Assumption modification

Chanwoo Lee, November 27, 2020

1 New assumption

(A) [Boundary noise] There exist constants $\alpha \in (0,1]$ and $C > 0$, such that

$$\max_{\pi \in \Pi'} \mathbb{P}_{\boldsymbol{X}} \left(|p(\boldsymbol{X}) - \pi| \le t/H \right) \le C \left(\frac{t}{H} \right)^{\frac{\alpha}{1-\alpha}}, \quad \text{for all } t \in [0,1] \text{ and all } H \in \mathbb{N}_+, \tag{1}$$

where $\Pi' \subset \Pi$ with $|\Pi| - |\Pi'| \le m$ and m is a finite number independent on H. When $\alpha = 1$, the inequality (1) reads $\mathbb{P}(|p(\boldsymbol{X}) - \pi| \le t/H) = 0$.

Theorem 1.1. The condition (1) is equivalent to

$$\mathbb{P}_{\boldsymbol{X}}[S \Delta S_{\text{bayes}}] \le \begin{cases} C_1 [R_\pi(S) - R_\pi(S_{\text{bayes}}(\pi))]^\alpha, & \text{if } \alpha \in (0,1), \\ C_2 H [R_\pi(S) - R_\pi(S_{\text{bayes}}(\pi)] & \text{if } \alpha = 1, \end{cases} \tag{2}$$

for all sets $S \in \mathbb{R}^{d_1 \times d_2}$, levels $\pi \in \Pi'$, and $H = |\Pi| \in \mathbb{N}_+$.

Remark 1. What are changed?

I excluded the worst case scenario of π using new set Π'. For example, consider the constant probability, ability case $p(\boldsymbol{X}) = p$. Suppose that $\pi_i = \frac{i}{H} \le p(\boldsymbol{x}) < \frac{i+1}{H} = \pi_{i+1}$ for some $i \in [H]$. In previous assumption, we cannot guarantee that $\mathbb{P}_{\boldsymbol{X}}(|p(\boldsymbol{X}) - \pi_i| \le t/H) = 0$ and $\mathbb{P}_{\boldsymbol{X}}(|p(\boldsymbol{X}) - \pi_{i+1}| \le t/H) = 0$ for all $t \in [0,1]$ because we can set p i/H or arbitrarily close to π_i or π_{i+1}. Simply excluding this two π_i and π_{i+1} from Π setting $\Pi' = \Pi \setminus \{\pi_i, \pi_{i+1}\}$ can solve the problem without changing convergence rate because finite number of exclusion is negligible.

If we exclude the worst case $\pi \in \Pi$ by adopting Π', then I start to believe that (1) is enough to characterize the most of probability function we are interested in. Theorem 1.1 shows that only when $\alpha = 1$, we need extra H to bound $\mathbb{P}(S \Delta S_{\text{bayes}})$. I think this happens because (1) has the different trend when $\alpha = 1$ by the term $\alpha/(1-\alpha)$.

Remark 2. Can this assumption cover logistic link case?

Yes, consider the case $p(\boldsymbol{X}) = e^x/(1+e^x)$ where $\boldsymbol{X} \sim N(0,1)$. Let $\Pi = \left\{\frac{i}{H}\right\}_{i=1}^{H-1}$ and $\Pi' = \left\{\frac{i}{H}\right\}_{i=2}^{H-2}$. The worst case in Π' is when $\pi = \frac{H-2}{H}$, In this case where $t = 1$, we have

$$\begin{aligned}
\mathbb{P}\left(\frac{H-3}{H} \le p(\boldsymbol{X}) \le \frac{H-1}{H} \right) &= \mathbb{P}\left(\log\left(\frac{H-3}{2}\right) \le \boldsymbol{X} \le \log(H-1) \right) \\
&= \int_{\log\left(\frac{H-3}{2}\right)}^{\log(H-1)} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&\le \frac{1}{\sqrt{2\pi}} e^{-\frac{\left[\log\left(\frac{H-3}{2}\right)\right]^2}{2}} \underbrace{\left(\log\left(\frac{H-3}{2}\right) - \log(H-1) \right)}_{(*)} \\
&\le \frac{1}{\sqrt{2\pi}} e^{-\frac{\log\left(\frac{H-3}{2}\right)}{2}} \log\left(\frac{2(H-1)}{H-3} \right) \\
&\le \frac{\log 3}{\sqrt{2\pi}} \sqrt{\frac{2}{H-3}} \\
&\le C \frac{1}{\sqrt{H}}.
\end{aligned}$$

