

# Rademacher complexity and consistency of the estimation

Chanwoo Lee, June 25, 2020

## 1 Rademacher complexity

Based on many lecture notes and papers related to Rademacher complexity, I find general theorem about the error bound.

**Theorem 1.1.** Let  $\ell$  and  $\mathcal{F}$  be a considered loss function and function space. From  $\{(\mathbf{X}_i, y_i)\}_{i=1}^n$  i.i.d. drawn samples, with probability at least  $1 - \delta$ , we have the following inequality.

$$\sup_{f \in \mathcal{F}} \left[ \mathbb{E}_{\mathbf{X}, y}(\ell(y, f(\mathbf{X}))) - \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(\mathbf{X}_i)) \right] \leq \mathcal{R}_n(\ell \circ \mathcal{F}) + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}},$$

where  $\ell \circ \mathcal{F} = \{\ell \circ f : (\mathbf{X}, y) \mapsto \ell(y, f(\mathbf{X})) : f \in \mathcal{F}\}$  and  $\mathcal{R}_n(\mathcal{G}) = 2\mathbb{E} \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(\mathbf{X}_i)$ .

In particular, when  $\mathcal{F}$  is a set of  $\{-1, 1\}$ -valued functions defined on  $\mathcal{X}$  and  $\ell(y, f(\mathbf{X})) = \mathbb{1}\{y \neq f(\mathbf{X})\}$ , one can show  $\mathcal{R}_n(\ell \circ \mathcal{F}) = \frac{\mathcal{R}_n(\mathcal{F})}{2}$  (you can check [1]) so that we have the following generalization error, which we based on: For all  $f \in \mathcal{F}$ ,

$$\mathbb{P}[Y \neq f(\mathbf{X})] \leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{y_i \neq f(\mathbf{X}_i)\} + \frac{\mathcal{R}_n(\mathcal{F})}{2} + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}. \quad (1)$$

**Remark 1.** From the definition of sup in Theorem 1.1, Equation (1) holds for any function in  $\mathcal{F}$ .

**Remark 2.** Equation (1) holds only when considered function class is a set of  $\{-1, 1\}$ -valued functions. So we cannot directly apply Rademacher complexity of linear predictors.

**Remark 3.** In [3], they bound the Rademacher complexity using entropy of  $\mathcal{F}$ . But I am not sure they consider  $\mathcal{F}$  as a set of  $\{-1, 1\}$ -valued functions. I think the reason of the authors using entropy is to find the general Rademacher complexity not confined in Euclidean space. In [2] where covering number is used for Rademacher complexity, I can check the authors use covering number for the Rademacher complexity in more general settings than Euclidean spaces.

I find a new way to utilize the Rademacher complexity of linear predictors such that

$$\mathcal{R}_n(\mathcal{F}_r) = 2\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\mathbf{X}_i) \leq \frac{2MG\sqrt{r}}{\sqrt{n}}, \quad (2)$$

where  $\mathcal{F}_r = \{f : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R} : f(\mathbf{X}) = \langle \mathbf{B}, \mathbf{X} \rangle \text{ with } \mathbf{B} \in \mathcal{B}\}$ ,  $\mathcal{B} = \{\mathbf{B} \in \mathbb{R}^{d_1 \times d_2} : \text{rank}(\mathbf{B}) \leq r, \lambda_1(\mathbf{B}) \leq M\}$ , and  $G = \max \|\mathbf{X}\|$ .

Change order between these two conditions.

"With probability at least ..., the following holds for all f ..."

**Theorem 1.2.** Let loss  $\varphi$  be  $L$ -Lipchitz and greater than 0/1 loss. For any  $f \in \mathcal{F}_r$ , with probability at least  $1 - \delta$ ,

Explain: 1. fix f first, with probability ...  $\Rightarrow$  CLT (pointwise consistency; easy)  
2. fix probability first ..., holds for all f  $\Rightarrow$  generalization error (uniform consistency)

$$\mathbb{P}[Y \neq f(\mathbf{X})] \leq \frac{\text{fixed number}}{n} \sum_{i=1}^n \varphi(y_i f(\mathbf{X}_i)) + \frac{2LMG\sqrt{r}}{\sqrt{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}.$$

1. (generalization error) worst-case discrepancy between risk and empirical risk. worst case w.r.t. f in the class F

2. (estimation error)  $\hat{f}$  vs.  $f_{\text{bayes}}$

loss ( $\hat{f}$ ) vs. loss ( $f_{\text{bayes}}$ ); loss = hinge + penalty

*Proof.* Note that

$$\mathbb{P}[Y \neq f(\mathbf{X})] = \mathbb{E}[\mathbb{1}\{y(f(\mathbf{X})) < 0\}] \leq \mathbb{E}(\varphi(yf(\mathbf{X}))) \leq \frac{1}{n} \sum_{i=1}^n \varphi(y_i f(\mathbf{X}_i)) + \mathcal{R}_n(\varphi \circ \mathcal{F}_r) + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}.$$

Theorem 1.1 is used in the last inequality. The Rademacher complexity term is bounded by the following inequality.

