

Rademacher complexity and tuning parameter

Chanwoo Lee, June 21, 2020

1 Rademacher complexity

Theorem 1.1. Let $K : \mathbb{R}^{d_1 \times d_2} \times \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^{d_1 \times d_1}$ be bounded with $\sqrt{\text{tr}(K(\mathbf{X}, \mathbf{X}))} \leq G$ and let $h : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^{d_1 \times d'_2}$ be a corresponding feature mapping such that $K(\mathbf{X}, \mathbf{X}') = h(\mathbf{X})h(\mathbf{X}')^T$. Define

$$\mathcal{F}_r = \{f : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R} : f(\mathbf{X}) = \langle \mathbf{B}, h(\mathbf{X}) \rangle \text{ with } \mathbf{B} \in \mathcal{B}\},$$

where $\mathcal{B} = \{\mathbf{B} \in \mathbb{R}^{d_1 \times d'_2} : \text{rank}(\mathbf{B}) \leq r, \lambda_1(\mathbf{B}) \leq M\}$. Then

$$\mathcal{R}_n(\mathcal{F}_r) \leq \frac{2MG\sqrt{r}}{\sqrt{n}}.$$

Proof. Since the hinge loss is an 1-Lipschitz function,

$$\mathcal{R}_n(\mathcal{F}_r) = 2\mathbb{E} \sup_{\mathbf{B} \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n \sigma_i (1 - y_i \langle \mathbf{B}, h(\mathbf{X}_i) \rangle)_+ \leq 2\mathbb{E} \sup_{\mathbf{B} \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle \mathbf{B}, h(\mathbf{X}_i) \rangle,$$

where $\{\sigma_i\}_{i=1}^n$ is independent Rademacher random variables with $\mathbb{P}(\sigma_i = \pm 1) = 1/2$. The result follows by observing

$$\begin{aligned} 2\mathbb{E} \sup_{\mathbf{B} \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle \mathbf{B}, h(\mathbf{X}_i) \rangle &= \frac{2}{n} \mathbb{E} \sup_{\mathbf{B} \in \mathcal{B}} \langle \mathbf{B}, \sum_{i=1}^n \sigma_i h(\mathbf{X}_i) \rangle \\ &\leq \frac{2}{n} \mathbb{E} \sup_{\mathbf{B} \in \mathcal{B}} \|\mathbf{B}\| \left\| \sum_{i=1}^n \sigma_i h(\mathbf{X}_i) \right\|, \text{ by Cauchy-Schwartz inequality} \\ &\leq \frac{2}{n} \mathbb{E} \sup_{\mathbf{B} \in \mathcal{B}} \lambda_1 \sqrt{r} \left\| \sum_{i=1}^n \sigma_i h(\mathbf{X}_i) \right\| \\ &\leq \frac{2M\sqrt{r}}{n} \mathbb{E} \left\| \sum_{i=1}^n \sigma_i h(\mathbf{X}_i) \right\| \\ &\leq \frac{2M\sqrt{r}}{n} \sqrt{\mathbb{E} \left\langle \sum_{i=1}^n \sigma_i h(\mathbf{X}_i), \sum_{i=1}^n \sigma_i h(\mathbf{X}_i) \right\rangle} \text{ by Jensen's inequality} \\ &\leq \frac{2M\sqrt{r}}{n} \sqrt{\sum_{i=1}^n \|h(\mathbf{X}_i)\|^2} \\ &= \frac{2M\sqrt{r}}{n} \sqrt{\sum_{i=1}^n \text{tr}(K(\mathbf{X}_i, \mathbf{X}_i))} \leq \frac{2MG\sqrt{r}}{\sqrt{n}}. \end{aligned}$$

□

Remark 1. If we choose K as a linear kernel, Theorem 1.1 reduces to the linear SMM Rademacher complexity case.

