Appendix of prediction and outlier detection in classification problems

Leying Guan†
Stanford University, Stanford, CA, USA
E-mail: lguan@stanford.edu
Robert Tibshirani
Stanford University, Stanford, CA, USA

A. Proofs of Theorems 1-2

Before we prove Theorems 1-2, we first state some Lemmas that will be useful in the proofs. Let $\hat{F}_k$ be the empirical distribution of $F_k$ for given samples based on the context. In this paper, when it comes to the empirical estimates, they are estimated with samples of size $cn$ for a constant $c > 0$ that depends on the context.

**Lemma 1.** Let $G$ and $\hat{G}$ be the CDF and empirical CDF of a univariate variable in $\mathbb{R}$. Let $V_n = \sup_t |G(t) - \hat{G}(t)|$, then, for a large enough constant $B$, we have $P(V_n \leq B \sqrt{\log n/n}) \rightarrow 1$.

**Proof.** Lemma 1 is a result of the classical empirical process theory (?), see also Lemma C.1 in Lei et al. (2013).

**Lemma 2.** Let $g(t)$ be the density of the univariate variable $t \in \mathbb{R}$. Let $\hat{g}(t)$ be its Gaussian kernel density estimation with bandwidth $h_n > \frac{\log n}{n}$. Suppose $g(t)$ is bounded and Hölder continuous with exponent $1 \geq \alpha > 0$, then, there exists a large enough constant $B$, such that with probability at least $1 - \frac{1}{n}$, we have

$$\|g(t) - \hat{g}(t)\|_{\infty} < B(h_n^{\frac{2}{\alpha}} + \sqrt{\frac{\log n}{nh_n}}).$$

**Proof.** Obviously, the Gaussian kernel $K(z)$ for $z \in \mathbb{R}$ satisfies Assumption 2-3 in (?) (the spherically symmetric and non-increasing Assumption, and the exponential decay Assumption), and $g(t)$ is bounded (Assumption 1 in ?)). Then, Lemma 2 is a special case of Theorem 2 in ?).

†Address for correspondence: Leying Guan, Department of Statistics, Stanford University, Stanford, CA, 94305, USA. E-mail: lguan@stanford.edu
A.1. Proof of Theorem 1

The proof follows the same procedure as in Lei et al. (2013). Let \( A_k \) be the accepted region for class \( k \) under the oracle BCOPS. Let \( R_{n,0} = P_{\text{test}}(x \notin A_n) \), \( R_{n,1} = \sup_{x \in A_n} |\hat{v}_k(x) - v_k(x)| \) and \( R_{n,2} = |Q(\alpha; v_k, F_k) - Q(\alpha; v_k, \hat{F}_k)| \). Then we have

\[
(\hat{A}_k \setminus A_k) \cap A_n = \{x | x \in A_n, v_k(x) < Q(\alpha; v_k, F_k), \hat{v}_k(x) \geq Q(\alpha; \hat{v}_k, \hat{F}_k)\} \subseteq \{x | Q(\alpha; v_k, F_k) - 2R_{n,1} - R_{n,2} \leq v_k(x) < Q(\alpha; v_k, F_k)\},
\]

and

\[
\int_{\hat{A}_k \setminus A_k} f_{\text{test}}(x)dx \leq \int_{(\hat{A}_k \setminus A_k) \cap A_n} f_{\text{test}}(x)dx + R_{n,0} \leq \int_{(\hat{A}_k \setminus A_k) \cap A_n} \frac{v_k(x)}{Q(\alpha; v_k, F_k) - 2R_{n,1} - R_{n,2}} f_{\text{test}}(x)dx + R_{n,0} \leq \frac{P_k((\hat{A}_k \setminus A_k) \cap A_n)}{Q(\alpha; v_k, F_k) - 2R_{n,1} - R_{n,2} + R_{n,0}} + R_{n,0}.
\]

By Assumption 2, for a large enough constant \( B_1 \), we have

\[
P(R_{n,1} \leq B_1(\frac{\log n}{n})^{\frac{\alpha}{2}}) \to 0, \quad P(R_{n,0} \leq B_1(\frac{\log n}{n})^{\frac{\alpha}{2}}) \to 0.
\]

Let \( G \) and \( \hat{G} \) be the CDF and empirical CDF of \( v_k(x) \). By Lemma 1 on the one hand, with probability approaching 1, for any constant \( \delta \) and a constant \( B_2 \) large enough, we have

\[
|\hat{G}(Q(\alpha - \delta; v_k, F_k)) - (\alpha - \delta)| \leq B_2 \sqrt{\frac{\log n}{n}}.
\]

On the other hand, by Assumption 4, we have

\[
\delta \geq c_1 |Q(\alpha; v_k, F_k) - Q(\alpha + \delta; v_k, F_k)|^\gamma.
\]

In other words, with probability approaching 1, the following is true

\[
Q(\alpha - B_2 \sqrt{\frac{\log n}{n}}; v_k, F_k) \leq Q(\alpha; v_k, \hat{F}_k) \leq Q(\alpha + B_2 \sqrt{\frac{\log n}{n}}; v_k, F_k).
\]

and

\[
|Q(\alpha + B_2 \sqrt{\frac{\log n}{n}}; v_k, F_k) - Q(\alpha; v_k, F_k)| \leq (\frac{B_2}{c_1}) \sqrt{\frac{\log n}{n}}^\gamma.
\]

Hence, \( R_{n,2} \leq (\frac{B_2}{c_1}) \sqrt{\frac{\log n}{n}}^\gamma \).

