Consistency of probability estimation

Chanwoo Lee, June 28, 2020

1 Linear case

Define

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

where $\mathcal{B} = \{\boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2} : \text{rank}(\boldsymbol{B}) \leq r, \lambda_1(\boldsymbol{B}) \leq M\}$. Let $\bar{f}_\pi$ be a Bayes rule. In addition, let $e_V(f, \bar{f}_\pi) = \mathbb{E}\{V(f, \boldsymbol{X}, y) - V(\bar{f}_\pi, \boldsymbol{X}, y)\}$ with $V(f, \boldsymbol{X}, y) = S(y)L\{yf(\boldsymbol{X})\}$.

Based on function class $\mathcal{F}_r$, we have the following theorem.

Theorem 1.1 (linear case)**.** *Assume that*

1. *For some positive sequence such that $s_n \to 0$ as $n \to \infty$, there exists $f_\pi^* \in \mathcal{F}_r$ such that $e_V(f_\pi^*, \bar{f}_\pi) \leq s_n$.*
2. *There exists $0 \leq \alpha < \infty$ and $a_1 > 0$ such that, for any sufficiently small $\delta > 0$,*

$$\sup_{\{f \in \mathcal{F} : e_V(f, \bar{f}_\pi) \leq \delta\}} \|sign(f) - sign(\bar{f}_\pi)\|_1 \leq a_1 \delta^\alpha,$$

3. *Considered feature space is uniformly bounded such that there exists $0 < G < \infty$ satisfying $\|\boldsymbol{X}\| \leq G$*

Then, for the estimator $\hat{p}$ obtained from our algorithm, there exists a constant a_2 such that

$$\mathbb{P}\left\{\|\hat{p} - p\|_1 \geq \frac{1}{2m} + \frac{1}{2}a_1(m+1)\delta_n^{2\alpha}\right\} \leq 15\exp\{-a_2 n(\lambda J_\pi^*)\},$$

provided that $\lambda^{-1} \geq \frac{rGJ_\pi^*}{2\delta_n^2}$ where $J_\pi^* = \max(J(f_\pi^*), 1)$ and $\delta_n = \max\left(\mathcal{O}\left(rG\exp\left(-\frac{\sqrt{n}}{rG}\right)^{2/3}\right), s_n\right)$.

Proof. We apply Theorem 3 in [2] to our case. First, notice that trucation on the loss function V is not needed by third assumption:

$$\|yf(\boldsymbol{X})\| = \|\langle B, \boldsymbol{X}\rangle\| \leq \|B\|\|\boldsymbol{X}\| \leq rGM,$$

which implies uniformly boundness of V. Let V be bounded by T. For the second equation of Assumption 2 in [2],

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

Therefore, β in [2] can be replaced by 1 from the following inequality.

$$\sup_{\{f \in \mathcal{F} : e_V(f, \bar{f}_\pi) \leq \delta\}} \text{var}\{V(f, \boldsymbol{X}, y) - V(\bar{f}_\pi, \boldsymbol{X}, y)\} \leq \sup_{\{f \in \mathcal{F} : e_V(f, \bar{f}_\pi) \leq \delta\}} Te_V(f, \bar{f}_\pi) \leq T\delta$$

Now we check Assumption 3 in [2]. Notice that

$$H_2(\epsilon, \mathcal{F}^V(k)) \le H_2(\epsilon, \mathcal{F}(k)) \le \log N_\infty(\epsilon, \mathcal{F}(k)), \tag{1}$$

because for functions f_ℓ and f_u, $\|V(f_\ell, \cdot) - V(f_u, \cdot)\|_2 \le \|f_\ell - f_u\|_2$. The last inequality in (1) is from Lemma 9.22 in [1]. From [3], we have

Gaussian design: entries of feature i.i.d Gaussian
X: lenght-d vector — > llXll_F ~ sqrt{d}
X: d-by-d matrix — > llXll_F ~ d, \lambda(X) = \sqrt{d}

$$\log N_\infty(\epsilon, \mathcal{F}(k)) \le \mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^2 \log\left(\frac{k}{\epsilon}\right)\right).$$

Fact: ll Xll_F <= \sqrt{r} lambda(X)

Therefore,

N { <B, X>: all B} —> E_X(N), where X ~ i.i.d. Gaussian

$$\phi(\epsilon, k) \approx \int_{\mathcal{O}(L)}^{\mathcal{O}(\sqrt{L})} \left(\frac{k}{\omega}\right) \sqrt{\log\left(\frac{k}{\omega}\right)} d\omega \approx \mathcal{O}\left(k \left(\log\left(\frac{k}{L}\right)\right)^{3/2}\right),$$

where $L = \min\{\epsilon^2 + \lambda(k/2-1)H_\pi^*, 1\}$. Solving Assumption 3 in [2] gives us $\epsilon_n^2 = \mathcal{O}\left(rG \exp\left(-\frac{\sqrt{n}}{rG}\right)^{2/3}\right)$ when $\epsilon_n^2 \ge \lambda r G J_\pi^*$. Plugging each variable into Theorem 3 proves the theorem. Notice that condition of λ is replaced because $\{\epsilon_n^2 \ge \lambda r G J_\pi^*\} \subset \{\epsilon_n^2 \ge 2\lambda J_\pi^*\}$ when $rG \ge 2$. □

Remark 1. In the proof, we use discrete version of the covering number based on ℓ observations defined by

$$N_\infty(\epsilon, \mathcal{F}, \ell) = \sup_{\{\boldsymbol{X}\}_{i=1}^\ell \subset \mathbb{R}^{d_1 \times d_2}} N_\infty(\epsilon, \mathcal{F}, \{\boldsymbol{X}\}_{i=1}^\ell).$$

We can check the following relationship in [4].

$$N_\infty(\mathcal{F}, \epsilon, \ell) \le N_\infty(\mathcal{F}, \epsilon) \le N_\infty\left(\mathcal{F}, \epsilon - \frac{cMG\sqrt{r}}{\ell^{1/d_1 d_2} - 1}, \ell\right), \text{ where } c > 0 \text{ is a constant} \tag{2}$$

(The above upper bound makes sense only for $\ell > (cMG\sqrt{r}/\epsilon + 1)^{d_1 d_2}$ i.e. for sufficiently large ℓ (2) holds). Therefore, the covering number $N_\infty(\epsilon, \mathcal{F})$ is almost equivalent to the covering number $N_\infty(\epsilon, \mathcal{F}, \ell)$.

2 Nonlinear case

Define

rank[B*t(B)] = r, and B*t(B) is d1-by-d1

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

where $h : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2'}$ and $\mathcal{B} = \{\sum_{i=1}^r \lambda_i u_i v_i^T : u_i \in \mathbb{R}^{d_1}, v_i \in \mathbb{R}^{d_2'}, \|u_i\| = \|v_i\| = 1, \text{ and } 0 < \lambda_r \le \cdots \le \lambda_1 \le M\}$. The reason of changing the set $\mathcal{B}$ from previous lecture note is that rank concept becomes ambiguous when $d_2' = \infty$. In this case, I am not sure that there is an unified way to calculate the covering number. Above all, we cannot use the covering number of linear class because Equation (2) becomes meaningless when extended feature space has infinity dimension. Therefore, the covering number of kernel case should be calculated individually.

References

[1] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

[2] Junhui Wang, Xiaotong Shen, and Yufeng Liu. Probability estimation for large-margin classifiers. *Biometrika*, 95(1):149–167, March 2008.

[3] Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

[4] Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739 – 767, 2002.