Fast-rate PAC-Bayes Generalization Bounds via Shifted Rademacher Processes

Jun Yang*, Shengyang Sun*, Daniel M. Roy

*Equal Contribution University of Toronto, Vector Institute

Contributions

Bridging Rademacher complexity and PAC-Bayesian Theory

- We extend the work of Kakade, Sridharan, and Tewari [1] to connect between Rademacher complexity and state-of-the-art PAC-Bayesian theory.
- Matching the fast rate PAC-Bayes bound by Catoni [2].
- Deriving a new fast-rate PAC-Bayes bound in terms of the **"flatness"** of the empirical risk surface on which the posterior concentrates.
- A new framework for deriving fast-rate PAC-Bayes bounds.

PAC-Bayes

- Data distribution $\mathcal{D}$ over the labeled space $\mathcal{Z}$; Hypothesis class $\mathcal{H}$; Binary loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \{0,1\}$; Loss class $\mathcal{F} := \{\ell(h,\cdot) : h \in \mathcal{H}\}$ of functions from $\mathcal{Z} \to \{0,1\}$. i.i.d. training set $S = (z_1, \cdots, z_m) \sim \mathcal{D}^m$.
- Gibbs classifiers. Distributions P on $\mathcal{F}$ as randomized classifiers that classify each new example according to a hypothesis drawn independently from P. For Gibbs classifiers, the (expected) risk and the (expected) empirical risk are defined respectively as:

$$\mathcal{L}_{\mathcal{D}}(P) := \mathbb{E}_{f\sim P}\mathcal{L}_{\mathcal{D}}(f) = \mathbb{E}_{z\sim\mathcal{D}}\mathbb{E}_{f\sim P}[f(z)] = \mathbb{E}_{z\sim\mathcal{D}}\mathbb{E}_P f(z)$$

$$\hat{\mathcal{L}}_S(P) := \mathbb{E}_{f\sim P}\hat{\mathcal{L}}_S(f) = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{f\sim P}[f(z_i)] = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_P f(z_i).$$

Theorem 1 (PAC-Bayes Bound by McAllester [3].) *For any prior distribution P over $\mathcal{F}$, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of training data $S \sim \mathcal{D}^m$, for all distributions Q over $\mathcal{F}$,*

$$\mathcal{L}_{\mathcal{D}}(Q) \le \hat{\mathcal{L}}_S(Q) + \sqrt{\frac{\mathrm{KL}\left(Q||P\right) + \log\frac{m}{\delta}}{2(m-1)}}. \tag{1}$$

Theorem 2 (Fast-Rate PAC-Bayes Bound by Catoni [2, Thm 1.2.6].) *For any prior distribution P over $\mathcal{F}$, for any $\delta \in (0,1)$ and $C > 0$, with probability at least $1-\delta$ over draws of training data $S \sim \mathcal{D}^m$, for all distributions Q over $\mathcal{F}$,*

$$\mathcal{L}_{\mathcal{D}}(Q) \le \frac{1}{1-e^{-C}}\left[C\hat{\mathcal{L}}_S(Q) + \frac{\mathrm{KL}\left(Q||P\right) + \log\frac{1}{\delta}}{m}\right]. \tag{2}$$

References

[1] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*, pages 793–800, 2008.

[2] Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Lecture Notes – Monograph Series*. Institute of Mathematical Statistics, 2007.

[3] David A. McAllester. PAC-Bayesian model averaging. In *Conference on Learning Theory*, pages 164–170, 1999.

[4] Nikita Zhivotovskiy and Steve Hanneke. Localization of VC classes: Beyond local Rademacher complexities. *Theoretical Computer Science*, 742:27–49, 2018.