Corollary 1.1. *Assume the same condition in Theorem 1.1. Then, with probability at least $1 - \delta$, the generalization error of low-rank SMM is*

$$\mathbb{P}\{Y^{new} \neq \text{sign}(\hat{f}(\mathbf{X}^{new}))\} \leq \text{training error} + \frac{MG\sqrt{r}}{\sqrt{n}} + \sqrt{\frac{\ln(\frac{1}{\delta})}{2n}}.$$

2 Tuning parameter

In the probability estimation, we assume that we selected the optimal tuning parameter λ and the rank r . In practice, the tuning parameter selection can be done using an independent validation set or cross-validation. From the available dataset N , I propose to use one half for training and the other half for tuning, i.e. $n_{\text{train}} = n_{\text{tune}} = \frac{N}{2}$.

Detail procedure for parameter tuning is as follows

1. We obtain the tuning grid $\{(\lambda_i, r_j) : \lambda_1 < \dots < \lambda_M, r_j = j \in \{1, \dots, \min(d_1, d_2)\}\}$
2. For a given (λ_i, r_j) we obtain probability estimates

$$\hat{p}^{(\lambda_i, r_j)}(\mathbf{X}_k) = \hat{\mathbb{P}}^{(\lambda_i, r_j)}(y = 1 | \mathbf{X}_k), \quad k = 1, \dots, n_{\text{tune}}.$$

3. We evaluate the log-likelihood

$$L(\lambda_i, r_j) = \sum_{k=1}^{n_{\text{tune}}} \log(\hat{p}^{(\lambda_i, r_j)}(\mathbf{X}_k)).$$

4. We choose the optimal tuning parameter $(\lambda_{\hat{i}}, r_{\hat{j}})$ that minimizes BIC value based on the log-likelihood

$$(\hat{i}, \hat{j}) = \arg \min_{i,j} -2L(\lambda_i, r_j) + r_j(d_1 + d_2 - r_j) \log(n_{\text{tune}}).$$

This grid search might requires too many calculations because we have to perform $M \times \min(d_1, d_2)$ times probability estimation. One way to avoid this grid search is to find the best tuning parameters using profile method. First, fix rank r first and find the best λ . Second, find the best rank r fixing the obtained λ . In this way, we can reduce the number of trials to $M + \min(d_1, d_2)$.

3 Consistency of Probability estimation

Our estimation method is based on the following optimization problem.

$$\min_{f \in \mathcal{F}} n^{-1} \left[(1 - \pi) \sum_{y_i=1} (1 - y_i f(\mathbf{X}_i))_+ + \pi \sum_{y_i=-1} (1 - y_i f(\mathbf{X}_i))_+ \right] + \lambda J(f). \quad (1)$$

In (1), when $n \rightarrow \infty$, the first component approaches

$$\mathbb{E}[S(Y)(1 - Yf(\mathbf{X}))_+] \quad \text{where } S(Y) = 1 - \pi \text{ if } Y = 1, \text{ and } \pi \text{ otherwise.} \quad (2)$$

We prove that minimizing (2) with respect to f yields the Bayes rule $\bar{f}_\pi(\mathbf{X}) = \text{sign}(p(\mathbf{X}) - \pi)$ where $p(\mathbf{X}) = \mathbb{P}(Y = 1 | \mathbf{X})$. The following theorem is referred from [1]. Every argument works through

on our setting because this theorem is specified in terms of complexity of considered function space.

Define $e_V(f, \bar{f}_\pi) = \mathbb{E}\{V(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)\}$ where $V(f, \mathbf{X}, y) = S(y)(1 - yf(\mathbf{X}))_+$. There are three assumptions to be made for the theorem.

Assumption 1. *For some positive sequence such that $s_n \rightarrow 0$ as $n \rightarrow \infty$, there exists $f_\pi^* \in \mathcal{F}$ such that $e_V(f_\pi^*, \bar{f}_\pi) \leq s_n$.*

Assumption 1 ensures that the Bayes rule \bar{f}_π is well approximated by \mathcal{F} .

Define a truncated V by $V^T(f, \mathbf{X}, y) = V(f, \mathbf{X}, y)\mathbf{1}\{V(f, \mathbf{X}, y) \leq T\} + T\mathbf{1}\{V(f, \mathbf{X}, y) > T\}$ for some truncation constant T such that $\max\{V(\bar{f}_\pi, \mathbf{X}, y), V(f_\pi^*, \mathbf{X}, y)\} \leq T$ almost surely, and $e_{V^T}(f, \bar{f}_\pi) = \mathbb{E}\{V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)\}$.

