Main assumption and theorems

Chanwoo Lee, November 19, 2020

1 Main assumption

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

$$\mathbb{P}_{\boldsymbol{X}}\left(|p(\boldsymbol{X}) - \pi| \leq t\right) \leq C' t^{\frac{\alpha}{1-\alpha}}, \quad \text{for all } t \in (0, \delta(H)]. \tag{1}$$

This $\delta(H)$ can be independent or dependent on H.

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

$$\begin{aligned}\int_{\boldsymbol{X} \in S\Delta S_{\text{bayes}}} |p(\boldsymbol{X}) - \pi| d\mathbb{P}_{\boldsymbol{X}} &\geq t\left[\mathbb{P}_{\boldsymbol{X}}((S\Delta S_{\text{bayes}}) \cap A)\right] \qquad (2)\\ &\geq t\left(\mathbb{P}_{\boldsymbol{X}}\left(S\Delta S_{\text{bayes}}\right) - \mathbb{P}_{\boldsymbol{X}}(A^c)\right)\\ &\geq t\left(\mathbb{P}_{\boldsymbol{X}}\left(S\Delta S_{\text{bayes}}\right) - C' t^{\frac{\alpha}{1-\alpha}}\right), \quad \text{for all } t \in (0, \delta(H)].\end{aligned}$$

Combining the equality $R(S) - R(S_{\text{bayes}}) = 2\int_{\boldsymbol{X} \in S\Delta S_{\text{bayes}}} |p(\boldsymbol{X}) - \pi| d\mathbb{P}_{\boldsymbol{X}}$,

$$R(S) - R(S_{\text{bayes}}) \geq 2t\left(\mathbb{P}_{\boldsymbol{X}}\left(S\Delta S_{\text{bayes}}\right) - C' t^{\frac{\alpha}{1-\alpha}}\right), \quad \text{for all } t \in (0, \delta(H)]. \tag{3}$$

We maximize the lower bound of (3) with respect to t and obtain the optimal $t_{\text{opt}} \in (0, \delta(H)]$,

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

The corresponding lower bound of the inequality (3) becomes

$$R(S) - R(S_{\text{bayes}}) \geq \begin{cases} 2\alpha\delta(H)\mathbb{P}_{\boldsymbol{X}}\left(S\Delta S_{\text{bayes}}\right), & \text{if} \quad \mathbb{P}_{\boldsymbol{X}}\left(S\Delta S_{\text{bayes}}\right) \geq \frac{C'}{1-\alpha}\delta(H)^{\frac{\alpha}{1-\alpha}},\\ \\ 2\alpha\left(\frac{1-\alpha}{C'}\right)^{\frac{1-\alpha}{\alpha}}\mathbb{P}_{\boldsymbol{X}}^{\frac{1}{\alpha}}\left(S\Delta S_{\text{bayes}}\right), & \text{if} \quad \mathbb{P}_{\boldsymbol{X}}\left(S\Delta S_{\text{bayes}}\right) < \frac{C'}{1-\alpha}\delta(H)^{\frac{\alpha}{1-\alpha}}.\end{cases}$$

1.1 Case 1: $\delta(H)$ is independent on H

When $n \gg H$ or $\delta(H)$ is a constant, we can always have $\mathbb{P}_{\boldsymbol{X}}\left(S\Delta S_{\text{bayes}}\right) < \frac{C'}{1-\alpha}\delta(H)^{\frac{\alpha}{1-\alpha}}$ because left side converges to 0 faster than right side. So we can freely use

$$\mathbb{P}_{\boldsymbol{X}}\left(S\Delta S_{\text{bayes}}\right) \leq c_1(R(S) - R(S_{\text{bayes}}))^{\alpha}. \tag{4}$$

In addition, we can bound variance term based on (4) as

$$\text{Var}\left[\ell(yf(\boldsymbol{X})) - \ell(yf_{\text{bayes}}(\boldsymbol{X}))\right] \leq c_2(R(s) - R(S_{\text{bayes}}))^{\alpha}.$$

