Stability Enhanced Large-Margin Classifier Selection

Supplementary Materials

Will Wei Sun[1], Guang Cheng[2], Yufeng Liu[3]

In this online supplementary note, we provide all proofs, discuss the calculation of the transformation matrix, and provide a notation table.

S.1. Proof of Theorem 1:

Before we prove Theorem 1, we show an intermediate result in Lemma 2.

Lemma 2 *Suppose Assumptions (L1)–(L3) hold and* $\lambda_n = o(n^{-1/2})$. *Then we have* $\widehat{\boldsymbol{\theta}}_L \xrightarrow{P} \boldsymbol{\theta}_{0L}$ *and* $\widehat{D}_L \xrightarrow{P} D_{0L}$ *as* $n \to \infty$.

To show $\widehat{\boldsymbol{\theta}}_L \to \boldsymbol{\theta}_{0L}$, we apply Theorem 5.7 of van der Vaart (1998). Firstly, we show that, uniformly in $\boldsymbol{\theta}$, the empirical risk $O_{nL}(\boldsymbol{\theta})$ converges to the true risk $\mathcal{R}_L(\boldsymbol{\theta})$ in probability. Assumption (L3) guarantees that the loss function $L(yf(\boldsymbol{x};\boldsymbol{\theta}))$ is convex in $\boldsymbol{\theta}$, and it is easy to see that $O_{nL}(\boldsymbol{\theta})$ converges to $\mathcal{R}_L(\boldsymbol{\theta})$ for each $\boldsymbol{\theta}$. Then we have $\sup_{\boldsymbol{\theta}} |O_{nL}(\boldsymbol{\theta}) - \mathcal{R}_L(\boldsymbol{\theta})| \to 0$ in probability by uniform convergence Theorem for convex functions in Pollard (1991). Secondly, according to assumption (L2), we have that $\mathcal{R}_L(\boldsymbol{\theta})$ has a unique minimizer $\boldsymbol{\theta}_{0L}$. Therefore, we know that $\widehat{\boldsymbol{\theta}}_L$ converges to $\boldsymbol{\theta}_{0L}$ in probability. The consistency of $\widehat{D}(\widehat{\boldsymbol{\theta}}_L)$ can be obtained by the uniform law of large numbers. According to Assumption (L1), $p(\boldsymbol{x})$ is continuously differentiable, and hence $|y - \text{sign}\{f(\boldsymbol{x};\boldsymbol{\theta})\}| = |y - \text{sign}\{\tilde{\boldsymbol{x}}^T\boldsymbol{\theta}\}|$ is continuous in each $\boldsymbol{\theta}$ for almost all $\boldsymbol{x}$. This together with $|y - \text{sign}\{f(\boldsymbol{x};\boldsymbol{\theta})\}| \leq 2$ leads to uniform

[1]Assistant Professor, Department of Management Science, University of Miami, FL 33156, Email: wsun@bus.miami.edu. This work was carried out during Will's PhD period at Purdue University.

[2]Corresponding Author. Professor, Department of Statistics, Purdue University, West Lafayette, IN 47906, Email: chengg@purdue.edu. Partially supported by NSF Grant DMS-0906497, CAREER Award DMS-1151692, DMS-1418042 and Simons Foundation 305266.

[3]Professor, Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina Chapel Hill, NC 27599, Email: yfliu@email.unc.edu. Partially supported by NSF DMS-1407241, NIH/NCI P01 CA-142538, and NSF IIS 1632951.

convergence $\sup_{\boldsymbol{\theta}} |\widehat{D}(\boldsymbol{\theta}) - \frac{1}{2}E|y_0 - \text{sign}\{f(\boldsymbol{x}_0;\boldsymbol{\theta})\}|| \to 0$. Therefore, we have $\widehat{D}(\widehat{\boldsymbol{\theta}}_L) \to D_{0L}$ in probability. This concludes the proof of Lemma 2. ■

Proof of Theorem 1:

