Consistency of probability estimator

Chanwoo Lee, July 12, 2020

Lemma 1. *Let* $\mathcal{B}_r(k) = \{\boldsymbol{B} \in \mathbb{R}^{d_1\times d_2} : rank(B) \leq r, \|\boldsymbol{B}\|_F \leq k\}$. *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 $\boldsymbol{B} \in \mathcal{B}_r(k)$ in the form of $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^T$ where $\boldsymbol{U} \in \mathbb{R}^{d_1\times r}, \boldsymbol{V} \in \mathbb{R}^{d_2\times r}$ such that $\|\boldsymbol{U}\|_F \leq \sqrt{k}$ and $\|\boldsymbol{V}\|_F \leq \sqrt{k}$. We construct set of $\{U_i\}$ and $\{V_j\}$ such that for any $\boldsymbol{U}, \boldsymbol{V}$, there exist i, j such that $\|\boldsymbol{U} - \boldsymbol{U}_i\|_F \leq \epsilon/2\sqrt{k}$ and $\|\boldsymbol{V} - \boldsymbol{V}_j\|_F \leq \epsilon/2\sqrt{k}$. Then, epsilon balls with centers in $\{\boldsymbol{U}\boldsymbol{V}^T : \boldsymbol{U} \in \{\boldsymbol{U}_i\}, \boldsymbol{V} \in \{\boldsymbol{V}_j\}\}$ can cover $\mathcal{B}_r(k)$ because for any $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^T \in \mathcal{B}_r(k)$, we have $\boldsymbol{U}_i\boldsymbol{V}_j^T \in \{\boldsymbol{U}\boldsymbol{V}^T : \boldsymbol{U} \in \{\boldsymbol{U}_i\}, \boldsymbol{V} \in \{\boldsymbol{V}_j\}\}$ such that

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

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 $\{\boldsymbol{U}_i\}$ and $\mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^{r(d_2)}\right)$ from $\{\boldsymbol{V}_j\}$. □

Remark 1. This covering number bound is not the sharpest bound. There are several reasons for that. First, there are many representations of $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^T$ i.e. the representation is not unique for given $\boldsymbol{B}$, 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* $\mathcal{F}$ *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* $\mathcal{F}_r(k) = \{f : \mathbb{R}^{d_1\times d_2} \to \mathbb{R} : f(\boldsymbol{X}) = \langle \boldsymbol{B}, \boldsymbol{X}\rangle$ *for* $\boldsymbol{B} \in \mathcal{F}_r(k)\}$ *where* $\mathcal{B}_r(k) = \{\boldsymbol{B} \in \mathbb{R}^{d_1\times d_2} : rank(B) \leq r, \|\boldsymbol{B}\|_F \leq k\}$*. Suppose that there exists* $G > 0$ *such that* $\sqrt{\mathbb{E}\|\boldsymbol{X}\|_F^2} \leq G$*. 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 $f_{\boldsymbol{B}}(\boldsymbol{X}) = \langle \boldsymbol{B}, \boldsymbol{X}\rangle$. Notice that for any $\boldsymbol{B}_1, \boldsymbol{B}_2 \in \mathcal{B}_r(k)$,

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

Applying Proposition 1 with $F(\boldsymbol{X}) = \|\boldsymbol{X}\|_F$, $d(\boldsymbol{B}_1, \boldsymbol{B}_2) = \|\boldsymbol{B}_1 - \boldsymbol{B}_2\|_F$ 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_ℓ and f_u,

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

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

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

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

Proof.

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

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

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

from the condition on L. Combining Equation (1) and Equation (2) completes the proof. □

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

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

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

□

Theorem 0.1. *Assume that*

A.1 For some positive sequence such that $s_n \to 0$ as $n \to \infty$, 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 $0 \le \alpha < \infty$, $a_1 > 0$ such that, for any sufficiently small $\delta > 0$.

$$\sup_{\{f\in\mathcal{F}:e_{V^T}(f,\bar f_\pi)\le\delta\}} \|sign(f) - sign(\bar f_\pi)\|_1 \le a_1\delta^{\alpha}.$$

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

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

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

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

The second condition of the assumption is

$$\sup_{\{f\in\mathcal{F}:e_{V^T}(f,\bar f_\pi)\le\delta\}} \operatorname{var}\{V(f,\boldsymbol{X},y) - V(\bar f_\pi,\boldsymbol{X},y) \le a_2\delta^{\beta}\}.$$