Therefore, we have the ℓ_1 norm error

$$\begin{aligned}\mathbb{E}|p(\boldsymbol{X}) - \bar{p}(\boldsymbol{X})| &\leq \max_{\pi \in \Pi}\mathbb{P}\left[\bar{S}(\pi)\Delta S_{\text{bayes}}(\pi)\right] + \frac{1}{2H}\\ &\leq C\max_{\pi\in\Pi}\left[R_\pi(\bar{S}(\pi)) - R_\pi(S_{\text{bayes}}(\pi))\right]^{\alpha} + \frac{1}{2H}\end{aligned}$$

$$\leq C \max\left(\left(\frac{r(s_1+s_2)\log d}{n}\right)^{\alpha/(2-\alpha)}, a_n^\alpha\right) + \frac{1}{2H}.$$

In this case, we can make H arbitrarily big so that H does not affect convergence rate.

1.2 Case 2: $\delta(H)$ is dependent on H

1. When $0 < \alpha < 1$.

   We can choose optimal H with respect to n such that

   $$\max\left(\left(\frac{r(s_1+s_2)\log d}{n}\right)^{\alpha/(2-\alpha)}, a_n^\alpha\right) < \delta(H)^{\alpha/1-\alpha}$$

   Notice that the worst case of $\delta(H)$ is when $\delta(H) = \frac{1}{H}$. In this case, $H \approx n^{\frac{2-\alpha}{1-\alpha}}$ can make

   $$\mathbb{E}|p(\boldsymbol{X}) - \bar{p}(\boldsymbol{X})| \leq C \max\left(\left(\frac{r(s_1+s_2)\log d}{n}\right)^{\alpha/(2-\alpha)}, a_n^\alpha\right) + \left(\frac{1}{n}\right)^{(2-\alpha)/(1-\alpha)}.$$

2. When $\alpha = 1$.

   This case is the worst case among all the cases. From (2), we have

   $$\mathbb{P}_{\boldsymbol{X}}\left(S\Delta S_{\text{bayes}}\right) \leq \frac{c_1}{\delta(H)}(R(S) - R(S_{\text{bayes}})). \tag{5}$$

   In addition, we can bound variance term based on (5) as

   $$\text{Var}\left[\ell(yf(\boldsymbol{X})) - \ell(yf_{\text{bayes}}(\boldsymbol{X}))\right] \leq \frac{c_2}{\delta(H)}(R(s) - R(S_{\text{bayes}})).$$

   I realized that this variance bound can be actually tight in some cases and will talk about it in the next section. Therefore, we have to consider new bound of variance and expectation to calculate $\mathbb{P}_{\boldsymbol{X}}(S\Delta S_{\text{bayes}})$. My calculation has

   $$\mathbb{P}_{\boldsymbol{X}}(S\Delta S_{\text{bayes}}) \leq C \max\left(\left(\frac{r(s_1+s_2)\log d}{n\delta(H)^2}\right), a_n\right).$$

   which implies

   $$\mathbb{E}|p(\boldsymbol{X}) - \bar{p}(\boldsymbol{X})| \leq C \max\left(\left(\frac{r(s_1+s_2)\log d}{n\delta(H)^2}\right), a_n\right) + \frac{1}{H}.$$

   In the constant probability case $p(\boldsymbol{X}) = p$, we have $\delta(H) = 1/H$ and $\alpha = 1$. Therefore, we have

   $$\mathbb{E}|p(\boldsymbol{X}) - \bar{p}(\boldsymbol{X})| \leq C \max\left(\left(\frac{H^2 r(s_1+s_2)\log d}{n}\right), a_n\right) + \frac{1}{H} \leq \mathcal{O}(1/n^{1/3}).$$