We next prove (4) in three steps. Let $M_i(\boldsymbol{\theta}_{0L}) = \nabla_{\boldsymbol{\theta}} L(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0L}}$. In step 1, we show that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_L - \boldsymbol{\theta}_{0L}) = -n^{-1/2} H(\boldsymbol{\theta}_{0L})^{-1} \sum_{i=1}^{n} M_i(\boldsymbol{\theta}_{0L}) + o_P(1) \tag{S.1}$$

by applying Theorem 2.1 in Hjort and Pollard (1993). Denote $Z = (\boldsymbol{X}^T, Y)$ and $\Delta\boldsymbol{\theta} = (\Delta b, \Delta\boldsymbol{w}^T)^T$. Taylor expansion leads to

$$L(Y f(\boldsymbol{X};\boldsymbol{\theta}_{0L} + \Delta\boldsymbol{\theta})) - L(Y f(\boldsymbol{X};\boldsymbol{\theta}_{0L})) = M(\boldsymbol{\theta}_{0L})^T \Delta\boldsymbol{\theta} + R(Z, \Delta\boldsymbol{\theta}), \tag{S.2}$$

where

$$M(\boldsymbol{\theta}_{0L}) = \nabla_{\boldsymbol{\theta}} L(Y f(\boldsymbol{X};\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0L}}; \quad R(Z, \Delta\boldsymbol{\theta}) = \frac{(\Delta\boldsymbol{\theta})^T \Big(\nabla^2_{\boldsymbol{\theta}} L(Y f(\boldsymbol{X};\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0L}} \Big) \Delta\boldsymbol{\theta}}{2} + o(\|\Delta\boldsymbol{\theta}\|^2).$$

According to Assumption (L1), it is easy to check that $E(M(\boldsymbol{\theta}_{0L})) = \nabla_{\boldsymbol{\theta}} \mathcal{R}_L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0L}} = 0$, and

$$E[R(Z, \Delta\boldsymbol{\theta})] = \frac{1}{2}(\Delta\boldsymbol{\theta})^T H(\boldsymbol{\theta}_{0L})(\Delta\boldsymbol{\theta}) + o(\|\Delta\boldsymbol{\theta}\|^2); \quad E[R^2(Z, \Delta\boldsymbol{\theta})] = o(\|\Delta\boldsymbol{\theta}\|^3).$$

Denote $s = (b_s, \boldsymbol{w}_s^T)^T$, $Z_i = (\boldsymbol{X}_i^T, Y_i)$, and

$$\begin{aligned} A_n(s) \;=\; & \sum_{i=1}^{n} \Big\{ L(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0L} + s/\sqrt{n})) - L(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0L})) \Big\} \\ & + \lambda_n(\boldsymbol{w}_{0L} + \boldsymbol{w}_s/\sqrt{n})^T(\boldsymbol{w}_{0L} + \boldsymbol{w}_s/\sqrt{n}) - \lambda_n \boldsymbol{w}_{0L}^T \boldsymbol{w}_{0L}. \end{aligned}$$

Note that $A_n(s)$ is minimized when $s = \sqrt{n}(\widehat{\boldsymbol{\theta}}_L - \boldsymbol{\theta}_{0L})$ and $nE[R(Z, s/\sqrt{n})] = \frac{1}{2}s^T H(\boldsymbol{\theta}_{0L}) s +$

$o(\|s\|^2)$. Based on the above Taylor expansion (S.2), we have

$$\begin{aligned} A_n(s) &= \sum_{i=1}^{n}\Big\{M_i(\boldsymbol{\theta}_{0L})^T s/\sqrt{n} + R(Z_i, s/\sqrt{n}) - ER(Z_i, s/\sqrt{n})\Big\} + nE[R(Z, s/\sqrt{n})] + \lambda_n \boldsymbol{w}_s^T \boldsymbol{w}_s \\ &= U_n^T s + \frac{1}{2} s^T H(\boldsymbol{\theta}_{0L}) s + o(\|s\|^2) + \sum_{i=1}^{n}\Big\{R(Z_i, s/\sqrt{n}) - ER(Z_i, s/\sqrt{n})\Big\} + \lambda_n \boldsymbol{w}_s^T \boldsymbol{w}_s, \end{aligned}$$

where $U_n = n^{-1/2}\sum_{i=1}^{n} M_i(\boldsymbol{\theta}_{0L})$. Note that $\sum_{i=1}^{n}\{R(Z_i, s/\sqrt{n}) - ER(Z_i, s/\sqrt{n})\} \to 0$, and $\lambda_n \boldsymbol{w}_s^T \boldsymbol{w}_s \to 0$ since $\lambda_n \to 0$ and $\boldsymbol{w}_s$ is bounded. In addition, Hessian matrix $H(\boldsymbol{\theta}_{0L})$ is positive definite due to Assumption (L5). Therefore, we can conclude that (S.1) holds by Theorem 2.1 in Hjort and Pollard (1993).