Therefore, we have $\alpha = 1/3$. If we include $\pi = \frac{H-1}{H}$, the $(*)$ part cannot be calculated well. This might leads to use extra H terms but haven't checked yet. Therefore, introduction of new set Π' makes things easier.

Remark 3. Why do we need t/H instead of simple t?

If we do not use t/H for bounding $|p(\boldsymbol{X})-\pi| \le t/H$, we cannot make it clear that $|\Pi'|$ is finitely different from $|\Pi|$ independently on H. By adopting t/H instead of t, we can successfully exclude the finite number of the worst $\pi \in \Pi$ which is closest to concentrated mass of $p(\boldsymbol{X})$. For example, consider $p(\boldsymbol{X}) = \sum_{i=1}^{M} \frac{i}{M} \mathbb{1}_{\boldsymbol{X} \in G_i}$, where $\{G_i\}_{i=1}^{M}$ are partition of $\mathbb{R}^{d_1 \times d_2}$. Then the the number of π's, that are excluded from Π' is at most $2M$, which is negligible because $\frac{2M}{H} \to 0$ as $H \to \infty$.

Remark 4. What kinds of probability function has $\alpha \in (1/2, 1)$?

For ease of notation, I restate (1) with $\beta \in (1/2, 1)$ as

$$\max_{\pi \in \Pi'} \mathbb{P}_{\boldsymbol{X}} \left(|p(\boldsymbol{X}) - \pi| \le t/H \right) \le C \left(\frac{t}{H} \right)^{\beta}, \quad \text{for all } t \in [0,1] \text{ and all } H \in \mathbb{N}_+.$$

Assume that $\boldsymbol{X}$ is from Unif$[0,1]$ for easy calculation and $p(\boldsymbol{X})$ is monotonically increasing or decreasing for sufficiently small $1/H$-neighborhood around $\pi \in \Pi'$. Notice that

$$\frac{\mathbb{P}(\pi - t/H \le p(\boldsymbol{X}) \le \pi + t/H)}{2t/H} = \frac{|p^{-1}(\pi + t/H) - p^{-1}(\pi - t/H)|}{2t/H} \le C(t/H)^{\beta - 1}.$$

Since $\beta > 1$, the derivative of p^{-1} is 0 at all $\pi \in \Pi'$. In other words, $p(\boldsymbol{X})$ is not differentiable at most points in Π. Since (1) holds for arbitrary $H \in \mathbb{N}_+$, we can conclude that $p(\boldsymbol{X})$ is not differentiable at most rational points.

We can find such function $p(\boldsymbol{X})$ because there is a function which is continuous everywhere but differentiable nowhere such as Weierstrass function defined by

$$f(x) = \sum_{n=0}^{\infty} \frac{1}{2^n} \sin\left(2^n x\right).$$

The following question is "Can we have probability estimation whose convergence rate is greater than $\mathcal{O}(1/\sqrt{n})$?"

My answer is no. Notice we assume that our function class must have a sequence of functions $f_n^* \in \mathcal{F}(r, s_1, s_2)$ for which the surrogate excess risk vanishes; i.e., $R_{\ell,\pi}(f_n^*) - R_{\ell,\pi}(f_{\text{bayes}}) \le a_n \to 0$ for all $\pi \in \Pi'$ and for some vanishing sequence $a_n \to 0$ as $n, d \to \infty$ where $f_{\text{bayes}} = \text{sign}(p(\boldsymbol{X}) - \pi)$. However, if $p(\boldsymbol{X})$ has $\beta > 1$, I think that our linear classifier cannot have good approximation of such complicated function so that the this assumption cannot hold in the case $\beta > 1$. The argument to find such linear classifier in the main draft is based on monotonicity of the link function on $p(\boldsymbol{X})$.