For the numerator of equation (1), we use equations (2)-(3) and apply Assumption 4 again, for a large enough constant \( B_3 \), we have,

\[
P_k((\hat{A}_k \setminus A_k) \cap A_n) \leq P_k(v_k, \alpha - 2R_{n,1} - R_{n,2} \leq v_k(x) \leq v_k, \alpha) \leq (c_2(2R_{n,1} + R_{n,2}))^\gamma \leq B_3(\sqrt{\frac{\log n}{n}} + (\frac{\log n}{n})^{\frac{\alpha}{2}}).
\]
For the denominator of equation (1), when $\alpha$ is a positive constant, since $v_k(x) = f_k(x)$ is non-zero as long as $f_k(x)$ is non-zero, we must have that $Q(\alpha; v_k, F_k)$ is also a positive constant, and we can always take $n$ large enough, such that

$$Q(\alpha; v_k, F_k) - 2R_{n,1} - R_{n,2} \geq \frac{1}{2}Q(\alpha; v_k, F_k) > 0. \tag{4}$$

and that for a constant $B$ large enough, with probability approaching 1:

$$\int_{\hat{A}_k \setminus A_k} f_{\text{test}}(x) dx \leq B \frac{K}{n} \min(\gamma_1, \beta_2, 1)^2. \tag{5}$$

Combining eqs.(4)-(5), with probability approaching 1, we have

$$\int (|\hat{C}(x) - |C(x)|| f_{\text{test}}(x) dx \leq \sum_k \int_{\hat{A}_k \setminus A_k} f_{\text{test}}(x) dx \leq B \frac{K}{n} \min(\gamma_1, \beta_2, 1)^2. \tag{6}$$

with probability approaching 1 for a large enough constant $B$.

A.2. Proof of Theorem 2

(a): In section 4.2, we have only defined $P_{l,k}$ for $k = 1, \ldots, K$, here we include the class $R$ as well following the same definition: $P_{l,R} = P_R(\eta_l(x) \in S_l)$ for convenience. We first show that

$$\hat{\pi} \to \tilde{\pi} + \epsilon \Sigma^{-1} P^T P_R.$$

Since $\Sigma$ is invertible with smallest eigenvalue $\sigma_{\min} \geq c > 0$ for a constant $c$, it is sufficient to show

$$\hat{P}_{l,k} \overset{p}{\to} P_{l,k}, \quad \forall k = 1, \ldots, K, R, \quad l = 1, \ldots, K.$$

Recall the definition of $\hat{P}_{l,k}$:

$$\hat{P}_{l,k} := P_{F_k}(\eta_l(x) \in \hat{S}_l) = (\hat{P}_{l,k} - \tilde{P}_{l,k}) + \tilde{P}_{l,k},$$

where

- $\tilde{P}_{l,k} := P_{F_k}(\eta_l(x) \in \hat{S}_l)$.
- $\hat{S}_l := \{ t : \hat{g}_{l,t}(t) \geq \hat{g}_{l,\zeta} \}, \hat{g}_{l,t}(t)$ is the kernel estimation of $g_{l,t}(t)$ and $\hat{g}_{l,\zeta} := Q(\zeta; \hat{g}_{l,t} \circ \eta_l, F_l)$.

We first show that

$$\hat{P}_{l,k} \overset{p}{\to} P_{l,k}, \quad \forall k = 1, \ldots, K, R, \quad l = 1, \ldots, K.$$

Let $\Delta = \hat{P}_{l,k} - P_{l,k} = \Delta_1 - \Delta_2$ where $\Delta_1 = \int_t g_{l,k}(t) 1_{\hat{g}_{l,\zeta} \leq \hat{g}_{l,t}(t)} 1_{\hat{g}_{l,t}(t) < g_{l,\zeta}}$ and $\Delta_2 = \int_t g_{l,k}(t) 1_{\hat{g}_{l,\zeta} \leq \hat{g}_{l,t}(t)} 1_{\hat{g}_{l,t}(t) < g_{l,\zeta}}$. We now prove $\Delta_1 \overset{p}{\to} 0$. Let $R_{n,1} = \|\hat{g}_{l,t}(t) - g_{l,t}(t)\|_{\infty}$ and
Notice that

\[ R_{n,2} = |g_{l,\zeta} - g_{l,\zeta}|. \]

