

# Consistency of probability estimator

Chanwoo Lee, July 16, 2020

**Lemma 1.** Let  $\mathcal{B}_r(k) = \{\mathbf{B} \in \mathbb{R}^{d_1 \times d_2} : \text{rank}(\mathbf{B}) \leq r, \|\mathbf{B}\|_F \leq k\}$ . Then  $N(\epsilon, \mathcal{B}_r(k), \|\cdot\|_F) \leq \mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^{r(d_1+d_2)}\right)$ .

*Proof.* Consider  $\mathbf{B} \in \mathcal{B}_r(k)$  in the form of  $\mathbf{B} = \mathbf{U}\mathbf{V}^T$  where  $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ ,  $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$  such that  $\|\mathbf{U}\|_F \leq \sqrt{k}$  and  $\|\mathbf{V}\|_F \leq \sqrt{k}$ . We construct set of  $\{\mathbf{U}_i\}$  and  $\{\mathbf{V}_j\}$  such that for any  $\mathbf{U}, \mathbf{V}$ , there exist  $i, j$  such that  $\|\mathbf{U} - \mathbf{U}_i\|_F \leq \epsilon/2\sqrt{k}$  and  $\|\mathbf{V} - \mathbf{V}_j\|_F \leq \epsilon/2\sqrt{k}$ . Then, epsilon balls with centers in  $\{\mathbf{U}\mathbf{V}^T : \mathbf{U} \in \{\mathbf{U}_i\}, \mathbf{V} \in \{\mathbf{V}_j\}\}$  can cover  $\mathcal{B}_r(k)$  because for any  $\mathbf{B} = \mathbf{U}\mathbf{V}^T \in \mathcal{B}_r(k)$ , we have  $\mathbf{U}_i\mathbf{V}_j^T \in \{\mathbf{U}\mathbf{V}^T : \mathbf{U} \in \{\mathbf{U}_i\}, \mathbf{V} \in \{\mathbf{V}_j\}\}$  such that

$$\begin{aligned} \|\mathbf{U}\mathbf{V}^T - \mathbf{U}_i\mathbf{V}_j^T\|_F &\leq \|\mathbf{U}\mathbf{V}^T - \mathbf{U}\mathbf{V}_j^T\|_F + \|\mathbf{U}\mathbf{V}_j^T - \mathbf{U}_i\mathbf{V}_j^T\|_F \\ &\leq \|\mathbf{U}\|_F \|\mathbf{V} - \mathbf{V}_j\|_F + \|\mathbf{V}_j\|_F \|\mathbf{U} - \mathbf{U}_i\|_F \\ &\leq \sqrt{k} \frac{\epsilon}{2\sqrt{k}} + \sqrt{k} \frac{\epsilon}{2\sqrt{k}} \leq \epsilon. \end{aligned}$$

Therefore, the covering number of  $N(\epsilon, \mathcal{B}_r(k), \|\cdot\|_F) \leq \mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^{r(d_1+d_2)}\right)$ , where  $\mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^{r(d_1)}\right)$  comes from  $\{\mathbf{U}_i\}$  and  $\mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^{r(d_2)}\right)$  from  $\{\mathbf{V}_j\}$ .  $\square$

**Remark 1.** This covering number bound is not the sharpest bound. There are several reasons for that. First, there are many representations of  $\mathbf{B} = \mathbf{U}\mathbf{V}^T$  i.e. the representation is not unique for given  $\mathbf{B}$ , which means there might be redundant centers in the set. In addition, when considered matrices are full rank ( $r = \min(d_1, d_2)$ ), this bound is slightly greater than the covering number bound of coefficient  $\mathcal{B}(k)$  only with norm constraint. However, the covering bound in Lemma 1 is small enough to show benefit of low rank structure.

