PAC-Bayes-Empirical-Bernstein Inequality

Ilya Tolstikhin[1], Yevgeny Seldin[2,3]

[1]Computing Centre of Russian Academy of Sciences, [2]Queensland University of Technology, [3]UC Berkeley

Short Summary

PAC-Bayes-kl (Seeger '02) — State-of-the-art PAC-Bayesian bound

PAC-Bayes-Bernstein (Seldin et. al. '12) — Tighter, but depends on the expected variance (unknown)

Empirical Bernstein (Maurer & Pontil '09) + PAC-Bayes = PAC-Bayesian Bound on the Variance NEW

PAC-Bayesian Bound on the Variance + PAC-Bayes-Bernstein = PAC-Bayes-Empirical-Bernstein NEW

PAC-Bayes-Empirical-Bernstein depends on the empirical variance & tighter than PAC-Bayes-kl

Definitions

$\mathcal{X}$ — input space

$\mathcal{Y}$ — output space

$S = \{(x_i, y_i)\}_{i=1}^n$ — training sample $\overset{iid}{\sim} \mathcal{D}^n$

$h : \mathcal{X} \to \mathcal{Y}$ — hypothesis

$\mathcal{H}$ — hypothesis space

$\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ — loss function

$L(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(y, h(x))]$ — expected loss

$L_n(h) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, h(x_i))$ — empirical loss

$\mathbb{V}(h) = \mathrm{Var}_{(x,y)\sim\mathcal{D}}[\ell(y, h(x))]$ — variance

$\mathbb{V}_n(h) = \frac{1}{n-1}\sum_{i=1}^n \big(\ell(y_i, h(x_i)) - L_n(h)\big)^2$ — unbiased estimate of the variance

---

Randomized prediction rules:

ρ — a distribution over $\mathcal{H}$

For every point x:

1. Draw $h \sim \rho$
2. Return $h(x)$

$L(\rho) = \mathbb{E}_{h\sim\rho}[L(h)]$ — expected loss of ρ

$L_n(\rho) = \mathbb{E}_{h\sim\rho}[L_n(h)]$ — empirical loss of ρ

$\mathbb{V}(\rho) = \mathbb{E}_{h\sim\rho}[\mathbb{V}(h)]$ — average variance

$\mathbb{V}_n(\rho) = \mathbb{E}_{h\sim\rho}[\mathbb{V}_n(h)]$ — average empirical variance

---

For distributions ρ and π over $\mathcal{H}$ define:

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

For $p, q \in [0, 1]$ define:

$$\mathrm{kl}(q\|p) = \mathrm{KL}([q, 1-q]\|[p, 1-p]) \underset{\text{Pinsker}}{\geq} 2(q-p)^2$$

State-of-the-art PAC-Bayesian Inequalities

Fix a reference distribution π over $\mathcal{H}$. Then for any $\delta \in (0, 1)$, with probability greater than $1-\delta$ over a draw of S, the following inequalities hold for all distributions ρ over $\mathcal{H}$ simultaneously:

PAC-Bayes-kl **Inequality** (Seeger, 2002, Maurer, 2004)

$$\mathrm{kl}\big(L_n(\rho)\|L(\rho)\big) \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n} \qquad \textbf{(PB-kl)}$$

Relaxed PB-kl (McAllester, 2003) (uses the inequality $p \leq q + \sqrt{2q\mathrm{kl}(q\|p)} + 2\mathrm{kl}(q\|p)$ for $p > q$):

$$L(\rho) \leq L_n(\rho) + \sqrt{\frac{2L_n(\rho)\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}} + \frac{2\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}$$

PAC-Bayes-Bernstein inequality (Seldin et al., 2012) (**Simplified version**)
Under some technical conditions

$$L(\rho) \leq L_n(\rho) + \sqrt{\frac{3\mathbb{V}(\rho)\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{o(\dots)}{\delta}\right)}{n}} \qquad \textbf{(PB-B)}$$

The result also holds if $\mathbb{V}(\rho)$ is replaced by $\bar{V}(\rho)$ satisfying $\mathbb{V}(\rho) \leq \bar{V}(\rho) \leq \frac{1}{4}$ for all ρ.