In step 2, we show that $W_L = \sqrt{n}\{\widehat{D}(\widehat{\boldsymbol{\theta}}_L) - D_{0L}\} \to N(0, E(\psi_1^2))$. As shown in Jiang et al. (2008), the class of functions $\mathcal{G}_{\boldsymbol{\theta}}(\delta) = \Big\{|Y - \text{sign}\{f(\boldsymbol{X};\boldsymbol{\theta})\}| : \|\boldsymbol{\theta} - \boldsymbol{\theta}_{0L}\| \le \delta\Big\}$ is a P-Donsker class for any fixed $0 < \delta < \infty$. This together with (S.1) and consistency of $\widehat{\boldsymbol{\theta}}_L$ implies that

$$\begin{aligned} & \sqrt{n}\Big(\widehat{D}(\widehat{\boldsymbol{\theta}}_L) - D_{0L}\Big) \\ = \;& \sqrt{n}\Big(\widehat{D}(\widehat{\boldsymbol{\theta}}_L) - \widehat{D}(\boldsymbol{\theta}_{0L})\Big) + \sqrt{n}\Big(\widehat{D}(\boldsymbol{\theta}_{0L}) - D_{0L}\Big) \\ \stackrel{d}{=} \;& \sqrt{n}\dot{d}(\boldsymbol{\theta}_{0L})^T(\widehat{\boldsymbol{\theta}}_L - \boldsymbol{\theta}_{0L}) + \sqrt{n}\Big(\widehat{D}(\boldsymbol{\theta}_{0L}) - D_{0L}\Big) \\ \stackrel{d}{=} \;& n^{-1/2}\sum_{i=1}^{n}\Big\{\frac{1}{2}|Y_i - \text{sign}\{f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0L})\}| - D_{0L} - \dot{d}(\boldsymbol{\theta}_{0L})^T H(\boldsymbol{\theta}_{0L})^{-1} M_1(\boldsymbol{\theta}_{0L})\Big\} \\ = \;& n^{-1/2}\sum_{i=1}^{n}\psi_i \xrightarrow{d} N(0, E(\psi_1^2)), \end{aligned}$$

where " $\stackrel{d}{=}$ " means asymptotical equivalence in the distributional sense.

In step 3, the distribution of $\mathcal{W}_L = n^{1/2}\{\widehat{\mathcal{D}}_L - D_{0L}\}$ is asymptotically equivalent to that of W_L as shown in Theorem 3 in Jiang et al. (2008). This concludes the proof of Theorem 1. ■

S.2. Proof of Theorem 2

According to Appendix D in Jiang et al. (2008), we have

$$W_1^* \stackrel{d}{=} n^{-1/2} \sum_{i=1}^{n} \psi_{i1}(G_i - 1) \text{ and } W_2^* \stackrel{d}{=} n^{-1/2} \sum_{i=1}^{n} \psi_{i2}(G_i - 1),$$

where $\psi_{ij} = \frac{1}{2}|Y_i - \text{sign}\{f(\boldsymbol{X}_i; \boldsymbol{\theta}_{0j})\}| - D_{0j} - \dot{d}(\boldsymbol{\theta}_{0j})^T H(\boldsymbol{\theta}_{0j})^{-1} M_i(\boldsymbol{\theta}_{0j})$, for $j = 1, 2$. Recall that " $\stackrel{d}{=}$ " means the distributional equivalence. As shown in Jiang et al. (2008), conditional on the data, W_j^* converges to a normal with mean 0 and variance $n^{-1} \sum_{i=1}^{n} \psi_{ij}^2$ for $j = 1, 2$. Note that

$$W_2^* - W_1^* \stackrel{d}{=} n^{-1/2} \sum_{i=1}^{n} (\psi_{i2} - \psi_{i1})(G_i - 1).$$