**Proposition 1** (Thm 9.23 in [1]). Suppose the class of functions  $\mathcal{F} = \{f_t : t \in T\}$  satisfies,

$$|f_s(x) - f_t(x)| \leq d(s, t)F(x),$$

for some metric  $d$  on  $T$ , some real function  $F$  on the sample space  $\mathcal{X}$ . Then, for any norm  $\|\cdot\|$ ,

$$N_{[]} (2\epsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq N(\epsilon, T, d).$$

**Lemma 2.** Let  $\mathcal{F}_r(k) = \{f : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R} : f(\mathbf{X}) = \langle \mathbf{B}, \mathbf{X} \rangle \text{ for } \mathbf{B} \in \mathcal{B}_r(k)\}$  where  $\mathcal{B}_r(k) = \{\mathbf{B} \in \mathbb{R}^{d_1 \times d_2} : \text{rank}(\mathbf{B}) \leq r, \|\mathbf{B}\|_F \leq k\}$ . Suppose that there exists  $G > 0$  such that  $\sqrt{\mathbb{E}\|\mathbf{X}\|_F^2} \leq G$ . Then the bracketing number  $N_{[]}(\epsilon, \mathcal{F}_r^V(k), \|\cdot\|_2)$  is bounded by

$$\log N_{[]}(\epsilon, \mathcal{F}_r^V(k), \|\cdot\|_2) \leq \mathcal{O}\left(r(d_1 + d_2) \log\left(\frac{Gk}{\epsilon}\right)\right).$$

*Proof.* Let  $f_{\mathbf{B}}(\mathbf{X}) = \langle \mathbf{B}, \mathbf{X} \rangle$ . Notice that for any  $\mathbf{B}_1, \mathbf{B}_2 \in \mathcal{B}_r(k)$ ,

$$|f_{\mathbf{B}_1}(\mathbf{X}) - f_{\mathbf{B}_2}(\mathbf{X})| = |\langle \mathbf{B}_1 - \mathbf{B}_2, \mathbf{X} \rangle| \leq \|\mathbf{B}_1 - \mathbf{B}_2\|_F \|\mathbf{X}\|_F.$$

Applying Proposition 1 with  $F(\mathbf{X}) = \|\mathbf{X}\|_F$ ,  $d(\mathbf{B}_1, \mathbf{B}_2) = \|\mathbf{B}_1 - \mathbf{B}_2\|_F$  and  $\|\cdot\| = \|\cdot\|_2$ , we have

$$N_{[]}(\epsilon, \mathcal{F}_r(k), \|\cdot\|_2) \leq N\left(\frac{\epsilon}{2\|F\|_2}, \mathcal{B}_r(k), \|\cdot\|_F\right) \leq N\left(\frac{\epsilon}{2G}, \mathcal{B}_r(k), \|\cdot\|_F\right).$$

From Lemma 1, we have the covering number  $N(\epsilon, \mathcal{B}_r(k), \|\cdot\|_F) \leq \mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^{r(d_1+d_2)}\right)$ . Note that, for functions  $f_\ell$  and  $f_u$ ,

$$\|V^T(f_\ell, \cdot) - V^T(f_u, \cdot)\|_2^2 \leq \|f_\ell - f_u\|_2^2$$

implying that  $N_{[]}(\epsilon, \mathcal{F}^V(k), \|\cdot\|_2) \leq N_{[]}(\epsilon, \mathcal{F}(k), \|\cdot\|_2) \leq N\left(\frac{\epsilon}{2G}, \mathcal{B}_r(k), \|\cdot\|_F\right) \leq \mathcal{O}\left(r(d_1 + d_2) \log\left(\frac{Gk}{\epsilon}\right)\right)$ .  $\square$

**Lemma 3.** Let  $k > 0$  be a given constant. If  $\frac{1}{Ke} > L > 0$ , we have

$$\int_{\mathcal{O}(L)}^{\mathcal{O}(L^{\beta/2})} \sqrt{\log\left(\frac{k}{\omega}\right)} d\omega \leq \mathcal{O}\left(\sqrt{L^\beta \log\left(\frac{k}{\sqrt{L}^\beta}\right)}\right),$$

where  $\beta = 1 \wedge \alpha$  and  $\alpha > 0$ .

