Supplementary Material of *Binary Classification with Karmic, Threshold-Quasi-Concave Metrics*

In this document, we include some supplementary materials for *Binary Classification with Karmic, Threshold-Quasi-Concave Metrics*. Throughout the document, we follow the notations in the main paper. We will use bolded $\boldsymbol{C}$ as the confusion matrix and C as absolute constants where the exact value might change from line to line.

A Proofs in Section 3

Proof of Theorem 3.1. We consider a continuous extension for the space of classifiers: $\mathcal{F} = \{f : \mathcal{X} \mapsto [-1, 1]\}$. Let $\mu(X)$ be the marginal distribution of X with respect to $\mathbb{P}$. For any classifier $f \in \mathcal{F}$ and $\mu(X)$, the confusion matrix is given by its entries: true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). These quantities can be represented in terms of expectation as:

$$\begin{aligned}
\mathrm{TP}(f, \mathbb{P}) &= \mathbb{P}(Y = +1, f = +1) = \int \eta(x) \left[\frac{1 + f(x)}{2}\right] d\mu \\
\mathrm{FP}(f, \mathbb{P}) &= \mathbb{P}(Y = -1, f = +1) = \int (1 - \eta(x)) \left[\frac{1 + f(x)}{2}\right] d\mu \\
\mathrm{FN}(f, \mathbb{P}) &= \mathbb{P}(Y = +1, f = -1) = \int \eta(x) \left[\frac{1 - f(x)}{2}\right] d\mu \\
\mathrm{TN}(f, \mathbb{P}) &= \mathbb{P}(Y = -1, f = -1) = \int (1 - \eta(x)) \left[\frac{1 - f(x)}{2}\right] d\mu
\end{aligned} \tag{A.1}$$

Let $\pi = \mathbb{P}(Y = 1) = \int \eta(x) d\mu$ denote the marginal distribution of Y.

The confusion matrix is continuous and Frechét differentiable with respect to classifiers $f \in \mathcal{F}$ with the derivatives given pointwise by:

$$[\nabla_f \boldsymbol{C}(f, \mathbb{P})]_x = \frac{1}{2}(\eta(x), 1 - \eta(x), -\eta(x), -(1 - \eta(x)))d\mu(x)$$

Recall that for any $\mathcal{U}$ given by some function $\mathcal{G} : [0, 1]^4 \mapsto \mathbb{R}$:

$$\mathcal{U}(f, \mathbb{P}) \equiv \mathcal{G}(\boldsymbol{C}(f, \mathbb{P})).$$

Let $\nabla \mathcal{G} = (g_1, g_2, g_3, g_4)^T$, and note that g_i depend on the classifier f via the confusion matrix. Applying the chain rule, the Frechét derivative may be computed as:

$$\begin{aligned}
[\nabla \mathcal{U}(f, \mathbb{P})]_x &= \nabla \mathcal{G}(\boldsymbol{C}(f))^T \cdot [\nabla \boldsymbol{C}(f)]_x \\
&= \frac{1}{2}(g_1(\boldsymbol{C}(f))\eta(x) + g_2(\boldsymbol{C}(f))(1 - \eta(x)) - g_3(\boldsymbol{C}(f))\eta(x) - g_4(\boldsymbol{C}(f))(1 - \eta(x)))d\mu \\
&= \frac{1}{2}\left(\nabla \mathcal{G}(\boldsymbol{C}(f))^T(1, -1, -1, 1)^T \eta(x) - \nabla \mathcal{G}(\boldsymbol{C}(f))^T(0, -1, 0, 1)^T\right) d\mu
\end{aligned}$$

By Assumption 1, we know that $\nabla\mathcal{G}(\boldsymbol{C}(f))^T(1,-1,-1,1)^T > 0$. The first order optimality condition holds for any optimal point f^*: $\langle\nabla\mathcal{U}(f^*), f^* - f\rangle \geq 0, \quad \forall f \in \mathcal{F}$, which is equivalent to:

$$\int_{\mathcal{X}} [\nabla\mathcal{U}(f^*)]_x (f^*(x) - f(x)) \geq 0, \quad \forall f \in \mathcal{F}. \tag{A.2}$$

For any $f \in \mathcal{F}$, define the "critical set of $\mathcal{G}(f, P)$" where the utility has zero derivative:

$$A_3(f) = \left\{ x : \eta(x) = \frac{\nabla\mathcal{G}(\boldsymbol{C}(f))^T(0,-1,0,1)^T}{\nabla\mathcal{G}(\boldsymbol{C}(f))^T(1,-1,-1,1)^T} \right\}.$$

We will refer to $A_3^* := A_3(f_{\mathcal{F}}^*)$ as the Bayes critical set. Similarly, define

$$\begin{aligned} A_1(f) &= \{x \in \mathcal{X} : [\nabla\mathcal{U}(f)]_x > 0\} \\ &= \left\{ x \in \mathcal{X} : \eta(x) > \frac{\nabla\mathcal{G}(\boldsymbol{C}(f))^T(0,-1,0,1)^T}{\nabla\mathcal{G}(\boldsymbol{C}(f))^T(1,-1,-1,1)^T} \right\}, \end{aligned}$$

and

$$\begin{aligned} A_2(f) &= \{x \in \mathcal{X} : [\nabla\mathcal{U}(f)]_x < 0\} \\ &= \left\{ x \in \mathcal{X} : \eta(x) < \frac{\nabla\mathcal{G}(\boldsymbol{C}(f))^T(0,-1,0,1)^T}{\nabla\mathcal{G}(\boldsymbol{C}(f))^T(1,-1,-1,1)^T} \right\}. \end{aligned}$$

It is easily seen that $A_1(f) \cup A_2(f) \cup A_3(f) = \mathcal{X}$. Hence Eq. (A.2) is equivalent to

$$\int_{A_1} [\nabla\mathcal{U}(f^*)]_x (f^*(x) - f(x)) + \int_{A_2} [\nabla\mathcal{U}(f^*)]_x (f^*(x) - f(x)) \geq 0, \, \forall f \in \mathcal{F}. \tag{A.3}$$

We now claim $f^*(x) = 1$ on A_1 and $f^*(x) = -1$ on A_2. To see this, note that for any optimal point f^*, if there exists a subset $U \subset \mathcal{X}$ on which it fails to satisfy the claim, without loss of generality, assume $U \subset A_1^*$, we are going to show $\mu(U) = 0$. By assumption, $f^*(x) < 1$ on U. Let $f(x) = \mathbb{I}_U(x) + f^*\mathbb{I}_{U^c}$, and plug it into (A.3), we have

$$\int_U [\nabla\mathcal{U}(f^*)]_x (f^*(x) - 1) \geq 0$$

The integrand is strictly negative so $\mu(U) = 0$. Thus, $f^*(x) = \begin{cases} 1 & x \in A_1^* \\ -1 & x \in A_2^* \end{cases}$ holds almost everywhere. In this paper we focus on distributions where the critical set of $\mathcal{U}(f, \mathbb{P})$ satisfies $P(A_3(f)) = 0$. For instance, this occurs when the the conditional distribution $\eta(\cdot)$ is injective and the marginal instance distribution μ is continuous.

□

Proof of Theorem 3.2. The proof follows directly from the definition of $\mathcal{V}_\eta(\delta, \mathbb{P})$ and strictly quasi-concave. □

We proceed to the proof of Corollary 3.1. Before that, we introduce the following lemma to characterize the derivative of $\mathcal{V}$ for any given η.