Assumption 2. *There exist constants $0 \leq \alpha < \infty, 0 \leq \beta \leq 1, a_1 > 0$ and $a_2 > 0$ such that, for any sufficiently small $\delta > 0$,*

$$\begin{aligned} \sup_{\{f \in \mathcal{F}: e_{V^T}(f, \bar{f}_\pi) \leq \delta\}} \| \text{sign}(f) - \text{sign}(\bar{f}_\pi) \|_1 &\leq a_1 \delta^\alpha, \\ \sup_{\{f \in \mathcal{F}: e_{V^T}(f, \bar{f}_\pi) \leq \delta\}} \text{var}\{V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)\} &\leq a_2 \delta^\beta. \end{aligned}$$

Assumption 2 describe local smoothness within a neighborhood of \bar{f}_π .

We define the L_2 metric entropy with bracketing that measures the cardinality of \mathcal{F} . Given any $\epsilon > 0$, define $\{(f_m^\ell, f_m^u)\}_{m=1}^M$ to be an ϵ -bracketing function set of \mathcal{F} if for any $f \in \mathcal{F}$, there exists an m such that $f_m^\ell \leq f \leq f_m^u$ and $\|f_m^\ell - f_m^u\|_2 \leq \epsilon$ for $m = 1, \dots, M$. Then L_2 -metric entropy with bracketing $H_2(\epsilon, \mathcal{F})$ is defined as the logarithm of the cardinality of the smallest ϵ -bracketing function set of \mathcal{F} . Let $\mathcal{F}^V(k) = \{V^T(f, \mathbf{X}, y) - V(f_\pi^*, \mathbf{X}, y) : f \in \mathcal{F}(k)\}$ where $\mathcal{F}(k) = \{f \in \mathcal{F} : \frac{1}{2}\|f\|_k^2 \leq k\}$ and $J_\pi^* = \max\{J(f_\pi^*), 1\}$.

Assumption 3. *For some constant $a_3, a_4, a_5 > 0$, and $\epsilon_n > 0$,*

$$\sup_{k \geq 2} \int_{a_4 L}^{\sqrt{a_3 L^\beta}} \sqrt{H_2(\omega, \mathcal{F}^V(k))} d\omega / L \leq a_5 \sqrt{\epsilon_n}, \text{ where } L = L(\epsilon, \lambda, k) = \min\{\epsilon^2 + \lambda(k/2 - 1)J_\pi^*, 1\}.$$

Theorem 3.1. *Under Assumptions 1-3, for the estimator \hat{p} obtained from our method, there exists a constant $a_6 > 0$ such that*

$$\mathbb{P} \left\{ \|\hat{p} - p\|_1 \geq \frac{1}{2m} + \frac{1}{2} a_1 (m+1) \delta_n^{2\alpha} \right\} \leq 15 \exp\{-a_6 n (\lambda J_\pi^*)^{2-\beta}\},$$

provided that $\lambda^{-1} \geq 4\delta_n^{-2} J_\pi^$, where $\delta_n^2 = \min\{\max(\epsilon_n^2, s_n), 1\}$*

4 Covering number bounds of linear function classes

From theorems in [2], we can calculate the entropy of linear function class in our setting. This following lemma might be helpful for checking assumptions in Section 3.

Lemma 1. Define $\mathcal{F} = \{f : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R} : f(\mathbf{X}) = \langle \mathbf{B}, \mathbf{X} \rangle \text{ with } \mathbf{B} \in \mathcal{B}\}$ under the condition that $\|\mathbf{X}\| \leq G$, there exists constraints $c, c' > 0$ such that for all $n \in \mathbb{N}$ and all $\epsilon > 0$,

$$\log_2 H_2(\epsilon, \mathcal{F}) \leq \left\lfloor \frac{M^2 G^2 r}{\epsilon^2} \right\rfloor \log_2(2d_1 d_2 + 1).$$

References

- [1] Junhui Wang, Xiaotong Shen, and Yufeng Liu. Probability estimation for large-margin classifiers. *Biometrika*, 95(1):149–167, March 2008.
- [2] Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.