*Proof.*

$$\begin{aligned} \int_{\mathcal{O}(L)}^{\mathcal{O}(L^{\beta/2})} \sqrt{\log\left(\frac{k}{\omega}\right)} - \frac{1}{2\sqrt{\log\left(\frac{k}{\omega}\right)}} d\omega &= k \left[ \omega \sqrt{\log\left(\frac{1}{\omega}\right)} \right]_{\mathcal{O}(L/k)}^{\mathcal{O}(L^{\beta/2}/k)} \\ &= \mathcal{O}\left(\sqrt{L^\beta \log\left(\frac{k}{\sqrt{L}^\beta}\right)}\right) \end{aligned} \quad (1)$$

The first equality in (1) is from changing variable. Notice that

$$\int_{\mathcal{O}(L)}^{\mathcal{O}(L^{\beta/2})} \sqrt{\log\left(\frac{k}{\omega}\right)} - \frac{1}{2\sqrt{\log\left(\frac{k}{\omega}\right)}} d\omega \geq \int_{\mathcal{O}(L)}^{\mathcal{O}(L^{\beta/2})} \sqrt{\log\left(\frac{k}{\omega}\right)} - \mathcal{O}(1) d\omega, \quad (2)$$

from the condition on  $L$ . Combining Equation (1) and Equation (2) completes the proof.  $\square$

**Lemma 4.**  $\sqrt{\frac{d}{L^{2-\beta}} \log\left(\frac{k}{\sqrt{L}^\beta}\right)} \leq \sqrt{n}$  holds if  $L \leq \frac{\log(n/d)^{1/(2-\beta)} + 2\log(k)}{(n/d)^{1/(2-\beta)}}$  where  $\beta \leq 1$ .

*Proof.* Suppose  $L \leq \frac{\log(n/d)^{1/(2-\beta)} + 2\log(k)}{(n/d)^{1/(2-\beta)}}$ . By plugging in, we have

$$\begin{aligned} \sqrt{\frac{d}{L^{2-\beta}} \log\left(\frac{k}{\sqrt{L}^\beta}\right)} &\leq \sqrt{\frac{n}{(2-\beta) \log((n/d)^{1/(2-\beta)} k^2)} \left( \frac{\log((n/d)^{1/(2-\beta)} k^2) - \log \log((n/d)^{1/(2-\beta)} k^2)}{2} \right)} \\ &\leq \sqrt{n}. \end{aligned}$$

$\square$

**Theorem 0.1.** Assume that

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

A.2 There exist constant  $0 \leq \alpha < \infty$ ,  $a_1 > 0$  such that, for any sufficiently small  $\delta > 0$ .

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

A.3 Considered feature space is uniformly bounded such that there exists  $0 < G < \infty$  satisfying  $\sqrt{\mathbb{E}\|\mathbf{X}\|_F^2} \leq G$

Then, for the estimator  $\hat{p}$  obtained from our algorithm with function class  $\mathcal{F}_r(M)$ , there exists a constant  $a_2$  such that

$$\mathbb{P}\left\{\|\hat{p} - p\|_1 \geq \frac{1}{2m} + \frac{a_1}{2}(m+1)\delta_n^{2\alpha}\right\} \leq 15 \exp\{-a_2 n(\lambda J_\pi^*)^{2-\alpha \wedge 1}\}, \quad (3)$$

provided that  $\lambda^{-1} \geq \max\left(\frac{MJ_\pi^*}{2\delta_n^2}, \frac{4J_\pi^*}{\delta_n^2}\right)$  where  $J_\pi^* = \max(J(f_\pi^*), 1)$ ,  $\delta_n = \max\left(\mathcal{O}\left(\frac{\log(n/r(d_1+d_2))^{1/(2-\beta)} + 2\log(GM)}{(n/r(d_1+d_2))^{1/(2-\beta)}}\right), s_n\right)$  and  $\beta = \alpha \wedge 1$ .