Lemma A.1. *When $\eta(X)$ is fixed, $\mathcal{V}_\eta(\delta)$ is differentiable w.r.t. δ, and*

$$\mathcal{V}'_\eta(\delta, \mathbb{P}) = \left(\nabla\mathcal{G}(\boldsymbol{C})^T(-1,1,1,-1)^T\delta + \nabla\mathcal{G}(\boldsymbol{C})^T(0,-1,0,1)^T\right) p_\eta(\delta). \tag{A.4}$$

where p_η is the density associated with random variable $\eta(X)$.

Proof of Lemma A.1. By Assumption 2, $\eta(X)$ is absolutely continuous with respect to μ, hence has a density, denoted as $p_\eta(x)$. When $f(x) = \text{sign}\,(\eta(x) - \delta)$, we could conduct the change of variable $u = \eta(x)$ and have

$$\text{TP}(f, \mathbb{P}) = \int_{\{x:\eta(x)>\delta\}} \eta(x) p_X(x) dx = \int_{u=\delta}^{\infty} u \frac{p_\eta(u)}{\eta'(x)} \eta'(x) dx = \int_{\delta}^{\infty} u p_\eta(u) du \tag{A.5}$$

By Leibniz integral rule, we take the partial derivative with respect to δ and have,

$$\frac{\partial \text{TP}(f)}{\partial \delta} = -\delta p_\eta(\delta) \tag{A.6}$$

Similarly for the other confusion matrix elements we have

$$\frac{\partial \text{FP}(f)}{\partial \delta} = -(1-\delta) p_\eta(\delta),\ \frac{\partial \text{FN}(f)}{\partial \delta} = \delta p_\eta(\delta),\ \frac{\partial \text{TN}(f)}{\partial \delta} = (1-\delta) p_\eta(\delta)$$

By chain rule,

$$\mathcal{V}'_\eta(\delta, \mathbb{P}) = \left(\nabla \mathcal{G}(\boldsymbol{C})^T (-1, 1, 1, -1)^T \delta + \nabla \mathcal{G}(\boldsymbol{C})^T (0, -1, 0, 1)^T\right) p_\eta(\delta). \tag{A.7}$$

□

Now we are in position to prove Proposition 3.1.

Proof of Proposition 3.1. For some $\eta(x)$ and threshold δ, define $f(x) = \text{sign}\,(\eta(x) - \delta)$. Recall that we define $v(\delta) = (-\delta, -(1-\delta), \delta, 1-\delta)^T$. For convenience of the analysis, we also define $H_\eta(\delta, \mathbb{P}) = \nabla \mathcal{G}(\boldsymbol{C}(f, \mathbb{P}))^T v(\delta)$. We drop the dependency of $\mathbb{P}$ when there is no confusion.

If $\mathcal{G}$ is ratio of linear Let $\mathcal{G} = \frac{\boldsymbol{a}^T \boldsymbol{C}}{\boldsymbol{b}^T \boldsymbol{C}}$ for two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^4$. Taking derivative of $\mathcal{G}$ we have

$$\nabla \mathcal{G}(\boldsymbol{C}) = \frac{(\boldsymbol{b}^T \boldsymbol{C}) \boldsymbol{a} - (\boldsymbol{a}^T \boldsymbol{C}) \boldsymbol{b}}{(\boldsymbol{b}^T \boldsymbol{C})^2} \tag{A.8}$$

where the denominator is always positive by assumption. Now for the function H, with Eq. (A.8) plugged in, we have

$$H_\eta(\delta) = \frac{(\boldsymbol{b}^T \boldsymbol{C})(\boldsymbol{a}^T v(\delta)) - (\boldsymbol{a}^T \boldsymbol{C})(\boldsymbol{b}^T v(\delta))}{(\boldsymbol{b}^T \boldsymbol{C})^2}$$

Since all we care about is the sign of H, so we focus on the numerator of it and denote $\tilde{H}_\eta(\delta) = (\boldsymbol{b}^T \boldsymbol{C})(\boldsymbol{a}^T v(\delta)) - (\boldsymbol{a}^T \boldsymbol{C})(\boldsymbol{b}^T v(\delta))$. Taking derivative of $\tilde{H}$ we have

$$\begin{aligned} \tilde{H}'_\eta(\delta) =& (\boldsymbol{b}^T \boldsymbol{C})(\boldsymbol{a}^T (-1, 1, 1, -1)^T) - (\boldsymbol{a}^T \boldsymbol{C})(\boldsymbol{b}^T (-1, 1, 1, -1)^T) \\ =& \nabla \mathcal{G}(\boldsymbol{C}(f))^T (-1, 1, 1, -1)^T (\boldsymbol{b}^T \boldsymbol{C})^2 \leq -C_B (\boldsymbol{b}^T \boldsymbol{C})^2 \end{aligned} \tag{A.9}$$

By Assumption 1 and the assumption (b) in the statement of the proposition, we have $\tilde{H}'_\eta(\delta) < 0, \forall \delta \in [0, 1]$. It is also easy to check that $\tilde{H}(0) > 0$ and $\tilde{H}(1) < 0$. Hence $\tilde{H}_\eta(\delta, \mathbb{P}) = 0$ has a unique solution, which implies equation $H_\eta(\delta, \mathbb{P}) = 0$ has a unique solution δ^*. Also by the monotonicity of $\tilde{H}$, we have $\text{sign}\,(H_\eta(\delta)) = \text{sign}\left(\tilde{H}_\eta(\delta)\right) = 1, \forall \delta < \delta^*$ and $\text{sign}\,(H_\eta(\delta)) = \text{sign}\left(\tilde{H}_\eta(\delta)\right) = -1, \forall \delta > \delta^*$. Hence δ^* is the maximizer of $\mathcal{V}_\eta(\delta)$.

If $\mathcal{G}$ is concave Taking derivative of H, we have

$$H'(\delta) = v(\delta)\nabla^2\mathcal{G}(\boldsymbol{C}(f))v(\delta)^T + \nabla\mathcal{G}(\boldsymbol{C}(f))^T(-1,1,1,-1)^T$$

By Assumption 1, we know that

$$H'(\delta) \leq \nabla\mathcal{G}(\boldsymbol{C}(f))^T(-1,1,1,-1)^T \leq -C_B \tag{A.10}$$

Also note that $H_\eta(0) > 0$ and $H_\eta(1) < 0$, by strict monotonicity of H, we know that $H_\eta(\delta) = 0$ has one unique solution $\delta^* \in (0,1)$. By continuity, there exists ρ_0 such that $\delta \in [\rho_0, 1-\rho_0]$. Let δ^* be the solution of Equation (4), then $\forall t < \delta^*$, $\mathcal{V}'_\eta(\delta) < 0$, and $\forall t > \delta^*$, $\mathcal{V}'_\eta(\delta) > 0$. Hence for all $t < \delta^*$, $H(t) > H(\delta^*) = 0$, which implies $\mathcal{V}'(t) > 0$. Similarly, $\mathcal{V}'(t) < 0$ when $t > \delta^*$. Note p_η is non-zero everywhere, so $\operatorname{sign}(V'(\delta)) = \operatorname{sign}(H(\delta))$.