Shifted Rademacher Processes

- We work with processes of the form $\{\frac{1}{m}\sum_{i=1}^m \epsilon_i' f(z_i)\}_{f\in\mathcal{F}}$ where the variables $\{\epsilon_i'\}$ are independent from S, i.i.d., and take two values with equal probability.
- The shifted Rademacher variables, $\{\epsilon_i'\}$, are not necessarily zero mean. When they take values in $\{\pm 1\}$, we obtain a standard Rademacher process.
- Define $\mathcal{G}_\kappa := \{\mathbb{E}_Q f(\cdot) : \mathrm{KL}\left(Q||P\right) \le \kappa\}$, and, by an abuse of notation, let $\mathcal{L}_{\mathcal{D}}(g)$ and $\hat{\mathcal{L}}_S(g)$ denote $\mathbb{E}_{z\sim\mathcal{D}}[g(z)]$ and $\frac{1}{m}\sum_{i=1}^m g(z_i)$, respectively.

Symmetrization and Slow-rate PAC-Bayes Bound

- By McDiarmid's inequality, with probability at least $1-\delta$,

$$\sup_{g\in\mathcal{G}_\kappa}\left[\mathcal{L}_{\mathcal{D}}(g) - \hat{\mathcal{L}}_S(g)\right] \le \mathbb{E}_S \sup_{g\in\mathcal{G}_\kappa}\left[\mathcal{L}_{\mathcal{D}}(g) - \hat{\mathcal{L}}_S(g)\right] + \sqrt{\frac{\log(1/\delta)}{m}}. \tag{3}$$

- By symmetrization, with i.i.d. Rademacher random variables $\mathbb{P}(\epsilon_i = +1) = \mathbb{P}(\epsilon_i = -1) = 1/2$,

$$\mathbb{E}_S \sup_{g\in\mathcal{G}_\kappa}\left[\mathcal{L}_{\mathcal{D}}(g) - \hat{\mathcal{L}}_S(g)\right] \le 2\mathbb{E}_S\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}_\kappa}\left[\frac{1}{m}\sum_{i=1}^m \epsilon_i g(z_i)\right]. \tag{4}$$

- Kakade et al. [1] use union bound over κ, bounding Rademacher complexity

$$\mathbb{E}_S\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}_\kappa}\left[\frac{1}{m}\sum_{i=1}^m \epsilon_i g(z_i)\right] = \mathcal{O}(\sqrt{\kappa/m}). \tag{5}$$

Theorem 3 (PAC-Bayes Bound via Rademacher Complexity by Kakade et al. [1]) *for every prior P over $\mathcal{F}$, with probability at least $1-\delta$ over draws of training data $S \sim \mathcal{D}^m$, for all distribution Q over $\mathcal{F}$,*

$$\mathcal{L}_{\mathcal{D}}(Q) \le \hat{\mathcal{L}}_S(Q) + 4.5\sqrt{\frac{\max\{\mathrm{KL}\left(Q||P\right), 2\}}{m}} + \sqrt{\frac{\log(1/\delta)}{m}}. \tag{6}$$

Shifted Symmetrization and Fast-rate PAC-Bayes Bound

- Using shifted symmetrization in deviation by Zhivotovskiy and Hanneke [4],

$$\mathbb{P}_S\left(\sup_{g\in\mathcal{G}_\kappa}\mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g) \ge t\right) \le 4\mathbb{P}_{S,\epsilon}\left(\sup_{g\in\mathcal{G}_\kappa}\left[\frac{1+\frac{c'}{2}}{m}\sum_{i=1}^m \epsilon_i' g(z_i)\right] \ge \frac{t'}{2}\right). \tag{7}$$

  where $c > c_2 > 0$, $c' = \frac{c-c_2}{1+c_2}$, $t' = \frac{t}{2(1+c_2)}$, and $\epsilon_i' := \epsilon_i - \frac{c'}{2+c'}$ are shifted Rademacher random variables.
- Further bounding the shifted Rademacher process and performing union bounds over κ, we obtain a fast-rate PAC-Bayes bound matching Catoni's,

