PAC-Bayes and Information Complexity

Pradeep Kr. Banerjee[1] Guido Montúfar[1,2]

[1]Max Planck Institute for Mathematics in the Sciences (MPI MIS), Leipzig
[2]University of California, Los Angeles (UCLA)

European Research Council — Established by the European Commission

OpenReview

Max Planck Institute for Mathematics in the Sciences

UCLA

Overview

We point out that a number of well-known PAC-Bayesian-style and information-theoretic (IT) generalization bounds for randomized learning algorithms can be derived under a common framework starting from a fundamental *information exponential inequality*.

Three key ideas guide our discussion:

1. The lesser the information revealed by an algorithm about its input, the better the generalization.
2. Data-dependent priors entail tighter generalization bounds.
3. Optimizing such bounds is a natural recipe for designing new learning algorithms.

General formulation of learning problem

Training data S → Learning algorithm ($P_{W|S}$) → Hypothesis W

- Examples domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ of instances and labels
- Hypothesis space $\mathcal{W}$, and a fixed loss function $\ell : \mathcal{W} \times \mathcal{Z} \to [0, \infty)$
- A *learning algorithm*, which is a Markov kernel $P_{W|S}$ with
  - Input: Training data $S = (Z_1, \ldots, Z_n)$, $\quad Z_i \overset{\text{i.i.d.}}{\sim} \mu$
  - Output: hypothesis $W \in \mathcal{W}$, which is a random element of $\mathcal{W}$
- *True risk* of a hypothesis $w \in \mathcal{W}$ on μ, $L_\mu(w) := \mathbb{E}_\mu[\ell(w, Z)]$
- *Empirical risk* on the training sample S, $L_S(w) := \frac{1}{n}\sum_{i=1}^n \ell(w, Z_i)$

Goal is to control the *generalization error*, $\mathrm{g}(W, S) := L_\mu(W) - L_S(W)$, either in expectation or with high probability.

- The expected generalization error can be written as the difference of two expectations of the same loss function,

$$\mathbb{E}_{SW}[\mathrm{g}(W, S)] = \mathbb{E}_{P_S \otimes P_W}[L_S(W)] - \mathbb{E}_{P_{SW}}[L_S(W)],$$

where $P_{SW} = \mu^{\otimes n} \otimes P_{W|S}$.

Key insight. The expected generalization error reflects the dependence between the input data and the output hypothesis, and this dependence can be measured by their *mutual information* (MI).

The information exponential inequality

- For any $\beta > 0$, we define the *annealed expectation*, $M_\beta(w) = -\beta^{-1} \ln \mathbb{E}_\mu[e^{-\beta\ell(w,Z)}]$, which acts as a surrogate for $L_\mu(w)$.

***Lemma* 1** (Information exponential inequality, IEI [Zhang, 2006])**.** *For any prior Q over $\mathcal{W}$, any real-valued loss ℓ, and any posterior distribution $P \ll Q$ over $\mathcal{W}$ that depends on an i.i.d. training sample S, we have $\mathbb{E}_S \exp\{n\beta\mathbb{E}_P[M_\beta(W) - L_S(W)] - D(P\|Q)\} \le 1$.*

- The IEI implies bounds both in probability and in expectation for the quantity $n\beta\mathbb{E}_P[M_\beta(W) - L_S(W)] - D(P\|Q)$, and is the key tool for showing our main result:

***Theorem* 2.** *Let Q be a prior distribution over $\mathcal{W}$ that does not depend on S, and let ℓ be a real-valued loss function on $\mathcal{W} \times \mathcal{Z}$. Suppose that there exist a convex function $\psi : \mathbb{R}_{\ge 0} \to \mathbb{R}$ satisfying $\psi(0) = \psi'(0) = 0$, such that $\sup_{w\in\mathcal{W}}[L_\mu(w) - M_\beta(w)] \le \frac{\psi(\beta)}{\beta}$, $\forall \beta > 0$. Then, for any $\beta > 0$, and $\delta \in (0, 1]$, with probability of at least $1 - \delta$ over the choice of $S \sim \mu^{\otimes n}$, for all distributions $P \ll Q$ over $\mathcal{W}$ (even such that depend on S), we have*

$$\mathbb{E}_P[\mathrm{g}(W, S)] \le \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta}\right) + \frac{\psi(\beta)}{\beta}. \qquad (1)$$

Moreover, we have the following bound in expectation:

$$\mathbb{E}_{SW}[\mathrm{g}(W, S)] \le \psi^{*-1}\left(\frac{D(P\|Q|P_S)}{n}\right), \qquad (2)$$

where ψ^{*-1} is the inverse of the Fenchel-Legendre dual of ψ.

Recovering known IT and PAC-Bayes bounds

- Under a *sub-gaussian* loss assumption, we recover the **MI-based bound** due to [Xu and Raginsky, 2017]:

***Corollary* 3.** *If $\ell(w, Z)$ is σ-sub-Gaussian under μ for all $w \in \mathcal{W}$, then $\mathbb{E}_{SW}[\mathrm{g}(W, S)] \le \sqrt{2\sigma^2 I(S; W)/n}$.*

- Under a *sub-gamma* loss assumption, fixing $\beta = 1$ in (1), we recover the **PAC-Bayesian bound** due to [Germain et al., 2016]:

***Corollary* 4.** *If $\ell(w, Z)$ (σ, c)-sub-gamma with $c < 1$, then with probability of at least $1 - \delta$ over the choice of $S \sim \mu^{\otimes n}$, for all $P \ll Q$ over $\mathcal{W}$, $\mathbb{E}_P[\mathrm{g}(W, S)] \le \frac{1}{n}(D(P\|Q) + \ln(1/\delta)) + \frac{\sigma^2}{2(1-c)}$.*

Differentially private data-dependent priors

- To have a good control over the KL term in (1), it is desirable that Q be "aligned" with the data-dependent posterior P.

Key insight. Choosing Q based on S in a differentially private fashion allows us to treat Q "as if" it is independent of S.

- We have the following result:

***Theorem* 5.** *Let $\mathcal{K}(\mathcal{S}, \mathcal{W})$ denote the set of Markov kernels from $\mathcal{S}$ to $\mathcal{W}$. Let $Q^0 \in \mathcal{K}(\mathcal{S}, \mathcal{W})$ be an $(\epsilon, 0)$-differentially private algorithm. Let ℓ be a real-valued loss on $\mathcal{W} \times \mathcal{Z}$, let $\beta > 0$, and let $\delta \in (0, 1]$. Then with probability of at least $1 - \delta$ over the choice of $S \sim \mu^{\otimes n}$, for all distributions P over $\mathcal{W}$,*

$$\mathbb{E}_P[M_\beta(W)] \le \mathbb{E}_P[L_S(W)] + \frac{D(P\|Q^0(S)) + \ln\frac{2}{\delta} + \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\ln\frac{4}{\delta}}}{n\beta}.$$

Information complexity minimization (ICM)

- Given a prior, choosing a posterior to minimize a PAC-Bayesian bound gives rise to a method called *information complexity minimization*.
- Practical examples of ICM for learning with neural networks, e.g., Entropy-SGD [Chaudhari et al., 2017], can be viewed as optimization schemes that search for "flat minima" solutions.
- We show a PAC-Bayes bound motivated by an Occam's factor argument in relation to flat minima.

References

[Chaudhari et al., 2017] Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. (2017). Entropy-SGD: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations*.

[Germain et al., 2016] Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. (2016). PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892.

[Xu and Raginsky, 2017] Xu, A. and Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533.

[Zhang, 2006] Zhang, T. (2006). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321.