Now we can show that $\mathcal{V}(\delta)$ is quasi-concave by definition. Note $\mathcal{V}(\delta)$ is increasing before δ^* and decreasing afterwards. For $x, y \in [0,1]$, if $\mathcal{V}(y) \geq \mathcal{V}(x)$, and $x \leq \delta^*$, then $y \geq x$, and $\mathcal{V}'(x)(y-x) \geq 0$. If $\mathcal{V}(y) \geq \mathcal{V}(x)$, and $x \geq \delta^*$, then $y \leq x$, and we still have $\mathcal{V}'(x)(y-x) \geq 0$. For the strictness, by Assumption 2, $\mathcal{V}(\delta) = 0$ if and only if $H(\delta) = 0$, which is only achieved at δ^*. □

B Proofs in Section 4

Proof of Lemma C.7. By the update rule of Algorithm 2 and condition 2, we have $\hat{\delta} \in (\delta^* - \gamma - \frac{\epsilon_0}{2}, \delta^* + \gamma + \frac{\epsilon_0}{2})$.

Now the difference in utility can be bounded as follows.

$$\begin{aligned}
&|\mathcal{U}(\operatorname{sign}(\hat{\eta}(X) - \delta^*), \mathbb{P}) - \mathcal{U}(\operatorname{sign}\left(\hat{\eta}(X) - \hat{\delta}\right), \mathbb{P})| \\
\leq &\underbrace{\left|\int_{\delta^*-\gamma-\frac{\epsilon_0}{2}}^{\delta^*-\gamma} \mathcal{V}'_{\hat{\eta}}(\delta, \mathbb{P})d\delta\right| + \left|\int_{\delta^*+\gamma}^{\delta^*+\gamma+\frac{\epsilon_0}{2}} \mathcal{V}'_{\hat{\eta}}(\delta, \mathbb{P})d\delta\right|}_{\mathrm{I}} + \underbrace{\left|\int_{\delta^*-\gamma}^{\delta^*+\gamma} \mathcal{V}'_{\hat{\eta}}(\delta, \mathbb{P})d\delta\right|}_{\mathrm{II}}.
\end{aligned} \tag{B.1}$$

By condition 1, $\mathrm{I} \leq L_v\epsilon_0$. Eq. (B.1) can be bounded by the flatness of $\mathcal{V}_{\hat{\eta}}(\delta)$. By condition 3, we have,

$$\left|\int_{\delta^*-\gamma}^{\delta^*+\gamma} \mathcal{V}'_{\hat{\eta}}(\delta, \mathbb{P})d\delta\right| \leq \int_{\delta^*-\gamma}^{\delta^*+\gamma} |\mathcal{V}'_{\hat{\eta}}(\delta, \mathbb{P})|d\delta \leq 2\gamma \cdot c_1\gamma = 2c_1\gamma^2. \tag{B.2}$$

Taking $\epsilon_0 = \frac{c_1}{L_v}\gamma^2$ and combining Eqs. (B.1) and (B.2), we have

$$\mathcal{U}(\operatorname{sign}(\hat{\eta}(X) - \delta^*), \mathbb{P}) - \mathcal{U}(\operatorname{sign}\left(\hat{\eta}(X) - \hat{\delta}\right), \mathbb{P}) \leq 3c_1\gamma^2.$$

□

C Proofs for the rates of convergence

In this section we present the proofs of result in Section 5.

The following lemma transforms the convergence in probability as stated in Assumption 4 into the convergence in expectation, resulting in the following lemma.

Lemma C.1. *If Assumption 4 holds, then*

$$\mathbb{E}\int |\eta_n - \eta| d\mu(X) \leq \frac{c_3}{\sqrt{a_n}}.$$

Proof of Lemma C.1. Let $\mathcal{S}$ be the sample set. By Fubini's theorem,

$$\begin{aligned}\mathbb{E}\int|\eta_n-\eta|d\mu =&\mathbb{E}_{\mathcal{S}}\mathbb{E}_X|\eta_n(X)-\eta(X)|\\ =&\mathbb{E}_X\mathbb{E}_{\mathcal{S}}|\eta_n(X)-\eta(X)|\\ =&\int_0^\infty P_X(\mathbb{E}_{\mathcal{S}}|\eta_n(X)-\eta(X)|\geq t)dt \end{aligned}\tag{C.1}$$

Now for each $t>0$, define $T(t)=\{x\in\mathcal{X}:\sup P(|\eta_n(x)-\eta(x)|)\geq t)\leq C_1\exp(-C_2a_nt^2)\}$. By assumption $P(T(t)^c)=0$.

$$\begin{aligned}P_X(\mathbb{E}_{\mathcal{S}}|\eta_n(X)-\eta(X)|\geq t)\leq&P(T(t)^c)+P(X\in T(t),\mathbb{E}_{\mathcal{S}}|\eta_n(X)-\eta(X)|\geq t)\\ \leq&P(\sup_{\mathcal{S}}|\eta_n(X)-\eta(X)|\geq t)\\ \leq&C_1\exp(-C_2a_nt^2)\end{aligned}$$

Plugging back to Eq. (C.1), we have

$$\mathbb{E}\int|\eta_n-\eta|d\mu\leq\int_0^\infty C_1\exp(-C_2a_nt^2)dt\leq\frac{C_3}{\sqrt{a_n}}$$

□

Lemma C.2. *Assume Assumption 4 is satisfied, and $f_n=sign\,(\hat{\eta}_n-\delta^*)$. Then*

$$\mathbb{E}\left[|\eta-\delta^*|1(f_n\neq f^*)\right]\leq C_3a_n^{-\frac{1+\alpha}{2}}.$$

Proof. For some $\epsilon>0$, define events

$$A_0=\{x\in\mathcal{X}:0\leq|\eta-\delta^*|\leq\epsilon\},\qquad A_j=\{x\in\mathcal{X}:2^{j-1}\epsilon<|\eta-\delta^*|\leq2^j\epsilon\},\forall j\geq1.$$

Note that $f_n\neq f^*$ implies $|\hat{\eta}_n-\eta|>|\eta-\delta^*|$, we have

$$\begin{aligned}\mathbb{E}|\eta-\delta^*|1(f_n\neq f^*)=&\sum_{j=0}^\infty\mathbb{E}|\eta-\delta^*|1(f_n\neq f^*)1(X\in A_j)\\ \leq&\epsilon P(0\leq|\eta-\delta^*|\leq\epsilon)+\sum_{j=1}^\infty\mathbb{E}|\eta-\delta^*|1(|\hat{\eta}_n-\eta|>|\eta-\delta^*|)1(X\in A_j)\\ \leq&\epsilon P(0\leq|\eta-\delta^*|\leq\epsilon)+\sum_{j=1}^\infty\mathbb{E}2^j\epsilon1(|\hat{\eta}_n-\eta|>|\eta-\delta^*|)1(X\in A_j)\\ \leq&\epsilon P(0\leq|\eta-\delta^*|\leq\epsilon)+\sum_{j=1}^\infty2^j\epsilon C_1\exp(-C_2a_n(2^{j-1}\epsilon)^2)P(X\in A_j)\\ \leq&C_0\epsilon^{1+\alpha}+C_0C_1\sum_{j=1}^\infty2^j\epsilon\exp(-C_2a_n(2^{j-1}\epsilon)^2)(2^j\epsilon)^\alpha\end{aligned}$$

Take $\epsilon=\frac{1}{\sqrt{a_n}}$, we have

$$\begin{aligned}\mathbb{E}|\eta-\delta^*|1(f_n\neq f^*)\leq&a_n^{-\frac{1+\alpha}{2}}C_0\left(1+C_1\sum_{j=1}^\infty(2^j)^{1+\alpha}\exp(-C_22^{2j-2})\right)\\ \leq&C_3a_n^{-\frac{1+\alpha}{2}}.\end{aligned}$$

□

Proof of Lemma 5.1. Denote $C_G = \nabla\mathcal{G}(\boldsymbol{C}^*)^T(1,-1,-1,1)^T$. By Assumption 1, $C_G > 0$. Also denote $C_H = \max_f \|\nabla^2\mathcal{G}(\boldsymbol{C}(f))\|_{op} > 0$. C_H is a constant because of Assumption 1 and the fact that $[0,1]^4$ is compact.