Here, $(\psi_{i2} - \psi_{i1})$'s, $i = 1, \ldots, n$, are i.i.d random vectors with $E(\psi_{i2} - \psi_{i1}) = 0$ and $E|\psi_{i2} - \psi_{i1}|^2 < \infty$. Independent of $(\psi_{i2} - \psi_{i1})$, $(G_i - 1)$'s are i.i.d random variables with mean 0 and variance 1. Since $(\psi_{i2} - \psi_{i1})$ depends on the sample $(\boldsymbol{x}_i, y_i)$, Lemma 2.9.5 in van der Vaart and Wellner (1996) implies that, conditional on the data,

$$n^{-1/2} \sum_{i=1}^{n} (\psi_{i2} - \psi_{i1})(G_i - 1) \stackrel{d}{\Longrightarrow} N(0, Var(\psi_{12} - \psi_{11})). \tag{S.3}$$

Next, as shown in Theorem 1, $W_1 \stackrel{d}{=} n^{-1/2} \sum_{i=1}^{n} \psi_{i1}$ and $W_2 \stackrel{d}{=} n^{-1/2} \sum_{i=1}^{n} \psi_{i2}$, therefore,

$$W_2 - W_1 \stackrel{d}{=} n^{-1/2} \sum_{i=1}^{n} (\psi_{i2} - \psi_{i1}) \stackrel{d}{\longrightarrow} N(0, Var(\psi_{12} - \psi_{11})).$$

This together with (S.3) and the asymptotic equivalence of W_L and $\mathcal{W}_L$ (Jiang et al. 2008) lead to the asymptotic equivalence between $\mathcal{W}_{\Delta 12}$ and $W_{\Delta_{12}}^*$, which concludes the proof. ■

S.3. Calculation of the transformation matrix in Section 3.2

Given a d dimensional hyperplane $f(\boldsymbol{x}; \boldsymbol{\theta}) = b + w_1 x_1 + \cdots + w_d x_d = 0$, we aim to find a transformation matrix $R \in \mathbb{R}^{d \times d}$ such that the transformed hyperplane $f(\boldsymbol{x}; \boldsymbol{\theta}^\dagger) = b^\dagger +$

$w_1^\dagger x_1 + \cdots + w_d^\dagger x_d = 0$ is parallel to $\mathcal{X}_1, \ldots, \mathcal{X}_{d-1}$, where $(w_1^\dagger, \cdots, w_d^\dagger)^T = R(w_1, \cdots, w_d)^T$ and $b^\dagger = b$. Here, we implicitly assume that $w_d \neq 0$.

We construct a class of linearly independent vectors spanning the hyperplane:

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ -\frac{w_1}{w_d} \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ -\frac{w_2}{w_d} \end{bmatrix} \quad \cdots \quad \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ -\frac{w_{d-1}}{w_d} \end{bmatrix}.$$

Denote these vectors as v_1, v_2,...,v_{d-1}. Then, by Gram-Schmidt process, we can produce the following orthogonal vectors $\bar{v}_1$, $\bar{v}_2$,..., $\bar{v}_{d-1}$:

$$\begin{aligned} \bar{v}_1 &= v_1, \\ \bar{v}_2 &= v_2 - \tfrac{<v_2, \bar{v}_1>}{<\bar{v}_1, \bar{v}_1>} \bar{v}_1, \\ \bar{v}_{d-1} &= v_{d-1} - \tfrac{<v_{d-1}, \bar{v}_1>}{<\bar{v}_1, \bar{v}_1>} \bar{v}_1 - \cdots - \tfrac{<v_{d-1}, \bar{v}_{d-2}>}{<\bar{v}_{d-2}, \bar{v}_{d-2}>} \bar{v}_{d-2}, \end{aligned}$$