Then,

\[ \Delta_1 = \int_t g_{l,k}(t) I_{g_{l,\zeta} + \hat{g}_l(t) \leq g_{l,1}(t) \leq g_{l,\zeta}} \]
\[ \leq \int_t g_{l,k}(t) I_{g_{l,\zeta} + \hat{g}_l(t) - R_n,2 \leq g_{l,1}(t) \leq g_{l,\zeta}} \]
\[ \leq \left( \max_t g_{l,k}(t) \right) P_t(g_{l,\zeta} + \hat{g}_l(t) - R_n,1 \leq g_{l,1}(t) \leq R_n,2 \leq g_{l,1}(t) \leq g_{l,\zeta}). \]

Notice that

- Under Assumption 3, let the constant \( \alpha > 0 \) be the Hölder exponent for \( g_{l,1}(t) \), we apply Lemma 2 and have that for a large enough constant \( B \),

\[ P(R_{n,1} \geq B(\frac{\log n}{nh_n} + h_n^a)) \to 0 \Rightarrow R_{n,1} \xrightarrow{p} 0. \]  

(6)

- Also, \( R_{n,2} = |\hat{g}_{l,\zeta} - Q(\zeta; g_{l,1}, \hat{F}_1) + Q(\zeta; g_{l,1}, \hat{F}_1) - g_{l,\zeta}| \leq R_{n,1} + |Q(\zeta; g_{l,1}, \hat{F}_1) - g_{l,\zeta}|. \)

Let \( G \) and \( \hat{G} \) be the CDF and empirical CDF of \( g_{l,1}(\eta(x)) \) in class 1. Apply Lemma 1 and Assumption 4, we know that there exist positive constants \( B \) and \( \delta_0 \), such that, with probability approaching 1, we have

\[ |G(Q(\zeta - \delta; g_{l,1}, \hat{F}_1)) - (\zeta - \delta)| \leq \|G - \hat{G}\|_\infty \leq B\sqrt{\frac{\log n}{n}}, \forall \delta \in (-\delta_0, \delta_0), \]
\[ |g_{l,\zeta} - Q(\zeta; g_{l,1}, \hat{F}_1)| \leq B(\frac{\log n}{n})^{\frac{1}{2}}. \]

As a consequence, we have \( R_{n,2} \xrightarrow{p} 0. \)

Again, by Assumption 4, for a large enough constant \( B \), we have that, with probability approaching 1,

\[ \Delta_1 \leq B \left( \frac{\max_t g_{l,k}(t)}{|g_{l,\zeta} - R_{n,1} - R_{n,2}|} \right) (R_{n,1} + R_{n,2}) \gamma \xrightarrow{p} 0. \]  

(7)

By symmetry, \( \Delta_2 \xrightarrow{p} 0 \) will follow the same argument argument. Hence, we have \( \Delta \xrightarrow{p} 0. \)

Next, we show that

\[ |\hat{P}_{l,k} - \hat{P}_{l,k}| \xrightarrow{p} 0. \]

Let \( G \) and \( \hat{G} \) be the CDF and empirical CDF of \( g_{l,1}(\eta(x)) \) in class \( k \), this is true with the argument below:

\[ |\hat{P}_{l,k} - \hat{P}_{l,k}| = |P_{F_k}(\hat{g}_{l,1}(\eta(x)) \leq \hat{g}_{l,\zeta}) - P_{F_k}(\hat{g}_{l,1}(\eta(x)) \leq \hat{g}_{l,\zeta})| \]
\[ \leq \max (|\hat{G}(\hat{g}_{l,\zeta} + R_{n,1}) - G(\hat{g}_{l,\zeta} - R_{n,1})|, |\hat{G}(\hat{g}_{l,\zeta} - R_{n,1}) - G(\hat{g}_{l,\zeta} + R_{n,1})|) \]
\[ \leq ||\hat{G} - G||_\infty + |G(\hat{g}_{l,\zeta} + R_{n,1}) - G(\hat{g}_{l,\zeta} - R_{n,1})|. \]
By Lemma 1, we have that \( \| \hat{G} - G \|_\infty < B \sqrt{\frac{\log n}{n}} \) for a large enough constant \( B \) with probability approaching 1. Following the same argument as eq. (7), we know that for a large enough constant \( B \):

\[
|G(\hat{g}_{l,\zeta} + R_{n,1}) - G(\hat{g}_{l,\zeta} - R_{n,1})| \leq B \frac{\max g_{l,k}(t)}{|g_{l,\zeta} - R_{n,2} - R_{n,1}|} (2R_{n,1})^\gamma \xrightarrow{p} 0.
\]

Hence, we also have \( |\hat{P}_{l,k} - \tilde{P}_{l,k}| \xrightarrow{p} 0 \) and we have thus proved our statement.

(b): When \( P_R = 0 \), from part (a), we know that \( \hat{\pi} \to \tilde{\pi} \).