ell_h(Z_i) - L_n(P_n))^2]$ is the **empirical variance**.

For **our bound**, we have $R_n = V_n$ and $R'_n = V'_n$, where

$$\mathrm{COMP}_n = \mathrm{KL}(P_n \| P(Z_{\leq m})) + \mathrm{KL}(P_n \| P(Z_{>m})), \tag{2}$$

$$V'_n := \frac{1}{n}\sum_{i=1}^{m} \ell_{\hat{h}_{>i}}(Z_i)^2 + \frac{1}{n}\sum_{j=m+1}^{n} \ell_{\hat{h}_{<j}}(Z_j)^2, \tag{3}$$

$$V_n := \frac{1}{n}\mathbb{E}_{h\sim P_n}\left[\sum_{i=1}^{m}(\ell_h(Z_i) - \ell_{\hat{h}_{>i}}(Z_i))^2 + \sum_{j=m+1}^{n}(\ell_h(Z_j) - \ell_{\hat{h}_{<j}}(Z_j))^2\right].$$

Informed Priors and Stability

We managed to replace the typical $\mathrm{KL}(P_n \| P_0)$ term in other bounds by the COMP_n in (2); we are essentially using each half of the data to build **"informed priors"**; in this case, $P(Z_{\leq m})$ and $P(Z_{>m})$.

When the algorithm P is **sufficiently stable**, $\mathrm{COMP}_n \ll \mathrm{KL}(P_n \| P_0)$.

Other bounds can also be applied in a way to replace the KL term by the COMP_n in (2): e.g., an "informed" Maurer's bound becomes:

$$\mathrm{kl}(L(P_n), L_n(P_n)) \leq \frac{\mathrm{COMP}_n + \ln \frac{4\sqrt{m(n-m)}}{\delta}}{n}, \tag{4}$$

with probability at least $1-\delta$, for any fixed $\delta \in]0,1[$ and $m \in [0..n]$.

A Bound Based on Online Estimators

Our bound is based on the errors of the **online estimators** $(\hat{h}_{>i})$ and $(\hat{h}_{<j})$ which converge to the final $(\hat{h}_{\leq n})$ based on the full sample.

If P_n is concentrated around $\hat{h}_{\leq n}$; $\ell_{\hat{h}_{<j}}(Z_j) \simeq \ell_{\hat{h}_{\leq n}}(Z_j)$, $m < j \leq n$; and $\ell_{\hat{h}_{>i}}(Z_i) \simeq \ell_{\hat{h}_{\leq n}}(Z_i)$, $1 \leq i \leq m$, then $V_n \simeq 0$, leaving in our bound only the **lower order** term $O(\mathrm{COMP}_n/n)$ and the **complexity-free** term $O(\sqrt{V'_n/n})$. (The latter is of order $O(\sqrt{L(P_n)/n})$ w.h.p.)

Relation to the Excess Risk

Unlike other PAC-Bayesian bounds, ours can be related to excess risk bounds under the Bernstein condition which characterizes the "easiness" of the learning problem:

Definition 1 (Bernstein Condition). *A learning problem satisfies the (β, B)-Bernstein condition, for $\beta \in [0,1]$ and $B > 0$, if for all $h \in \mathcal{H}$,*

$$\mathbb{E}_{Z\sim\mathbf{D}}[(\ell_h(Z) - \ell_{h_*}(Z))^2] \leq B \cdot \mathbb{E}_{Z\sim\mathbf{D}}[\ell_h(Z) - \ell_{h_*}(Z)]^\beta,$$

where $h_* \in \arg\inf_{h\in\mathcal{H}} \mathbb{E}_{Z\sim\mathbf{D}}[\ell_h(Z)]$ is a risk minimizer within $\mathrm{cl}\,\mathcal{H}$.

Theorem 1 (Informal). *Let $m = \lceil n/2 \rceil$. Under a (β, B)-Bernstein condition, for any learning algorithm P and estimator $\hat{h}$ such that $\hat{h}_{>i} = \hat{h}_{>m}$ and $\hat{h}_{<j} = \hat{h}_{\leq m}$, for $1 \leq i \leq m < j \leq n$, the term $\sqrt{\frac{V_n \cdot \mathrm{COMP}_n}{n}}$ is of order*

$$\bar{L}(P_n) + \bar{L}(\hat{h}_{>m}) + \bar{L}(\hat{h}_{\leq m}) + (\mathrm{COMP}_n/n)^{\frac{1}{2-\beta}} \quad (\textit{log-factors omitted})$$

with high probability, where $\bar{L}(\cdot) := L(\cdot) - L(h_*)$ is the excess risk.

A New Concentration Inequality

Our new PAC-Bayesian bound is based on the following new concentration inequality:

Lemma 1. [Key result: *un*-expected Bernstein] *Let $X \sim \mathbf{D}$ be a random variable bounded from above by $b > 0$ almost surely, and let $\vartheta(u) := (-\ln(1-u) - u)/u^2$. For all $0 < \eta < 1/b$, we have (a)*

$$\mathbb{E}\left[e^{\eta(\mathbb{E}[X]-X)-\eta c \cdot X^2}\right] \leq 1, \quad \textit{for all } c \geq \eta \cdot \vartheta(\eta b). \tag{5}$$

(b) the result is tight: if $c < \eta \cdot \vartheta(\eta b)$, then $\exists \mathbf{D}$, for which (5) breaks.

Lemma 1 is reminiscent of the following slight variation of Bernstein's inequality; let X be any random variable bounded from *below* by $-b$, and let $\kappa(x) := (e^x - x - 1)/x^2$. For all $\eta > 0$, we have

$$\mathbb{E}\left[e^{\eta(\mathbb{E}[X]-X)-\eta c \cdot \mathbb{E}[X^2]}\right] \leq 1, \quad \text{for all } c \geq \eta \cdot \kappa(\eta b). \tag{6}$$

Note that the **un-expected Bernstein Lemma** 1 has the X^2 **lifted out** of the expectation.

Conclusion and Future Work

The main goal of this paper was to introduce a new PAC-Bayesian bound based on a new proof technique. In future work, we plan to put the bound to real practical use by applying it to deep neural networks.