By the Taylor expansion around $\boldsymbol{C}^*$, there exists $\tilde{\boldsymbol{C}} = \alpha\boldsymbol{C}^* + (1-\alpha)\boldsymbol{C}_n$ for some $\alpha \in [0,1]$, such that

$$\mathcal{G}(\boldsymbol{C}^*) - \mathcal{G}(\boldsymbol{C}_n) = \nabla\mathcal{G}(\boldsymbol{C}^*)^T(\boldsymbol{C}^* - \boldsymbol{C}_n) + (\boldsymbol{C}^* - \boldsymbol{C}_n)^T\nabla^2\mathcal{G}(\tilde{\boldsymbol{C}})(\boldsymbol{C}^* - \boldsymbol{C}_n)$$

Expanding the first term, we have

$$\begin{aligned}
&\nabla\mathcal{G}(\boldsymbol{C}^*)^T(\boldsymbol{C}^* - \boldsymbol{C}_n)\\
=&g_1(\boldsymbol{C}^*)(\mathrm{TP}(f^*) - \mathrm{TP}(f_n)) + g_2(\boldsymbol{C}^*)(\mathrm{FP}(f^*) - \mathrm{FP}(f_n))\\
&\quad + g_3(\boldsymbol{C}^*)(\mathrm{FN}(f^*) - \mathrm{FN}(f_n)) + g_4(\boldsymbol{C}^*)(\mathrm{TN}(f^*) - \mathrm{TN}(f_n))\\
=&(g_1(\boldsymbol{C}^*) - g_3(\boldsymbol{C}^*))(P(Y=1, f^*=1, f_n=-1) - P(Y=1, f^*=-1, f_n=1))\\
&\quad + (g_4(\boldsymbol{C}^*) - g_2(\boldsymbol{C}^*))(P(Y=-1, f^*=-1, f_n=1) - P(Y=-1, f^*=-1, f_n=1)) \qquad \text{(C.2)}
\end{aligned}$$

Recall by Theorem 3.2, $\frac{g_4(\boldsymbol{C}^*) - g_2(\boldsymbol{C}^*)}{g_4(\boldsymbol{C}^*) - g_2(\boldsymbol{C}^*) + g_1(\boldsymbol{C}^*) - g_3(\boldsymbol{C}^*)} = \delta^*$, and $f^*(x) = \mathrm{sign}\,(\eta(x) - \delta^*)$. Hence Eq. (C.2) further equals to

$$\begin{aligned}
&(g_1(\boldsymbol{C}^*) - g_3(\boldsymbol{C}^*) + g_4(\boldsymbol{C}^*) - g_2(\boldsymbol{C}^*))[(1-\delta^*)(\mathbb{E}\eta 1(f^*=1, f_n=-1) - \mathbb{E}\eta 1(f^*=-1, f_n=1))\\
&\quad + \delta^*(\mathbb{E}(1-\eta)1(f^*=-1, f_n=1) - \mathbb{E}(1-\eta)1(f^*=-1, f_n=1))]\\
=&(g_1(\boldsymbol{C}^*) - g_3(\boldsymbol{C}^*) + g_4(\boldsymbol{C}^*) - g_2(\boldsymbol{C}^*))[\mathbb{E}(\eta-\delta^*)1(\eta>\delta^*, f_n=-1) + \mathbb{E}(-\eta+\delta^*)1(\eta<\delta^*, f_n=1)]\\
=&C_G\mathbb{E}|\eta - \delta^*|1(f_n \neq f^*) \qquad \text{(C.3)}
\end{aligned}$$

For the second order term, note $\mathrm{TP}(f^*) - \mathrm{TP}(f_n) = \mathrm{FN}(f_n) - \mathrm{FN}(f^*)$ and $\mathrm{TN}(f^*) - \mathrm{TN}(f_n) = \mathrm{FP}(f_n) - \mathrm{FP}(f^*)$, we have

$$\begin{aligned}
|(\boldsymbol{C}^* - \boldsymbol{C}_n)^T\nabla^2\mathcal{G}(\tilde{\boldsymbol{C}})(\boldsymbol{C}^* - \boldsymbol{C}_n)| \leq&\|\nabla^2\mathcal{G}(\tilde{\boldsymbol{C}})\|_{op}\cdot\|\boldsymbol{C}^* - \boldsymbol{C}_n\|^2\\
\leq&4\|\nabla^2\mathcal{G}(\tilde{\boldsymbol{C}})\|_{op}\cdot((\mathrm{TP}(f^*) - \mathrm{TP}(f_n))^2 + (\mathrm{TN}(f^*) - \mathrm{TN}(f_n))^2)\\
\leq&\frac{4C_H}{\min\{\delta^*, 1-\delta^*\}}\cdot((1-\delta^*)^2(\mathrm{TP}(f^*) - \mathrm{TP}(f_n))^2 + {\delta^*}^2(\mathrm{TN}(f^*) - \mathrm{TN}(f_n))^2)\\
\leq&\frac{4C_H}{\min\{\delta^*, 1-\delta^*\}}(\mathbb{E}|\eta - \delta^*|1(f_n \neq f^*))^2 \qquad \text{(C.4)}
\end{aligned}$$

By Lemma C.2, when $a_n \geq \left(\frac{8C_H}{C_3C_G\min\{\delta^*, 1-\delta^*\}}\right)^2$, we have

$$\mathbb{E}|\eta - \delta^*|1(f_n \neq f^*) \leq \frac{C_G\min\{\delta^*, 1-\delta^*\}}{8C_H} \qquad \text{(C.5)}$$

Combining Eq. (C.5) with Eqs. (C.3) and (C.4) completes the proof. □

Proof of Lemma 5.2. The proof of Lemma 5.2 can be obtained by combining Lemma 5.1 and Lemma C.2. □

C.1 Convergence of confusion matrix with fixed threshold

In this section we prove the following lemma, which bounds the norm of difference of confusion matrices when two classifiers share the same threshold.

Lemma C.3. *Let $\hat{\eta}(\cdot)$ be an estimator of $\eta(\cdot)$. For any $\delta \in (0,1)$, define $\hat{f} = sign\,(\hat{\eta} - \delta)$ and $f = sign\,(\eta - \delta)$. Then with constant $c_8 > 0$,*

$$\|\boldsymbol{C}(f, \mathbb{P}) - \boldsymbol{C}(\hat{f}, \mathbb{P})\| \leq c_8\int|\eta - \hat{\eta}|d\mu.$$

When comparing two threshold-form classifiers with same threshold and different conditional probability estimators, we observe that, the difference in the confusion matrix is bounded by the distance between the conditional probability used in the classifier. We start with several lemmas and the main proof is to be found at the end of this section. To bridge the two classifier, we define the following auxiliary variable, thresholding at which makes the probability of being predicted positive by $\hat{\eta}$ equals to the true probability of being positive. Define δ_{part} as

$$\mathbb{P}(\{X : \hat{\eta}(X) < \delta_{\text{part}}\}) = \mathbb{P}(\{X : \eta(X) < \delta\}). \tag{C.6}$$

Then the thresholding classifier defined by $\hat{\eta}$ and δ_{part} has the following relationship with the ground truth true positive.