Therefore, we can conclude that for one π, we have good classifier that has really fast convergence rate. However, for the probability estimation, we need to control whole elements of Π which result in stricter condition to be satisfied.

Proof. For ease of notation, we drop the argument π from $S_{\text{bayes}}(\pi)$ and $R_\pi(\cdot)$, and simply write S_{bayes} and $R(\cdot)$, respectively. The following identity is useful to relate the excess risk and set difference in classifiers,

$$\begin{aligned} R(S) - R(S_{\text{bayes}}) &= \mathbb{E}_{\boldsymbol{X},y}\left[w(y) \mathbb{1}(y \ne I(S)) \right] - \mathbb{E}_{\boldsymbol{X},y}\left[w(y) \mathbb{1}(y \ne I(S_{\text{bayes}})) \right] \\ &= \mathbb{E}_{\boldsymbol{X}}\left[(\pi - p(\boldsymbol{X})) \left(I(S) - I(S_{\text{bayes}}) \right) \right] \\ &= 2 \int_{\boldsymbol{X} \in S \Delta S_{\text{bayes}}} |p(\boldsymbol{X}) - \pi| d\mathbb{P}_{\boldsymbol{X}}. \end{aligned} \tag{3}$$

Now we will use the identity (3) to show the equivalence between (1) and (2). We divide the proof into two cases: $\alpha \in (0,1)$ and $\alpha = 1$.

Case 1: $\alpha \in (0,1)$.

(1) $\Rightarrow$ (2). Consider an arbitrary set $S \subset \mathbb{R}^{d_1 \times d_2}$. Let t be an arbitrary number in the interval $[0,1]$, and define the set $A = \{\boldsymbol{X} \colon |p(\boldsymbol{X}) - \pi| > t/H\}$. Based on the inequality (A2),

$$\begin{aligned}\int_{\boldsymbol{X} \in S\Delta S_{\text{bayes}}} |p(\boldsymbol{X}) - \pi| d\mathbb{P}_{\boldsymbol{X}} &\geq \frac{t}{H} \left[\mathbb{P}_{\boldsymbol{X}}((S\Delta S_{\text{bayes}}) \cap A)\right] \\ &\geq \frac{t}{H} \left(\mathbb{P}_{\boldsymbol{X}} \left(S\Delta S_{\text{bayes}}\right) - \mathbb{P}_{\boldsymbol{X}}(A^c)\right) \\ &\geq \frac{t}{H} \left(\mathbb{P}_{\boldsymbol{X}} \left(S\Delta S_{\text{bayes}}\right) - C\left(\frac{t}{H}\right)^{\frac{\alpha}{1-\alpha}}\right), \quad \text{for all } t \in [0,1].\end{aligned}$$

Combining the above inequality with the identity (3) yields

$$R(S) - R(S_{\text{bayes}}) \geq \frac{2t}{H} \left(\mathbb{P}_{\boldsymbol{X}} \left(S\Delta S_{\text{bayes}}\right) - C\left(\frac{t}{H}\right)^{\frac{\alpha}{1-\alpha}}\right), \quad \text{for all } t \in [0,1]. \tag{4}$$

We maximize the lower bound of (4) with respect to t and obtain the optimal $t_{\text{opt}} \in [0,1]$,

$$t_{\text{opt}} = \begin{cases} 1, & \text{if} \quad \mathbb{P}_{\boldsymbol{X}} \left(S\Delta S_{\text{bayes}}\right) \geq \frac{C}{1-\alpha} H^{\frac{\alpha}{1-\alpha}}, \\ \left[\frac{1-\alpha}{CH^{\frac{\alpha}{1-\alpha}}} \mathbb{P}_{\boldsymbol{X}} \left(S\Delta S_{\text{bayes}}\right)\right]^{\frac{1-\alpha}{\alpha}}, & \text{if} \quad \mathbb{P}_{\boldsymbol{X}} \left(S\Delta S_{\text{bayes}}\right) < \frac{C}{1-\alpha} H^{\frac{\alpha}{1-\alpha}}. \end{cases}$$