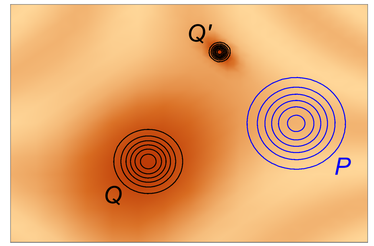
Theorem 4 (Matching Catoni's Fast Rate via Shifted Rademacher Processes) *For any given $c > 0$ and prior P over $\mathcal{F}$, there exist constants C_1, C_2, and C_3 such that, with probability at least $1-\delta$, for all distributions Q over $\mathcal{F}$,*

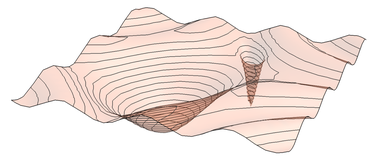
$$\mathcal{L}_{\mathcal{D}}(Q) \le (1+c)\hat{\mathcal{L}}_S(Q) + C_1\frac{\mathrm{KL}\left(Q||P\right)}{m} + C_2\frac{\log\frac{1}{\delta}}{m} + C_3\frac{1}{m}. \tag{8}$$

Fast-rate PAC-Bayes Bound in terms of "Flatness"

Definition (Notion of "Flatness") For given $h \in [0,1]$, the "h-flatness" of Q (w.r.t. S) is

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}_Q[f(z_i) - (1+h)\mathbb{E}_Q f(z_i)]^2. \tag{9}$$





- Note that flatness-augmented generalization can be written as

$$\mathcal{L}_{\mathcal{D}}(Q) - \hat{\mathcal{L}}_S(Q) - \frac{c}{m}\sum_{i=1}^m \mathbb{E}_Q[f(z_i) - (1+h)\mathbb{E}_Q f(z_i)]^2 = \mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g) + c(1-h^2)\hat{\mathcal{L}}_S(g^2). \tag{10}$$

- A shifted symmetrization inequality with "flatness",

$$\frac{1}{4}\mathbb{P}_S\left(\sup_{g\in\mathcal{G}_\kappa}\mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g) + c(1-h^2)\hat{\mathcal{L}}_S(g^2) \ge t\right) \le \mathbb{P}_{S,\epsilon}\left(\sup_{g\in\mathcal{G}_\kappa}\left[\frac{1}{m}\sum_{i=1}^m(\epsilon_i + \epsilon_i'')g(z_i) - \epsilon_i''(1-h^2)g^2(z_i)\right] \ge \frac{t}{4}\right), \tag{11}$$

  where $c' = \frac{c+c_2}{2}$, $c'' = \frac{c-c_2}{2}$, $\epsilon_i'' := \epsilon_i c' - c''$.
- Combining the new shifted symmetrization inequality and bounding the resulting shifted Rademacher process, we have the following result.

Theorem 5 (Fast Rate PAC-Bayes using "Flatness".) *For any given $c > 0$ and $h \in (0,1)$, with probability at least $1-\delta$ over random draws of training set $S \sim \mathcal{D}^m$, for all distributions Q over $\mathcal{F}$,*

$$\mathcal{L}_{\mathcal{D}}(Q) \le \hat{\mathcal{L}}_S(Q) + \frac{c}{m}\sum_{i=1}^m \mathbb{E}_Q[f(z_i) - (1+h)\mathbb{E}_Q f(z_i)]^2 \tag{12}$$

$$+ \frac{4}{Cm}\left[3\mathrm{KL}\left(Q||P\right) + \log\frac{1}{\delta} + 5\right], \tag{13}$$

where $C = \frac{2h^4 c}{1+16h^2 c}$.

Potential Directions

- Comparing the Rademacher-process and the direct PAC-Bayes approaches.
- Pursuing PAC-Bayes bounds via Talagrand's concentration inequalities.
- Extending the binary loss to bounded and unbounded losses.
- Using Rademacher-process techniques in the development of PAC-Bayesian analyses of large-scale neural networks trained by stochastic gradient descent.
- Studying the empirical performance in the context of large-scale neural networks.