Lemma C.4. *For any $\delta \in (0,1)$, conditional probability η and its estimator $\hat{\eta}$, let δ_{part} defined by Eq.* (C.6)*. Then $\left|\int_{\hat{\eta}>\delta_{part}} \hat{\eta} d\mu - \int_{\eta>\delta} \eta(x) d\mu\right| \leq \int_{\mathcal{X}} |\eta - \hat{\eta}| d\mu$.*

Proof. Define a partition based on Eq. (C.6).

$$\begin{aligned} B_1 &= \{X : \hat{\eta}(X) < \delta_{\text{part}}, \eta(X) > \delta\},\ B_2 = \{X : \hat{\eta}(X) > \delta_{\text{part}}, \eta(X) < \delta\} \\ B_3 &= \{X : \hat{\eta}(X) > \delta_{\text{part}}, \eta(X) > \delta\},\ B_4 = \{X : \hat{\eta}(X) < \delta_{\text{part}}, \eta(X) < \delta\} \end{aligned} \tag{C.7}$$

By definition of δ_{part}, $P(B_1 \cup B_3) = P(\eta > \delta) = P(\hat{\eta} > \delta_{\text{part}}) = P(B_2 \cup B_3)$, hence $P(A_1) = P(A_2)$. Now the left hand side can be represented as

$$\begin{aligned} \int_{B_2 \cup B_3} \hat{\eta} d\mu - \int_{B_1 \cup B_3} \eta d\mu &= \int_{B_2} \hat{\eta}(x) d\mu - \int_{B_1} \eta d\mu + \int_{B_3} (\hat{\eta} - \eta) d\mu \\ &\leq \int_{B_2} \hat{\eta}(x) d\mu - \int_{B_2} \eta d\mu + \int_{B_3} (\hat{\eta} - \eta) d\mu \\ &\leq \int_{B_2 \cup B_3} |\hat{\eta} - \eta| d\mu \end{aligned}$$

The first inequality is due to the fact that $\int_{B_1} \eta d\mu \geq \delta P(B_1) = \delta P(B_2) \geq \int_{B_2} \eta d\mu$. On the other hand, if we notice $\int_{B_1} \hat{\eta} d\mu \leq \delta_{\text{part}} P(B_1) = \delta_{\text{part}} P(B_2) \leq \int_{B_2} \hat{\eta} d\mu$, we have

$$\begin{aligned} \int_{B_2 \cup B_3} \hat{\eta} d\mu - \int_{B_1 \cup B_3} \eta d\mu &\geq \int_{B_1} \hat{\eta}(x) d\mu - \int_{B_1} \eta d\mu + \int_{B_3} (\hat{\eta} - \eta) d\mu \\ &\geq - \int_{B_1 \cup B_3} |\hat{\eta} - \eta| d\mu \end{aligned}$$

Combining both sides proves the lemma. □

Another key insight comes from the fact that when the estimation of the conditional probability is good enough, then the probability of thresholding at the same point will be sufficiently close. Formally speaking, let us define two random variables $Z_1 = \eta(X)$ and $Z_2 = \hat{\eta}(X)$. By definition of δ_{part}, $|\mathbb{P}(\hat{\eta} < \delta) - \mathbb{P}(\hat{\eta} < \delta_{\text{part}})| = |\mathbb{P}(\hat{\eta}(X) < \delta) - \mathbb{P}(\eta(X) < \delta)|$. The right hand side can be considered as the cumulative distribution function of Z_1 and Z_2 evaluated at the same point δ, which can be upper bounded by the Kolmogorov-Smirnov distance between the distributions of Z_1 and Z_2. We recall the definition below, let F_1, F_2 be the cumulative density function of Z_1, Z_2 respectively.

$$KS(F_1, F_2) = \sup_{x \in \mathbb{R}} |F_1(x) - F_2(x)|$$

The KS distance has close connection to the distances between characteristic functions. To be explicit, we cite the following lemma.

Lemma C.5 (Ushakov (1999) Theorem 2.9.3)**.** *Let $F_1(x)$ and $F_2(x)$ be two distribution functions with characteristic functions $\phi_1(t)$ and $\phi_2(t)$. Then for any positive T, the following inequality is true.*

$$KS(F_1, F_2) \leq \frac{1}{\pi}\int_{-T}^{T}\left|\frac{\phi_1(t)-\phi_2(t)}{t}\right|dt + \frac{1}{2T}\int_{-T}^{T}(|\phi_1(t)|+|\phi_2(t)|)dt$$

Equipped with Lemma C.5, we get the following.

Lemma C.6. *Consider two conditional probability functions $Z_1 = \eta(X)$ and $Z_2 = \hat{\eta}(X)$ of random variable X, and $\delta \in [0,1]$. If the characteristic functions of Z_1, Z_2 are $\phi_1(t)$ and $\phi_2(t)$ respectively, and ϕ_1, ϕ_2 are both absolutely integrable, then $\forall \delta \in [0,1]$,*

$$|\mathbb{P}(\eta(X) > \delta) - \mathbb{P}(\hat{\eta}(X) > \delta)| \leq C\int |\eta - \hat{\eta}| d\mu.$$

Proof. When $\phi_1(t), \phi_2(t)$ are absolutely integrable, $\lim_{T\to\infty}\int_{-T}^{T}(|\phi_1(t)|+|\phi_2(t)|)dt < \infty$. So by Lemma C.5 as $T \to \infty$,

$$\begin{aligned}
KS(F_1,F_2) \leq & \lim_{T\to\infty}\frac{1}{\pi}\int_{-T}^{T}\left|\frac{\phi_1(t)-\phi_2(t)}{t}\right|dt \\
\leq & \lim_{T\to\infty}\frac{1}{\pi}\int_{-T}^{T}\left|\frac{\mathbb{E}(e^{it\eta(X)})-\mathbb{E}(e^{it\hat{\eta}(X)})}{t}\right|dt \\
\leq & \lim_{T\to\infty}\frac{1}{\pi}\int_{-T}^{T}\mathbb{E}\left[\frac{|\cos(t\eta)-\cos(t\hat{\eta})+i\sin(t\eta)-\sin(t\hat{\eta})|}{t}\right]dt \\
\leq & \mathbb{E}\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{|\cos(t\eta)-\cos(t\hat{\eta})|+|\sin(t\eta)-\sin(t\hat{\eta})|}{t}dt \\
= & \frac{1}{\pi}\mathbb{E}\int_{-\infty}^{\infty}\frac{|2\sin(t(\eta-\hat{\eta})/2)\sin(t(\eta+\hat{\eta})/2)+2|\sin(t(\eta-\hat{\eta})/2)\cos(t(\eta+\hat{\eta})/2)|}{t}dt \\
\leq & \frac{4}{\pi}\mathbb{E}\int_{-\infty}^{\infty}\frac{|\sin(t(\eta-\hat{\eta})/2)|}{t}dt \\
\leq & C\int|\eta-\hat{\eta}|d\mu
\end{aligned}$$

The equality is built on the trigonometric identities. The inequality before that is due to the fact that characteristic function and KS distance are both finite, so we can change the order of integration and expectation. □

Now we are in position to prove Lemma C.3.