**Remark 2.** From the definition of  $\delta_n$ , we can check the larger  $\alpha$  we obtain, the sharper bound we can have. In addition, the sharpest  $\delta_n$  we can obtain is  $\mathcal{O}\left(\frac{\log n}{n}\right)$  when  $\alpha \geq 1$ .

*Proof.* We apply Theorem 3 in [2] to our case.

The second condition of the assumption is

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

Notice that

$$\begin{aligned} \text{var}\{V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)\} &\leq \mathbb{E}|V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)|^2 \\ &\leq T \mathbb{E}|V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)| \\ &= T(\lambda_1 + \lambda_2). \end{aligned}$$

where

$$\begin{aligned} \lambda_1 &= \mathbb{E}|S(y)(1 - \text{sign}(yf(\mathbf{X})) - V(\bar{f}_\pi, \mathbf{X}, y)| = \mathbb{E}|S(y)| |\text{sign}(f) - \text{sign}(\bar{f}_\pi)| \\ &\leq \|\text{sign}(f) - \text{sign}(\bar{f}_\pi)\|_1 \leq a_1 \delta^\alpha \quad \text{from A.2.} \end{aligned}$$

and

$$\begin{aligned} \lambda_2 &= \mathbb{E}[V^T(f, \mathbf{X}, y) - S(y)(1 - \text{sign}(yf(\mathbf{X})))] \\ &\leq e_{VT}(f, \bar{f}_\pi) + \mathbb{E}\{V(\bar{f}_\pi, \mathbf{X}, y) - S(y)(1 - \text{sign}(yf(\mathbf{X})))\} \\ \text{why 2 not 1?} &\leq 2e_{VT}(f, \bar{f}_\pi) \leq 2\delta \end{aligned}$$

Isn't it always  $\leq 0$ ?

b.c. expected hinge loss (Bayes  $\bar{f}$ ) = expected 0-1 loss (Bayes  $\bar{f}$ )  $\Leftarrow$  expected 0-1 loss (arbitrary  $\bar{f}$ )

Therefore,  $\beta$  in [2] can be replaced by  $1 \wedge \alpha$ .

Now we check Assumption 3 in [2]. From Lemma 2, we have

$$H_B(\epsilon, \mathcal{F}^V(k)) \leq \mathcal{O} \left( r(d_1 + d_2) \log \left( \frac{Gk}{\epsilon} \right) \right).$$

Therefore, we have the following equation from Lemma 3.

$$\phi(\epsilon, k) \approx \int_{\mathcal{O}(L)}^{\mathcal{O}(L^{\beta/2})} \sqrt{r(d_1 + d_2) \log \left( \frac{kG}{\omega} \right)} d\omega / L \lesssim \mathcal{O} \left( \sqrt{r(d_1 + d_2)} \left( \log \left( \frac{kG}{\sqrt{L^\beta}} \right) / L^{2-\beta} \right)^{1/2} \right),$$

where  $L = \min\{\epsilon^2 + \lambda(k/2 - 1)J_\pi^*, 1\}$ . Solving Assumption 3 in [2] gives us  $\epsilon_n^2 = \mathcal{O} \left( \frac{\log(n/r(d_1 + d_2))^{1/(2-\beta)} + 2\log(GM)}{(n/r(d_1 + d_2))^{1/(2-\beta)}} \right)$  by Lemma 4 when  $\epsilon_n^2 \geq \frac{\lambda M J_\pi^*}{2}$ . Plugging each variable into Theorem 3 proves the theorem.  $\square$

**Remark 3.** We show that the Assumption 2 is satisfied when there exists  $\eta > 0$  such that  $|\mathbb{P}(y = 1|\mathbf{X}) - \pi| \geq \eta$  almost surely with respect to distribution  $\mathbf{X}$ . Smooth parameter is  $a_1 = \frac{1}{\eta}$  and  $\alpha = 1$  in this case.