Applications and Motivation

Applications of PAC-Bayesian bounds

- Generalization bounds for learning algorithms
- Parameter tuning (substitute for cross-validation)
- New algorithms that minimize the bounds

Motivation for our work

- PB-B is tighter than PB-kl whenever $\mathbb{V}(\rho) \ll L_n(\rho)$
- But $\mathbb{V}(\rho)$ is inaccessible in practice...
- $L_n(\rho)$ is a potentially loose upper bound on $\mathbb{V}(\rho)$ $\big($since for $\ell \in [0, 1]$: $\frac{n}{n-1}L_n(\rho) \geq \mathbb{V}_n(\rho) \to \mathbb{V}(\rho)\big)$
- We need a tighter upper bound on $\mathbb{V}(\rho)$ that holds for all ρ simultaneously

PAC-Bayesian Bound on the Variance

Theorem 1. *Fix a reference distribution π over $\mathcal{H}$. Then for for any $\delta_1 \in (0, 1)$, and any $c_1 > 1$, with probability greater than $1 - \delta_1$ over a draw of S, for all distributions ρ over $\mathcal{H}$ simultaneously:*

$$\mathbb{V}(\rho) \leq \underbrace{\mathbb{V}_n(\rho) + (1 + c_1)\sqrt{\frac{\mathbb{V}_n(\rho)\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{\nu_1}{\delta_1}\right)}{2(n-1)}} + \frac{2c_1\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{\nu_1}{\delta_1}\right)}{n-1}}_{V_n(\rho)}, \qquad (1)$$

where $\nu_1 = \left\lceil \frac{1}{\ln c_1} \ln\left(\frac{1}{2}\sqrt{\frac{n-1}{\ln(1/\delta_1)}} + 1 + \frac{1}{2}\right)\right\rceil$.

PAC-Bayes-Empirical-Bernstein Inequality

Theorem 2. *Let $V_n(\rho)$ denote the right hand side of (1) (with $\delta_1 = \frac{\delta}{2}$) and let $\bar{V}_n(\rho) = \min\left(V_n(\rho), 1/4\right)$. Fix a reference distribution π over $\mathcal{H}$. Then for any $\delta > 0$ and for any $c_1, c_2 > 1$, with probability greater than $1 - \delta$ over a draw of S we have:*

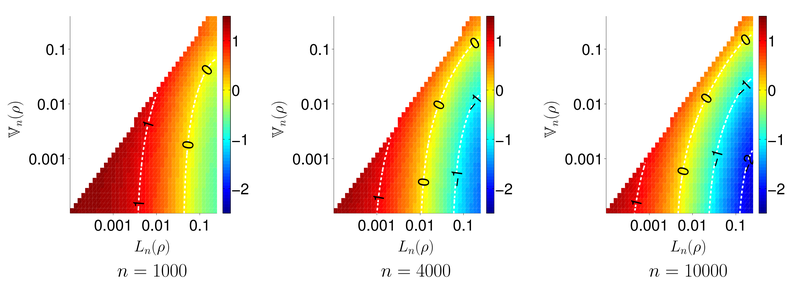
$$L(\rho) \leq L_n(\rho) + (1 + c_2)\sqrt{\frac{(e-2)\bar{V}_n(\rho)\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{2\nu_2}{\delta}\right)}{n}} \qquad \textbf{(PB-EB)}$$

simultaneously for all distributions ρ over $\mathcal{H}$ that satisfy $\sqrt{\frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{2\nu_2}{\delta}}{(e-2)\bar{V}_n(\rho)}} \leq \sqrt{n}$, *where* $\nu_2 = \left\lceil \frac{1}{\ln c_2} \ln\left(\sqrt{\frac{(e-2)n}{4\ln(2/\delta)}}\right)\right\rceil + 1$, *and for all other ρ we have:* $L(\rho) \leq L_n(\rho) + 2\frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{2\nu_2}{\delta}}{n}$.