Proof of Lemma C.3. It suffices to show that $|\mathrm{TP}(\mathrm{sign}\,(\eta(x)-\delta)) - \mathrm{TP}(\mathrm{sign}\,(\hat{\eta}(x)-\delta))| \leq C\int|\hat{\eta}-\eta|d\mu$, the upper bound for other entries can be shown similarly. Note here that in calculating $\mathrm{TP}(\mathrm{sign}\,(\hat{\eta}(x)-\delta))$ with

the estimated $\hat{\eta}$, the integrand is still with respect to the true η. Expanding both TP as integrals we have,

$$\begin{aligned}
&|\mathrm{TP}(\mathrm{sign}\,(\eta(x)-\delta)) - \mathrm{TP}(\mathrm{sign}\,(\hat{\eta}(x)-\delta))| \\
=&\left|\int_{\hat{\eta}>\delta}\eta(x)d\mu - \int_{\eta>\delta}\eta(x)d\mu\right| \\
\leq&\left|\int_{\hat{\eta}>\delta_{\mathrm{part}}}\eta(x)d\mu - \int_{\eta>\delta}\eta(x)d\mu\right| + \left|\int_{\min(\delta,\delta_{\mathrm{part}})<\hat{\eta}<\max(\delta,\delta_{\mathrm{part}})}\eta(x)d\mu\right| \\
\leq&\left|\int_{\hat{\eta}>\delta_{\mathrm{part}}}\eta(x)d\mu - \int_{\eta>\delta}\eta(x)d\mu\right| + |\mathbb{P}(\hat{\eta}<\delta) - \mathbb{P}(\hat{\eta}<\delta_{\mathrm{part}})| \\
\leq&\left|\int_{\hat{\eta}>\delta_{\mathrm{part}}}\hat{\eta}d\mu - \int_{\eta>\delta}\eta(x)d\mu\right| + \int_{\hat{\eta}>\delta_{\mathrm{part}}}|\hat{\eta}-\eta|d\mu + |\mathbb{P}(\hat{\eta}<\delta) - \mathbb{P}(\hat{\eta}<\delta_{\mathrm{part}})|
\end{aligned} \tag{C.8}$$

The first quantity in the last line is bounded by Lemma C.4, the second term is immediately upper bounded by $\int|\eta-\hat{\eta}|d\mu$, and the third term is bounded by Lemma C.6. Hence we have

$$\|\boldsymbol{C}(\mathrm{sign}\,(\eta(x)-\delta^*),\mathbb{P}) - \boldsymbol{C}(\mathrm{sign}\,(\hat{\eta}(x)-\delta^*),\mathbb{P})\| \leq C\int|\eta-\hat{\eta}|d\mu$$

□

C.2 Convergence rate of binary search

The proof of Lemma 5.3 relies on the following general lemma.

Lemma C.7. *If $\mathcal{V}_\eta(\delta,\mathbb{P})$ has a unique maximizer δ^*. Let $\hat{\delta}$ be the output of Algorithm 2 with input conditional probability function $\hat{\eta}$. If there exists constants L_v, c_1, γ such that*

1. *$\mathcal{V}_{\hat{\eta}}(\delta,\mathbb{P})$ is L_v-Lipschitz continuous for any $\hat{\eta}$ satisfying Assumption 2;*
2. *If $|\delta-\delta^*|\geq\gamma$, then $\mathrm{sign}\left(\mathcal{V}'_\eta(\delta,\mathbb{P})\right) = \mathrm{sign}\left(\mathcal{V}'_{\hat{\eta}}(\delta,\mathbb{P}_n)\right)$.*
3. *If $|\delta-\delta^*|<\gamma$, then $|\mathcal{V}'_\eta(\delta,\mathbb{P}_n)|\leq c_1\gamma$.*

Then with the tuning parameter ϵ_0 in Algorithm 2 set as $\epsilon_0 = c_1\gamma^2/L_v$, we have

$$\mathcal{U}(\mathit{sign}\,(\hat{\eta}(x)-\delta^*),\mathbb{P}) - \mathcal{U}(\mathit{sign}\left(\hat{\eta}(x)-\hat{\delta}\right),\mathbb{P}) \leq 3c_1\gamma^2$$

The intuition for the conditions is straightforward. We hope the empirical evaluated sign to be equal to the population value of the sign when the threshold is far from the optimum, hence the update has the correct direction. And the objective function is also smooth enough around the optimum, so that the cumulative difference can be bounded.

Proof of Lemma 5.3. It suffices to check the three conditions in Lemma C.7. Recall $v(\delta) = (-\delta, -(1-\delta), \delta, 1-\delta)$.

1. $\mathcal{V}_\eta(\delta,\mathbb{P})$ is Lipschitz continuous First by Assumption 2, $p_\eta(\cdot)$ is bounded everywhere. If $\mathcal{G}(\boldsymbol{C}) = \frac{\boldsymbol{a}^T\boldsymbol{C}}{\boldsymbol{b}^T\boldsymbol{C}}$, then $\mathcal{V}'_{\hat{\eta}}(\delta,\mathbb{P}) = \frac{(\boldsymbol{b}^T\boldsymbol{C})(\boldsymbol{a}^Tv(\delta))-(\boldsymbol{a}^T\boldsymbol{C})(\boldsymbol{b}^Tv(\delta))}{(\boldsymbol{b}^T\boldsymbol{C})^2}\cdot p_\eta(\delta)$. Note $|\boldsymbol{b}^T\boldsymbol{C}|>0$, and $\boldsymbol{C}$ is on a compact set, hence there exists constant c such that $|\boldsymbol{b}^T\boldsymbol{C}|\geq c$. Therefore, $|\mathcal{V}'_{\hat{\eta}}(\delta,\mathbb{P})|$ is bounded by some constant L_v.

If $\mathcal{G}$ is concave and twice continuously differentiable, $\mathcal{V}'_{\hat{\eta}}(\delta,\mathbb{P}) = \nabla\mathcal{G}(\boldsymbol{C})^Tv(\delta)p_\eta(\delta)$. Due to the fact that $\nabla\mathcal{G}(\boldsymbol{C})$ is continuous, we know it has an upper bound on confusion matrix domain $[0,1]^4$. Hence there exists constant L_v such that $\mathcal{V}_{\hat{\eta}}(\delta,\mathbb{P})$ is L_v-Lipschitz continuous.

2. $\Phi_\eta(\delta, \mathbb{P}) = \Phi_{\hat{\eta}}(\delta, \mathbb{P}_n)$ **when** $|\delta - \delta^*| \geq \gamma$ Recall the function H defined in the proof of Corollary 3.1. $\Phi_\eta(\delta, \mathbb{P}) = \text{sign}\left(\mathcal{V}'_\eta(\delta, \mathbb{P})\right) = \text{sign}\left(H_\eta(\delta, \mathbb{P})\right)$. Hence it suffices to show that $|H_\eta(\delta, \mathbb{P})| \geq |H_\eta(\delta, \mathbb{P}) - H_{\hat{\eta}}(\delta, \mathbb{P}_n)|$.