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

Therefore, (2) completes the theorem.  $\square$

**Remark 4.** We can apply the theorem with hinge loss or logistic loss with  $L = 1$  because  $\mathbb{1}\{yf(\mathbf{X}) < 0\} \leq \ell_{\text{hinge}}(yf(\mathbf{X}))$  and  $\mathbb{1}\{yf(\mathbf{X}) < 0\} \leq \ell_{\text{logistic}}(yf(\mathbf{X}))$

## 2 Consistency of the probability estimation

We have 3 main assumptions for the consistency of the probability estimation.

**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 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  $\delta > 0$ ,

$$\begin{aligned} \sup_{\{f \in \mathcal{F}: e_{VT}(f, \bar{f}_\pi) \leq \delta\}} \| \text{sign}(f) - \text{sign}(\bar{f}_\pi) \|_1 &\leq a_1 \delta^\alpha, \\ \sup_{\{f \in \mathcal{F}: e_{VT}(f, \bar{f}_\pi) \leq \delta\}} \text{var}\{V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)\} &\leq a_2 \delta^\beta. \end{aligned} \quad (3)$$

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

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

**Remark 5.** Equation (3) in Assumption 2 can be made interpretable. Consider the following equation.

$$\begin{aligned} \text{var}\{V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)\} &\leq \mathbb{E}|V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)|^2 \\ &\leq T \mathbb{E}|V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)| \\ &= T \|V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)\|. \end{aligned}$$

(3) can be replaced by

$$\sup_{\{f \in \mathcal{F}: e_{VT}(f, \bar{f}_\pi) \leq \delta\}} \|V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)\|_1 \leq a_2 \delta^\beta / T.$$

Therefore, the equations in Assumption 2 control local smoothness of the classifier function and truncated loss function.

**Remark 6.** Assumption 3 measures the complexity of considered function space. Notice that

$$H_2(\epsilon, \mathcal{F}^V(k)) \leq H_2(\epsilon, \mathcal{F}(k)) \leq H_\infty(\epsilon, \mathcal{F}(k)),$$

because for functions  $f_\ell$  and  $f_u$ ,  $\|V^T(f_\ell, \cdot) - V^T(f_u, \cdot)\|_2 \leq \|f_\ell - f_u\|_2 \leq \|f_\ell - f_u\|_\infty$ . I assume that  $H_2(\epsilon, \mathcal{F}^V(k))$  is replaced by  $H_s(\epsilon, \mathcal{F}(k))$  where  $s = 2$  or  $\infty$ , for better interpretation sacrificing weak assumption. Then, solving the equation in Assumption 3,

$$g(\sqrt{a_3 L \beta}) - g(a_4 L) = \sup_{k \geq 2} \int_{a_4 L}^{\sqrt{a_3 L \beta}} \sqrt{H_s(\omega, \mathcal{F}(k))} d\omega \leq a_5 \sqrt{n}, \quad (4)$$

can we find varepsilon as a function of g(n, r, d)

gives us the relation of  $\epsilon_n = g(n)$ , which determines the value  $\delta_n$  in the convergence rate in Theorem 2.1. Integration of entropy is closely related to upper bound of Rademacher complexity (Dudley's theorem) such that

$$\sqrt{\mathcal{H}(\text{rank } r)} = \text{function}(r) * \sqrt{\mathcal{H}(\text{full rank})}$$

$$\begin{aligned} \hat{\mathcal{R}}_n(\mathcal{F}) &\leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\epsilon}{4}}^{\infty} \sqrt{H_\infty(\omega, \mathcal{F})} d\omega, \text{ or} \\ \hat{\mathcal{R}}_n(\mathcal{F}) &\leq \inf_{\epsilon \leq 0} 4\epsilon + 12 \int_{\epsilon}^{\infty} \sqrt{\frac{H_2(\omega, \mathcal{F})}{n}} d\omega. \end{aligned} \quad (5)$$

Since when solving (4), we only consider  $\mathcal{O}(\max(g(\sqrt{L^\beta}), g(L)))$ , the upper bounds of (5) have the same order with the left side term of (4). Therefore, we can relate Rademacher complexity with Assumption 3 with stricter condition.

**Theorem 2.1.** 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  $\delta_n^2 = \min\{\max(\epsilon_n^2, s_n), 1\}$ . Simplified version of the above argument is

integration is used to bound the excess risk (sample, population):  
sample risk at  $\hat{f}$

$$\|\hat{p} - p\|_1 = \mathcal{O}_p \left\{ \frac{1}{m} + a_1 (m+1) \delta_n^{2\alpha} \right\}, \quad \text{vs. population risk at } \hat{f} \text{ bayes}$$

where  $\hat{f}$  is the minimizer of the penalized SVM

provided that  $n(\lambda J_\pi^*)^{2-\beta}$  is bounded away from 0.

## References

- [1] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.
- [2] Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with lipschitz functions. *Journal of Machine Learning Research*, 5(Jun):669–695, 2004.
- [3] Kush R Varshney and Alan S Willsky. Linear dimensionality reduction for margin-based classification: high-dimensional data and sensor networks. *IEEE Transactions on Signal Processing*, 59(6):2496–2512, 2011.