When PB-EB is tighter than PB-kl?

PB-EB is tighter under the "- -0- -" line (high L_n, low $\mathbb{V}_n$ region). The advantage increases as the sample size grows.

$\log_2\left(\frac{\text{PB-EB}-L_n(\rho)}{\text{PB-kl}-L_n(\rho)}\right)$: logarithm of the ratio of the complexity terms.



$n = 1000$ $\quad n = 4000$ $\quad n = 10000$

Experiments

Linear regression with absolute loss:

$\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$

$\mathcal{Y} = [-0.5, 0.5]$

$\mathcal{H} = \{h_w(x) = \langle x, w\rangle : w \in \mathbb{R}^d, \|w\|_2 \leq 0.5\}$

$\ell(y, y') = |y - y'|$

We solve ERM: $\hat{w} = \arg\min_{w\in\mathcal{H}} L_n(h_w)$

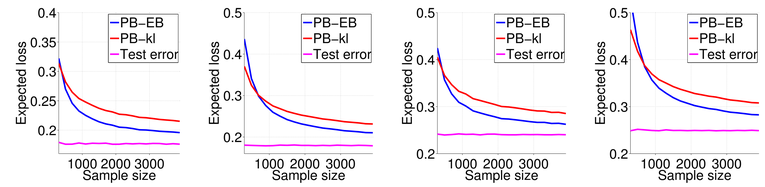
π is uniform over $\mathcal{H}$.

ρ is uniform over $\{w \in \mathbb{R}^d : \|w - \hat{w}\|_2 \leq \epsilon\}$

Set $\delta = \epsilon = 0.05$, $c_1 = c_2 = 1.15$

Synthetic datasets:

1. Draw $(x_1, \dots, x_n)$ i.i.d. uniformly from $\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$
2. 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 an independent noise



$d = 2$, $\|w_0\|_0 = 2$ $\quad d = 5$, $\|w_0\|_0 = 2$ $\quad d = 3$, $\|w_0\|_0 = 3$ $\quad d = 6$, $\|w_0\|_0 = 3$

UCI regression datasets:

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

Summary: PB-EB is tighter than PB-kl for all except very small n.

Proof Idea

Lemma 3 (PAC-Bayes). *For any function $f_n : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{R}$ and for any distribution π over $\mathcal{H}$, such that π is independent of S, with probability greater than $1-\delta$ over a random draw of S, for all distributions ρ over $\mathcal{H}$ simultaneously:*

$$\mathbb{E}_{h\sim\rho}\left[f_n(h, S)\right] \leq \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln\mathbb{E}_{h\sim\pi}\left[\mathbb{E}_{S'\sim\mathcal{D}^n}\left[e^{f_n(h,S')}\right]\right] \qquad (2)$$

Lemma 4 (Maurer and Pontil (2009)).

$$\mathbb{E}\left[e^{\lambda n(\mathbb{V}(h) - \mathbb{V}_n(h)) - \frac{\lambda^2}{2}\frac{n^2}{n-1}\mathbb{V}(h)}\right] \leq 1 \qquad (3)$$

Proof sketch of Thm 1: Take $f_n(h, S) = \lambda n\left(\mathbb{V}(h) - \mathbb{V}_n(h)\right) - \frac{\lambda^2}{2}\frac{n^2}{n-1}\mathbb{V}(h)$, substitute into (2), and apply (3). Massage the result and minimize it w.r.t. λ. The bound cannot be minimized simultaneously for all ρ by a single value of λ. Take a geometrically spaced grid of λ-s and replace the optimal λ by the closest λ from the grid (this gives c_1). Take a union bound over the grid (this gives ν_1).

Future Work

(1) Minimize the bound w.r.t. ρ; (2) More applications; (3) PAC-Bayes-kl-type bound for the variance.

References

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

A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. In *COLT*, 2009.

D. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1), 2003.

M. Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *JMLR*, 2002.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Trans. on Info. Theory*, 58, 2012.