Given $\hat{\eta}$ and δ fixed, we have $\mathbb{E}_X 1(\hat{\eta}(X) > \delta)\eta(X) = \text{TP}(\text{sign}\left(\hat{\eta}(X) - \delta\right), \mathbb{P})$. Note that $1(\hat{\eta}(X) > \delta)\eta(X) - \mathbb{E}1(\hat{\eta}(X) > \delta)\eta(X)$ is mean zero and absolutely bounded by 2. Hence by Bernstein's inequality, for $\epsilon > 0$,

$$P(|\text{TP}(\text{sign}\left(\hat{\eta}(x) - \delta^*\right), \mathbb{P}_n) - \text{TP}(\text{sign}\left(\hat{\eta}(x) - \delta^*\right), \mathbb{P})| > \epsilon) \leq C_1 \exp(-C_2 n\epsilon^2) \tag{C.9}$$

Similarly we get that for the entire confusion matrix, with probability greater than $1 - n^{-c_1}$,

$$\|\boldsymbol{C}(\text{sign}\left(\hat{\eta}(X) - \delta\right), \mathbb{P}) - \boldsymbol{C}(\text{sign}\left(\hat{\eta}(X) - \delta\right), \mathbb{P}_n)\| \leq c_2 \sqrt{\frac{\log n}{n}}$$

By similar reasoning as the above paragraph, there exists a constant c_3 such that

$$|H_{\hat{\eta}}(\delta, \mathbb{P}) - H_{\hat{\eta}}(\delta, \mathbb{P}_n)| \leq c_3 \|\boldsymbol{C}(\text{sign}\left(\hat{\eta} - \delta\right), \mathbb{P}) - \boldsymbol{C}(\text{sign}\left(\hat{\eta} - \delta\right), \mathbb{P}_n)\| \tag{C.10}$$

for both linear fractional functions and concave functions. Hence, there exists $c_4 \in \mathbb{R}$ such that

$$P\left(|H_{\hat{\eta}}(\delta, \mathbb{P}) - H_{\hat{\eta}}(\delta, \mathbb{P}_n)| \geq c_4 \sqrt{\frac{\log n}{n}}\right) \leq n^{-1} \tag{C.11}$$

On the other hand, by Lemma C.3, we have

$$\|\boldsymbol{C}(\text{sign}\left(\hat{\eta}(X) - \delta\right), \mathbb{P}) - \boldsymbol{C}(\text{sign}\left(\eta(X) - \delta\right), \mathbb{P})\| \leq c_5 \mathbb{E}_X |\hat{\eta} - \eta|$$

Hence by Lipschitzness of H and Assumption 4, we have with probability greater than $1 - a_n^{-1}$,

$$|H_{\hat{\eta}}(\delta) - H_\eta(\delta)| \leq c_5 \sqrt{\frac{\log a_n}{a_n}} \tag{C.12}$$

For concave functions, by Eq. (A.10), $|H'_\eta(\delta, \mathbb{P})| \geq C_B$. For linear fractional functions, by working out the derivative of H, we know it is bounded from below. There exists $c_6 > 0$, such that $\forall |\delta - \delta^*| \geq \gamma$, $|H_\eta(\delta)| \geq c_6 \gamma$. For $\gamma \geq \frac{c_4}{c_6}\sqrt{\frac{\log n}{n}} + \frac{c_5}{c_6}\sqrt{\frac{\log a_n}{a_n}})$, by Eqs. (C.11) and (C.12),

$$|H_\eta(\delta, \mathbb{P})| \geq |H_\eta(\delta, \mathbb{P}) - H_{\hat{\eta}}(\delta, \mathbb{P}_n)|$$

This implies all the empirically evaluated signs for $|\delta - \delta^*| \geq \gamma$ are correct, i.e., $\Phi_\eta(\delta, \mathbb{P}) = \Phi_{\hat{\eta}}(\delta, \mathbb{P}_n)$.

3. Smoothness around the optimum

$$|\mathcal{V}'_{\hat{\eta}}(\delta, \mathbb{P})| \leq |H_{\hat{\eta}}(\delta, \mathbb{P})| \cdot \max_\delta p_{\hat{\eta}}(\delta)$$

The gradient of H is a continuous function defined on a compact domain, and is bounded. Hence $H_\eta(\delta, \mathbb{P})$ is ℓ_0-Lipschitz continuous. We know that $H_\eta(\delta^*, \mathbb{P}) = 0$. So by Eq. (C.12),

$$\begin{aligned} H_{\hat{\eta}}(\delta, \mathbb{P}) \leq & |H_{\hat{\eta}}(\delta, \mathbb{P}) - H_\eta(\delta, \mathbb{P})| + |H_\eta(\delta, \mathbb{P})| \\ \leq & c_5 \sqrt{\frac{\log a_n}{a_n}} + \ell_0 \gamma \end{aligned}$$

Meanwhile, $\max_\delta p_{\hat{\eta}}(\delta)$ is bounded by Assumption 2. Take $\gamma = c_7 \frac{\log \tilde{n}}{\tilde{n}}$ with large enough constant c_7, all three conditions in Lemma C.7 are satisfied.

□

C.3 Rate of convergence for the two-step plug-in classifier with binary search

Proof of Theorem 5.1. Assume the classifier returned is $\hat{f}(x) = \text{sign}\,(\hat{\eta}(x) - \delta)$, where $\hat{\eta}$ is learned with the first n_1 samples $\{X_i^{(1)}, Y_i^{(1)}\}$, and δ is returned by Algorithm 2 with n_2 samples. By Lemma 5.2,

$$|\mathcal{U}(f^*, \mathbb{P}) - \mathcal{U}(\hat{\eta} - \delta^*, \mathbb{P})| \leq C_1 a_{n_1}^{-\frac{1+\alpha}{2}}$$

By Assumption 6 and Lemma 5.3, with probability at least $1 - \min\{a_{n_2}, n_2\}^{-c}$,

$$|\mathcal{U}(\hat{\eta} - \delta^*, \mathbb{P}) - \mathcal{U}(\hat{\eta} - \hat{\delta}, \mathbb{P})| \leq C_2 \frac{\log(\min\{a_{n_2}, n_2\})}{\min\{a_{n_2}, n_2\}}$$

Taking $n_1 = n_2 = \frac{n}{2}$, we have with probability at least $1 - \min\{a_n, n\}^{-c}$,

$$\begin{aligned}
\mathcal{U}(f^*, \mathbb{P}) - \mathcal{U}(\hat{f}, \mathbb{P}) \leq & |\mathcal{U}(\text{sign}\,(\eta - \delta^*), \mathbb{P}) - \mathcal{U}(\hat{\eta} - \delta^*, \mathbb{P})| + |\mathcal{U}(\hat{\eta} - \delta^*, \mathbb{P}) - \mathcal{U}(\hat{\eta} - \hat{\delta}, \mathbb{P})| \\
\leq & C_1 a_{n_1}^{-\frac{1+\alpha}{2}} + C_2 \max\left\{\frac{\log a_{n_2}}{a_{n_2}}, \frac{\log n_2}{n_2}\right\} \\
\leq & C_3 \max\left\{\frac{\log n}{n}, \frac{\log a_n}{a_n}, a_n^{-\frac{1+\alpha}{2}}\right\}
\end{aligned}$$

□

D Examples

In this section we present the proofs in Section 6.

D.1 Gaussian generative model

The following Lemma shows the asymmetric Gaussian generative model can be fitted via a logistic regression with coefficient μ and intercept $\log\left(\frac{\kappa}{1-\kappa}\right)$.