1.3 Discussion

1. Why we have sub-optimal probability estimation in constant probability case?

   Possible answer: when we derive upper bound of probability estimation error, we use

   $$\frac{1}{H}\sum_{\pi\in\Pi} \mathbb{P}(S\Delta S_{\text{bayes}}) \leq \max_{\pi\in\Pi} \mathbb{P}(S\Delta S_{\text{bayes}})$$

The worst case of classification error only happens around π such that $|p - \pi| = \mathcal{O}(1/H)$ and we use this worst case bound as upper bound. Since the best case classification error is $\mathcal{O}(1/n)$ while the worst case classification error is $\mathcal{O}(H^2/n)$, I conjecture that the averaged classification error is around $\mathcal{O}(H/n)$ which leads to the optimal probability estimation error. However, I cannot find good way to prove the conjecture.

2. Why large H can worsen the classification error?

   Possible answer: Poor classification error happens when $|p(\boldsymbol{X}) - \pi|$ is very small. Notice that what we really estimate in classification is $\{\boldsymbol{X} : \text{sign}(p(\boldsymbol{X}) - \pi)\}$. Therefore, the estimation becomes extremely hard when $p(\boldsymbol{X}) - \pi$ is around 0 because signal is too low. This is why we have extra $\delta(H)^2$ term in the new bound of $\mathbb{P}_{\boldsymbol{X}}(S \Delta S_{\text{bayes}})$ when $\mathbb{P}(|p(\boldsymbol{X}) - \pi| < t)$ is not small enough.

2 Expectation and variance bound

In the constant probability case such that $p(\boldsymbol{X}) = p$, I calculated main terms on the assumptions.

$$\mathbb{P}(S(\pi) \Delta S_{\text{bayes}}(\pi)) = \begin{cases} \mathbb{P}(\hat{p} < \pi) & \text{if } p \geq \pi, \\ \mathbb{P}(\hat{p} \geq \pi) & \text{if } p < \pi. \end{cases}$$

$$R_\pi(S) - R_\pi(S_{\text{bayes}}) = \begin{cases} (p - \pi)\mathbb{P}(\hat{p} < \pi) & \text{if } p \geq \pi, \\ (\pi - p)\mathbb{P}(\hat{p} \geq \pi) & \text{if } p < \pi. \end{cases}$$

Consider when $p \geq \pi$.

$$\begin{aligned} \text{Var}\left[\ell(yf(\boldsymbol{X})) - \ell(yf_{\text{bayes}}(\boldsymbol{X}))\right] &\leq \mathbb{E}\left[w(y)^2 \left(\mathbb{1}\{y \neq \text{sign}(\hat{p} - \pi)\} - \mathbb{1}\{y \neq \text{sign}(p - \pi)\}\right)^2\right] \\ &= \mathbb{E}\left[(1 - \pi)^2 p \mathbb{1}\{\hat{p} < \pi\} + \pi^2 (1 - p)(\mathbb{1}\{\hat{\geq}\pi\} - 1)^2\right] \\ &= \left((\pi - p)^2 + p(1 - p)\right) \mathbb{P}(\hat{p} < \pi). \end{aligned}$$

Similarly, when $p < \pi$, we have

$$\text{Var}\left[\ell(yf(\boldsymbol{X})) - \ell(yf_{\text{bayes}}(\boldsymbol{X}))\right] \leq \left((\pi - p)^2 + p(1 - p)\right) \mathbb{P}(\hat{p} \geq \pi).$$

Therefore, we can check two relationship.

$$\begin{aligned} \mathbb{P}(S(\pi) \Delta S_{\text{bayes}}(\pi)) &\leq \frac{c_1}{|p - \pi|}(R_\pi(S) - R_\pi(S_{\text{bayes}})) \\ \text{Var}\left[\ell(yf(\boldsymbol{X})) - \ell(yf_{\text{bayes}}(\boldsymbol{X}))\right] &\leq \frac{c_2}{|p - \pi|}(R_\pi(S) - R_\pi(S_{\text{bayes}})) \end{aligned}$$

Therefore, our bound in the previous section can be tight.