Notice that for sufficiently large n and H, we always have $\mathbb{P}_{\boldsymbol{X}} \left(S\Delta S_{\text{bayes}}\right) < \frac{C}{1-\alpha} H^{\frac{\alpha}{1-\alpha}}$. When N and H are not large enough, we can rescale C to satisfy $\mathbb{P}_{\boldsymbol{X}} \left(S\Delta S_{\text{bayes}}\right) < \frac{C}{1-\alpha} H^{\frac{\alpha}{1-\alpha}}$. Therefore, we have

$$R(S) - R(S_{\text{bayes}}) \geq 2\alpha \left(\frac{1-\alpha}{C}\right)^{\frac{1-\alpha}{\alpha}} \mathbb{P}_{\boldsymbol{X}}^{\frac{1}{\alpha}} \left(S\Delta S_{\text{bayes}}\right).$$

Finally, for all $\pi \in \Pi'$ and $H \in \mathbb{N}_+$

$$\mathbb{P}_{\boldsymbol{X}}(S\Delta S_{\text{bayes}}) \leq C_1 (R(S) - R(S_{\text{bayes}}))^{\alpha},$$

where we take $C_1 = \left(\frac{C}{1-\alpha}\right)^{1-\alpha} \left(\frac{1}{2\alpha}\right)^{\alpha} > 0$.

(2) $\Rightarrow$ (1). Let t be an arbitrary number in the interval $[0,1]$, and define the set $S = \{\boldsymbol{X} \colon p(\boldsymbol{X}) \in [\pi - t/H, \pi] \cup (\pi + t/H, 1]\}$. The set S satisfies

$$S\Delta S_{\text{bayes}} = \{\boldsymbol{X} \colon |p(\boldsymbol{X}) - \pi| \leq t/H\}.$$

Based on (3) and the definition of S,

$$R(S) - R(S_{\text{bayes}}) \leq \left(\frac{2t}{H}\right) \mathbb{P}_{\boldsymbol{X}} \left(S\Delta S_{\text{bayes}}\right).$$

Combining the above inequality with (2) gives

$$\mathbb{P}_{\boldsymbol{X}} \left(|p(\boldsymbol{X}) - \pi| \leq t/H\right) = \mathbb{P}_{\boldsymbol{X}} \left(S\Delta S_{\text{bayes}}\right) \leq C\left(\frac{t}{H}\right)^{\frac{\alpha}{1-\alpha}}, \tag{5}$$

assumed that the left hand side of (5) is non-zero (otherwise, the result is trivial), where we set $C = (C_1 2^{\alpha})^{\frac{1}{1-\alpha}}$. Because the above inequality holds for all $t \leq 1$, $\pi \in \Pi'$, and $H \in \mathbb{N}_+$, (1) holds

Case 2: $\alpha = 1$.

(1) $\Rightarrow$ (2). The inequality (4) now becomes

$$R(S) - R(S_{\text{bayes}}) \geq \frac{2t}{H} \mathbb{P}_{\boldsymbol{X}}(S \Delta S_{\text{bayes}}), \quad \text{for all } t \in [0,1], \pi \in \Pi'.$$

Therefore the inequality (2) holds with $C_2 = \frac{1}{2}$ and t = 1.

(2) $\Rightarrow$ (1). We replace C_2 by $\max(C_2, 1)$ in (2). The inequality (5) now becomes

$$\mathbb{P}_{\boldsymbol{X}}(S \Delta S_{\text{bayes}}) \leq \max(2C_2, 2) t \mathbb{P}_{\boldsymbol{X}}(S \Delta S_{\text{bayes}}), \quad \text{for all } t \in [0,1].$$

In particular, the inequality holds for all $H \in \mathbb{N}_+$ and all $\pi \in \Pi'$ which implies

$$\max_{\pi \in \Pi} \mathbb{P}_{\boldsymbol{X}} \left(|p(\boldsymbol{X}) - \pi| \leq t/H \right) = 0, \quad \text{for all } t \in [0,1].$$

Therefore the inequality (1) holds. $\square$