Lemma D.1. *Let $\mathbb{P}_{\mu,\kappa}$ be defined in Eq. (5). Denote $\gamma = \log\left(\frac{\kappa}{1-\kappa}\right)$. Then $\eta(X) = \frac{\exp(X^T\mu+\gamma)}{1+\exp(X^T\mu+\gamma)}$.*

Proof. By Bayes rule,

$$\begin{aligned}
\eta(X) = & P(Y = 1|X) \\
= & \frac{\kappa\phi(X - \frac{\mu}{2})}{\kappa\phi(X - \frac{\mu}{2}) + (1-\kappa)\phi(X + \frac{\mu}{2})} \\
= & \frac{\kappa \exp(\frac{X^T\mu}{2})}{\kappa \exp(\frac{1}{2}X^T\mu) + (1-\kappa)\exp(-\frac{1}{2}X^T\mu)} \\
= & \frac{\exp(X^T\mu + \gamma)}{1 + \exp(X^T\mu + \gamma)}
\end{aligned}$$

□

Lemma D.2. *Let $\mathbb{P}_{\mu,\kappa}$ as defined in Eq. (5) and $\ell(x) = \frac{e^x}{1+e^x}$. If $\|\mu\| = 2w$, then for any $\theta, \theta' \in \mathbb{R}^d$, we have*

$$\mathbb{E}_X|\ell(X\theta) - \ell(X\theta')| \leq w\|\theta - \theta'\|.$$

Proof of Lemma D.2. We know that $\ell(x)$ is Lipschitz continuous, since $\ell'(x) = \frac{e^x}{(1+e^x)^2} \leq 1$. Hence, for any $x \in \mathcal{X}$ and θ, θ' we have

$$|\ell(X\theta) - \ell(X\theta')| \leq \|X\| \cdot \|\theta - \theta'\|$$

By the assumption of the lemma we have $\mathbb{E}\|X\| \leq \frac{\|\mu\|}{2} < \infty$, hence

$$\mathbb{E}_X|\ell(X\theta) - \ell(X\theta')| \leq w\|\theta - \theta'\|$$

The claim is proved. □

Lemma D.2 states that in this generalized linear model, the convergence rate of the conditional probability is the same as convergence rate for parameter estimation.

Proof of Lemma 6.1. The consistency and asymptotic normality of the maximum likelihood estimator (MLE) of the parameters in generalized linear model has been well studied (e.g. see Fahrmeir and Kaufmann (1985)). With sample size n, under regularity conditions, the MLE converges in distribution to a Gaussian distribution whose mean is the true parameter and covariance matrix is the inverse of the Fisher information matrix.

$$\begin{aligned}
&\mathrm{Cov}(\hat{\theta}_{\mathrm{MLE}}) = X^T D X, \\
&d_{ii} = \frac{\exp(X_i\theta)}{(1+\exp(X_i\theta))^2}, \forall i \in [n]; \quad d_{ij} = 0, \forall i \neq j.
\end{aligned}$$

By the asymptotic normality, together with Lemma D.2, it is clear that $a_n = n$ in Assumption 4. □

Proof of Lemma 6.2. By Lemma D.1,

$$\begin{aligned}
\eta(X) - \delta =& \frac{\kappa \exp(\frac{X^T\mu}{2})}{\kappa \exp(\frac{X^T\mu}{2}) + (1-\kappa)\exp(-\frac{X^T\mu}{2})} - \delta \\
=& \frac{\kappa(1-\delta)\exp(\frac{X^T\mu}{2}) - \delta(1-\kappa)\exp(-\frac{X^T\mu}{2})}{\kappa \exp(\frac{X^T\mu}{2}) + (1-\kappa)\exp(-\frac{X^T\mu}{2})} \\
=& \frac{\kappa(1-\delta)\exp(X^T\mu) - \delta(1-\kappa)}{\kappa \exp(X^T\mu) + (1-\kappa)}
\end{aligned} \tag{D.1}$$

Hence for $t < \delta$, the margin probability

$$P_{X\sim\mathbb{P}_{\mu,\kappa}}(0 < |\eta(X) - \delta| < t) \leq P_{X\sim\mathbb{P}_{\mu,\kappa}}\left(\frac{1-\kappa}{\kappa} \cdot \frac{\delta - t}{1-\delta+t} < X^T\mu < \frac{1-\kappa}{\kappa} \cdot \frac{\delta+t}{1-\delta-t}\right)$$

Note the pdf of $X^T\mu$ is upper bounded by $\frac{1}{\sqrt{2\pi}w}$, we have

$$P_{X\sim\mathbb{P}_{\mu,\kappa}}(0 < |\eta(X) - \delta| < t) \leq \frac{2t}{(1-\delta+t)(1-\delta-t)} \cdot \frac{1-\kappa}{\sqrt{2\pi}\kappa w}$$

With $t < \min\{\delta, \frac{1-\delta}{2}\}$,

$$P_{X\sim\mathbb{P}_{\mu,\kappa}}(0 < |\eta(X) - \delta| < t) \leq \frac{t}{2(1-\delta)^2} \cdot \frac{1-\kappa}{\sqrt{2\pi}\kappa w}$$

Therefore there exists a large enough $C_0(\delta)$ such that $P_{X\sim\mathbb{P}_{\mu,\kappa}}(0 < |\eta(X) - \delta| < t) \leq C_0(\delta)t$ holds for $\forall t \in (0, 1]$. This, by definition, is the margin assumption with $\alpha = 1$. □

D.2 Non-parametric conditional probability estimators

The following two lemmas summarized the convergence rates for estimation β-Hölder class function with locally polynomial and kernel density estimators.

Lemma D.3 (Theorem 3.2 in Audibert et al. (2007))**.** *If a probability distribution $\mathbb{P}$ has a conditional probability that belongs to the β-Hölder family, and the marginal law of X satisfies $\mu_{\min} \leq \mu(x) \leq \mu_{\max}, \forall x \in supp(\mu)$. Then there exists constants $C_1, C_2, C_3 > 0$ such that for $h = n^{-1/(2\beta+d)}$, any $t > 0, n \geq 1$, the estimator η_n satisfies*

$$\sup_{\mathcal{S}} P(|\eta_n(x) - \eta(x)| \geq t) \leq C_1 \exp(-C_2 n^{2\beta/(2\beta+d)} t^2)$$

for almost all x.

Lemma D.4 (Theorem 5 in Jiang (2017))**.** *If a conditional probability η belongs to the β-Hölder class ($0 \leq \beta < 1$) is also bounded, spherically symmetric, and non-increasing and exponentially decay in terms of the norm, then there exists constants $C > 0$ such that the following holds with probability at least $1 - C_1/n$ for $h = n^{-1/(2\beta+d)}$,*

$$\sup_{x} \|\eta_n(x) - \eta(x)\|_\infty \leq C n^{-\frac{\beta}{2\beta+d}} \sqrt{\log n}$$

Note that Lemma D.3 and Lemma D.4 have different assumptions hence may be applied on different problems. Lemma D.3 makes assumption on the marginal distribution of X and Lemma D.4 assumes $\beta < 1$ and do not generalize to smoother classes.

Proof of Lemma 6.3. The proof follows directly from Lemma D.3 and Lemma D.4. □

References

Audibert, J.-Y., Tsybakov, A. B., et al. (2007). Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, pages 342–368.

Jiang, H. (2017). Uniform convergence rates for kernel density estimation. In *International Conference on Machine Learning*, pages 1694–1703.

Ushakov, N. G. (1999). *Selected topics in characteristic functions*. Walter de Gruyter.