where the inner product $< u, v >= \sum_{i=1}^d u_i v_i$ for $u = (u_1, \ldots, u_d)$ and $v = (v_1, \ldots, v_d)$. Denote $\bar{v}_d = [w_1, \cdots, w_d]^T$, which is orthogonal to every $\bar{v}_i$, $i = 1, \cdots, d-1$ by the above construction. In the end, we normalize $u_i = \bar{v}_i \|\bar{v}_i\|^{-1}$ for $i = 1, \cdots, d$, and define the orthogonal transformation matrix R as $[u_1, \ldots, u_d]^T$. By some elementary calculation, we can verify that that $w_i^\dagger = 0$ for $i = 1, \cdots, d-1$ but $w_d^\dagger \neq 0$ under the above construction. Therefore, the transformed hyperplane $f(\boldsymbol{x}; \boldsymbol{\theta}^\dagger)$ is parallel to $\mathcal{X}_1, \ldots, \mathcal{X}_{d-1}$. ■

S.4. Asymptotic Normality of $\widehat{\boldsymbol{\theta}}_\gamma$ and $\widehat{\mathcal{D}}_\gamma$ for LUM

This section establishes the asymptotic normality of $\widehat{\mathcal{D}}_\gamma$ and $\widehat{\boldsymbol{\theta}}_\gamma$ (with more explicit forms of the asymptotic variances) by verifying the conditions in Theorem 1, i.e., (L1)–(L5). In particular, we provide a set of sufficient conditions for the LUM, i.e., (L1) and (A1) below.

(A1) $\mathrm{Var}(\boldsymbol{X}|Y) \in \mathbb{R}^{d \times d}$ is a positive definite matrix for $Y \in \{1, -1\}$.

Assumption (A1) is needed to guarantee the uniqueness of the true minimizer $\boldsymbol{\theta}_{0\gamma}$. It is worth pointing out that the asymptotic normality of the estimated coefficients for SVM has also been established by Koo et al. (2008) under another set of assumptions.

Corollary 1 *Suppose Assumptions (L1) and (A1) hold and* $\lambda_n = o(n^{-1/2})$. *For each* $\gamma \in [0, 1]$,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_\gamma - \boldsymbol{\theta}_{0\gamma}) \xrightarrow{d} N(0, \Sigma_{0\gamma}) \quad \text{as } n \to \infty, \tag{S.4}$$

where $\Sigma_{0\gamma} = H(\boldsymbol{\theta}_{0\gamma})^{-1} G(\boldsymbol{\theta}_{0\gamma}) H(\boldsymbol{\theta}_{0\gamma})^{-1}$ *with* $G(\boldsymbol{\theta}_{0\gamma})$ *and* $H(\boldsymbol{\theta}_{0\gamma})$ *defined in (S.6) and (S.10) in the supplementary materials.*

In practice, direct estimation of $\Sigma_{0\gamma}$ in (S.4) is difficult because of the involvement of the Dirac delta function; see (S.8) and (S.9) in Section S.5 of the supplementary materials. Instead, we find that the perturbation-based resampling procedure proposed in Stage 1 works well.

Next we establish the asymptotic normality of $\widehat{\mathcal{D}}_\gamma$.

Corollary 2 *Suppose that the assumptions in Corollary 1 hold. We have, as* $n \to \infty$,

$$\sqrt{n}(\widehat{\mathcal{D}}_\gamma - D_{0\gamma}) \xrightarrow{d} N\Big(0, E(\psi_{1\gamma}^2)\Big), \tag{S.5}$$

where $\psi_{1\gamma} = \frac{1}{2}|Y_1 - sign\{f(\boldsymbol{X}_1; \boldsymbol{\theta}_{0\gamma})\}| - D_{0\gamma} - \dot{d}(\boldsymbol{\theta}_{0\gamma})^T H(\boldsymbol{\theta}_{0\gamma})^{-1} M_1(\boldsymbol{\theta}_{0\gamma})$, $\dot{d}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} E(\widehat{D}_\gamma(\boldsymbol{\theta}))$, *and*

$$M_1(\boldsymbol{\theta}_{0\gamma}) = -Y_1 \tilde{\boldsymbol{X}}_1 I_{\{Y_1 f(\boldsymbol{X}_1; \boldsymbol{\theta}_{0\gamma}) < \gamma\}} - \frac{(1-\gamma)^2 Y_1 \tilde{\boldsymbol{X}}_1 I_{\{Y_1 f(\boldsymbol{X}_1; \boldsymbol{\theta}_{0\gamma}) \geq \gamma\}}}{\Big(Y_1 f(\boldsymbol{X}_1; \boldsymbol{\theta}_{0\gamma}) - 2\gamma + 1\Big)^2}.$$