*Proof.*

$$\begin{aligned} e_{VT}(f, \bar{f}_\pi) &= \mathbb{E} [S(y)L(yf(\mathbf{X})) \wedge T - S(y)L(y\bar{f}_\pi(\mathbf{X}))] \\ &\geq \mathbb{E} [S(y)(1 - \text{sign}(yf(\mathbf{X}))) - S(y)(1 - \text{sign}(y\bar{f}_\pi(\mathbf{X})))] \\ &= \mathbb{E} [yS(y) (\text{sign}(\bar{f}_\pi) - \text{sign}(f))] \\ &= \mathbb{E} [\mathbb{E}(yS(y)|\mathbf{X}) (\text{sign}(\bar{f}_\pi) - \text{sign}(f))] \\ &= \mathbb{E} [|\mathbb{P}(y = 1|\mathbf{X}) - \pi| |\text{sign}(\bar{f}_\pi) - \text{sign}(f)|] \\ &\geq \eta \mathbb{E} |\text{sign}(\bar{f}_\pi) - \text{sign}(f)| = \eta \|\text{sign}(\bar{f}_\pi) - \text{sign}(f)\|_1. \end{aligned}$$

$\square$

The main part of the proof is the following inequality

$$\mathbb{E} [|f_\pi| |\text{sign}(f) - \text{sign}(\bar{f}_\pi)|] \geq \eta \mathbb{E} [|\text{sign}(f) - \text{sign}(\bar{f}_\pi)|]. \quad (4)$$

Therefore, we can replace the condition by

$$\text{For a given } \pi, \text{ there exists } \eta > 0 \text{ such that } \mathbb{E} [|f_\pi| \mathbf{1}_{\{\text{sign}(f) \neq \text{sign}(\bar{f}_\pi)\}}] \geq \eta \mathbb{E} [\mathbf{1}_{\{\text{sign}(f) \neq \text{sign}(\bar{f}_\pi)\}}].$$

**Example 1.** When ground truth  $p(\mathbf{X})$  is step function such that  $p(\mathbf{X}) = \sum_{k=1}^K c_k \mathbf{1}_{\{\mathbf{X} \in A_k\}}$ , then  $\eta = \min_k \{|c_k - \pi|\}$ .

**Example 2.** Assume that ground truth  $p(x) = x$  and  $x$  is a random variable from  $\text{Unif}(0, 1)$ . If considered function class is a set of functions with only one sign change, we show Assumption 2 holds with  $\alpha = 1/2, a_1 = 2$ .

*Proof.* We check two terms of Equation (4) in this case. If  $f^{-1}(0) = \pi$ , then the conclusion trivially holds. So we consider  $f^{-1}(0) \neq \pi$  case. Let  $f^{-1}(0) = \pi'$ . Notice that right side of Equation (4) is

$$\mathbb{E}|\text{sign}(f) - \text{sign}(\bar{f}_\pi)| = 2|\pi - \pi'|.$$

The left side of the equation is

$$\mathbb{E} [|f_\pi| |\text{sign}(f) - \text{sign}(\bar{f}_\pi)|] = \mathbb{E} [2|x - \pi| \mathbb{1}_{\{\pi \wedge \pi' < x < \pi \vee \pi'\}}] = \int_{\pi \wedge \pi'}^{\pi \vee \pi'} 2|x - \pi| dx = |\pi - \pi'|^2.$$

Therefore,  $e_{VT}(f, \bar{f}_\pi) \geq \mathbb{E} [|f_\pi| |\text{sign}(f) - \text{sign}(\bar{f}_\pi)|] = \frac{1}{4} (\mathbb{E} [|\text{sign}(f) - \text{sign}(\bar{f}_\pi)|])^2$ , which implies  $\alpha = 1/2, a_1 = 2$