Notice that

$$\begin{aligned}
\operatorname{var}\{V^T(f,\boldsymbol{X},y) - V(\bar f_\pi,\boldsymbol{X},y)\} &\le \mathbb{E}|V(^T f,\boldsymbol{X},y) - V(\bar f_\pi,\boldsymbol{X},y)|^2\\
&\le T\mathbb{E}|V^T(f,\boldsymbol{X},y) - V(\bar f_\pi,\boldsymbol{X},y)|\\
&= T\left(\lambda_1+\lambda_2\right).
\end{aligned}$$

where

$$\begin{aligned}
\lambda_1 &= \mathbb{E}\left|S(y)(1-\operatorname{sign}(yf(\boldsymbol{X})) - V(\bar f_\pi,\boldsymbol{X},y)\right| = \mathbb{E}|S(y)||\operatorname{sign}(f) - \operatorname{sign}(\bar f_\pi)|\\
&\le \|\operatorname{sign}(f) - \operatorname{sign}(\bar f_\pi)|\|_1 \le a_1\delta^{\alpha} \quad \text{ from } \textit{A.2}.
\end{aligned}$$

and

$$\begin{aligned}
\lambda_2 &= \mathbb{E}\left[V^T(f,\boldsymbol{X},y) - S(y)(1-\operatorname{sign}(yf(\boldsymbol{X}))\right]\\
&\le e_{V^T}(f,\bar f_\pi) + \mathbb{E}\left\{V(\bar f_\pi,\boldsymbol{X},y) - S(y)(1-\operatorname{sign}(yf(\boldsymbol{X}))\right\}\\
&\le 2e_{V^T}(f,\bar f_\pi) \le 2\delta
\end{aligned}$$

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

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

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

Therefore, we have the following equation from Lemma 3.

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

where $L = \min\{\epsilon^2+\lambda(k/2-1)H_\pi^*, 1\}$. Solving Assumption 3 in [2] gives us $\epsilon_n^2 = \mathcal{O}\left(\frac{\log(n/r(d_1+d_2))+2\log(GM)}{n/r(d_1+d_2)}\right)$ by Lemma 4 when $\epsilon_n^2 \ge \lambda G J_\pi^*$. Plugging each variable into Theorem 3 proves the theorem. Notice that condition of λ is replaced because $\{\epsilon_n^2 \ge \lambda G J_\pi^*\} \subset \{\epsilon_n^2 \ge 2\lambda J_\pi^*\}$ when $rG \ge 2$. $\square$

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

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

Proof.

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

□

The main part of the proof is the following inequality

$$\mathbb{E}\left[|f_\pi||\operatorname{sign}(f) - \operatorname{sign}(\bar{f}_\pi)|\right] \geq \eta\mathbb{E}\left[|\operatorname{sign}(f) - \operatorname{sign}(\bar{f}_\pi)|\right]. \tag{4}$$

Therefore, we can replace the condition by

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

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

Example 2. Assume that ground truth $p(x) = x$ and x is a random variable from 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

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

The left side of the equation is

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

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

□

Remark 3. 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 $\alpha = \frac{1}{\text{order}(f_\pi)+1}$ because if we consider each term of the condition (4), the left side is calculated as

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

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

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

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

Remark 4. *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 3. A.3 is related to the covering number which measures the complexity of considered function class $\mathcal{F}_r$.*

Remark 5. 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 6. We can think of our estimation method consisting of two parts.

S.1 Approximation of the target probability function

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

S.2 For each i, Estimation of sublevel set

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

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 $\boldsymbol{X}$. Therefore, we have the bound $\frac{1}{2m}$ for Equation (5). In the second step, two functions are m-step functions. Let $f(\boldsymbol{X}) = \sum_{i=1}^{m} \frac{1}{m} \mathbb{1}_{\{\boldsymbol{X}:\leq p(\boldsymbol{X})<\frac{i}{m}\}}$ and $g(\boldsymbol{X}) = \sum_{i=1}^{m} \frac{1}{m} \mathbb{1}_{\{\boldsymbol{X}:\text{sign}\left[\hat{f}_{\pi_i}(\boldsymbol{X})\right]=-1\}}$. Define $A_i = \{\boldsymbol{X} : f(\boldsymbol{X}) = \frac{i}{m}\} - \{\boldsymbol{X} : g(\boldsymbol{X}) = \frac{i}{m}\}$. 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 $frac12m$ 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.