Corollary 2 demonstrates that the K-CV error induced from each LUM loss function yields the desirable asymptotic property under Assumptions (L1) and (A1). It can be applied to justify the perturbation-based resampling procedure for LUM as shown in Theorem 2.

S.5. Proof of Corollary 1

It suffices to show that (A1) and (L1) imply Assumptions (L2)-(L5).

(L2). We first show that the minimizer $\boldsymbol{\theta}_{0\gamma}$ exists for each fixed γ. It is easy to see that $\mathcal{R}_\gamma(\boldsymbol{\theta})$ is continuous w.r.t. $\boldsymbol{\theta}$. We next show that, for any large enough M, the closed set $S(M) = \Big\{\boldsymbol{\theta} \in R^d : \mathcal{R}_\gamma(\boldsymbol{\theta}) \leq M\Big\}$ is bounded. When $yf(\boldsymbol{x}, \boldsymbol{\theta}) < \gamma$, we need to show $S(M) = \Big\{\boldsymbol{\theta} \in R^d : E[1 - Yf(\boldsymbol{X}; \boldsymbol{\theta})] \leq M\Big\}$ is contained in a box around the origin. Denote e_j as the vector with one in the j-th component and zero otherwise. Motivated by Rocha et al. (2009), we can show that, for any M, there exists a $\alpha_{j,M}$ such that any $\boldsymbol{\theta}$ satisfying $| < \boldsymbol{\theta}, e_j > | > \alpha_{j,M}$ leads to $E[(1 - Yf(\boldsymbol{X}; \boldsymbol{\theta})I_{(Yf(\boldsymbol{X};\boldsymbol{\theta})<\gamma)})] > M$. Similarly, when $yf(\boldsymbol{x}, \boldsymbol{\theta}) \geq \gamma$, $S(M)$ is contained in a sphere around the origin, that is, for any M, there exists a σ such that any $\boldsymbol{\theta}$ satisfying $| < \boldsymbol{\theta}, \boldsymbol{\theta} > | > \sigma$ leads to $E[\frac{(1-\gamma)^2}{Yf(\boldsymbol{X};\boldsymbol{\theta})-2\gamma+1} I_{(Yf(\boldsymbol{X};\boldsymbol{\theta})\geq\gamma)})] > M$. These imply the existence of $\boldsymbol{\theta}_{0\gamma}$. The uniqueness of $\boldsymbol{\theta}_{0\gamma}$ is implied by the positive definiteness of Hessian matrix as verified in (L5) below.

(L3). The loss function $L_\gamma(yf(\boldsymbol{x}; \boldsymbol{\theta}))$ is convex by noting that two segments of $L_\gamma(yf(\boldsymbol{x}; \boldsymbol{\theta}))$ are convex, and the sum of convex functions is convex.

(L4). The loss function $L_\gamma(yf(\boldsymbol{x}; \boldsymbol{\theta}))$ is not differentiable only on the set $\{\boldsymbol{x} : \tilde{\boldsymbol{x}}^T\boldsymbol{\theta} = \gamma \text{ or } \tilde{\boldsymbol{x}}^T\boldsymbol{\theta} = -\gamma\}$, which is assumed to be a zero probability event. Therefore, with probability one, it is differentiable with

$$\nabla_{\boldsymbol{\theta}} L_\gamma(yf(\boldsymbol{x}; \boldsymbol{\theta})) = -\tilde{\boldsymbol{x}} y I_{(y\tilde{\boldsymbol{x}}^T\boldsymbol{\theta}<\gamma)} - \frac{(1-\gamma)^2 \tilde{\boldsymbol{x}} y}{(y\tilde{\boldsymbol{x}}^T\boldsymbol{\theta} - 2\gamma + 1)^2} I_{(y\tilde{\boldsymbol{x}}^T\boldsymbol{\theta}\geq\gamma)},$$