□

**Remark 4.** In Example 1, the order of ground truth function is 0 and we obtain the smooth parameter  $\alpha = 1$ . In Example 2, the order of ground truth function is 1 and we have the smooth parameter  $\alpha = \frac{1}{2}$ . We can conjecture that the smooth parameter  $\alpha = \frac{1}{\text{order}(\bar{f}_\pi)+1}$  because if we consider each term of the condition (4), the left side is calculated as

$$L \stackrel{\text{def}}{=} \mathbb{E} [|f_\pi| \mathbb{1}_{\{\text{sign}(f) \neq \text{sign}(\bar{f}_\pi)\}}] = \int_{\{\text{sign}(f) \neq \text{sign}(\bar{f}_\pi)\}} |f_\pi| dF(x)$$

where  $F(x)$  is distribution of  $x$ . The right side is

$$R \stackrel{\text{def}}{=} \mathbb{E} [\mathbb{1}_{\{\text{sign}(f) \neq \text{sign}(\bar{f}_\pi)\}}] = \int_{\{\text{sign}(f) \neq \text{sign}(\bar{f}_\pi)\}} 1 dF(x)$$

If we consider the simple case where  $\{\text{sign}(f) \neq \text{sign}(\bar{f}_\pi)\}$  is an interval, we can easily see that  $L = \mathcal{O}((R)^{\text{order}(\bar{f}_\pi)+1})$  which explains the conjecture. Therefore, Assumption 2 consider features of ground truth probability.

**Remark 5.** *A.1 measures accuracy of approximation to the ground truth function from considered function class  $\mathcal{F}_r$ . A.2 considers the complexity of ground truth function as in Remark 4. A.3 is related to the covering number which measures the complexity of considered function class  $\mathcal{F}_r$ .*

**Remark 6.** Our theorem shows that a higher sample complexity is needed when the ground truth function has a high level of complexity or the candidate function class is either too small or too large. This reflect the trade off between A.1 and A.3.

**Remark 7.** We can think of our estimation method consisting of two parts.

### S.1 Approximation of the target probability function

$$\left\| p(\mathbf{X}) - \sum_{i=1}^m \frac{1}{m} \mathbb{1}_{\{\mathbf{X}: p(\mathbf{X}) < \frac{i}{m}\}} \right\|_1. \quad (5)$$

### S.2 For each $i$ , Estimation of sublevel set

$$\left\| \sum_{i=1}^m \frac{1}{m} \mathbb{1}_{\{\mathbf{X}: p(\mathbf{X}) < \frac{i}{m}\}} - \sum_{i=1}^m \frac{1}{m} \mathbb{1}_{\{\mathbf{X}: \text{sign}[\hat{f}_{\pi_i}(\mathbf{X})] = -1\}} \right\|_1. \quad (6)$$

Those estimation procedures are reflected in Theorem 0.1. In the first step, the maximum error of the approximation is  $\frac{1}{2m}$  at given  $\mathbf{X}$ . Therefore, we have the bound  $\frac{1}{2m}$  for Equation (5). In the second step, two functions are  $m$ -step functions. Let  $f(\mathbf{X}) = \sum_{i=1}^m \frac{1}{m} \mathbb{1}_{\{\mathbf{X}: p(\mathbf{X}) < \frac{i}{m}\}}$  and  $g(\mathbf{X}) = \sum_{i=1}^m \frac{1}{m} \mathbb{1}_{\{\mathbf{X}: \text{sign}[\hat{f}_{\pi_i}(\mathbf{X})] = -1\}}$ . Define  $A_i = \{\mathbf{X} : f(\mathbf{X}) = \frac{i}{m}\} - \{\mathbf{X} : g(\mathbf{X}) = \frac{i}{m}\}$ . Then total measure at which  $f$  and  $g$  disagree is at most  $m \max_i \mathbb{P}(A_i)$ . Therefore, we have bound  $m \max_i \mathbb{P}(A_i)$  bound for Equation (6). This shows why we have two terms  $\frac{1}{2m}$  and  $\frac{a_1}{2}(m+1)\delta_n^{2\alpha}$  in Equation (3).

## 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.