and hence

$$\begin{aligned} G(\boldsymbol{\theta}_{0\gamma}) &= E\Big[\nabla_{\boldsymbol{\theta}} L_\gamma(Yf(\boldsymbol{X}; \boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}} L_\gamma(Yf(\boldsymbol{X}; \boldsymbol{\theta}))^T|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0\gamma}}\Big] \\ &= E\Big\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T Y^2 I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}<\gamma)} + \frac{(1-\gamma)^4 \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T Y^2}{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma} - 2\gamma + 1)^4} I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\geq\gamma)}\Big\} \\ &= E\Big\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\Big[p(\boldsymbol{X})A(\boldsymbol{X}, \boldsymbol{\theta}_{0\gamma}) + (1 - p(\boldsymbol{X}))B(\boldsymbol{X}, \boldsymbol{\theta}_{0\gamma})\Big]\Big\}, \end{aligned} \tag{S.6}$$

where $A(\boldsymbol{X}, \boldsymbol{\theta}_{0\gamma})$ and $B(\boldsymbol{X}, \boldsymbol{\theta}_{0\gamma})$ are defined as

$$\begin{aligned} A(\boldsymbol{X}, \boldsymbol{\theta}_{0\gamma}) &= I_{(\tilde{\boldsymbol{X}}^T \boldsymbol{\theta}_{0\gamma} < \gamma)} + \frac{(1-\gamma)^4}{(\tilde{\boldsymbol{X}}^T \boldsymbol{\theta}_{0\gamma} - 2\gamma + 1)^4} I_{(\tilde{\boldsymbol{X}}^T \boldsymbol{\theta}_{0\gamma} \geq \gamma)}; \\ B(\boldsymbol{X}, \boldsymbol{\theta}_{0\gamma}) &= I_{(-\tilde{\boldsymbol{X}}^T \boldsymbol{\theta}_{0\gamma} < \gamma)} + \frac{(1-\gamma)^4}{(\tilde{\boldsymbol{X}}^T \boldsymbol{\theta}_{0\gamma} + 2\gamma - 1)^4} I_{(-\tilde{\boldsymbol{X}}^T \boldsymbol{\theta}_{0\gamma} \geq \gamma)}. \end{aligned}$$

Obviously, $|A(\boldsymbol{X}, \boldsymbol{\theta}_{0\gamma})|$ and $|B(\boldsymbol{X}, \boldsymbol{\theta}_{0\gamma})|$ are both bounded by one. Therefore, $G(\boldsymbol{\theta}_{0\gamma}) < \infty$ based on the moment condition of $\boldsymbol{X}$.

(L5). We prove it in three steps. First, we show the risk $R_\gamma(\boldsymbol{\theta})$ is bounded. For each fixed $\gamma \in [0, 1]$,

$$\begin{aligned} \mathcal{R}_\gamma(\boldsymbol{\theta}) \leq E\Big|L_\gamma(Yf(\boldsymbol{X}; \boldsymbol{\theta}))\Big| &= E\Big|(1 - Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}) I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta} < \gamma)} + \frac{(1-\gamma)^2}{Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta} - 2\gamma + 1} I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta} \geq \gamma)}\Big| \\ &\leq E\Big|(1 - Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}) I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta} < \gamma)}\Big| + E\Big|\frac{(1-\gamma)^2}{Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta} - 2\gamma + 1} I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta} \geq \gamma)}\Big| \\ &\leq E\Big|(1 - Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}) I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta} < 1)}\Big| + |1 - \gamma| < \infty, \end{aligned} \tag{S.7}$$

where the first term in (S.7) was shown to be bounded in Rocha et al. (2009).

Next, we derive the form of Hessian matrix. The moment assumption of $\boldsymbol{x}$ and the inequality $(y\tilde{\boldsymbol{x}}^T\boldsymbol{\theta} - 2\gamma + 1)^2 \leq (1-\gamma)^2$ lead to $E|\nabla_{\boldsymbol{\theta}} L_\gamma(Yf(\boldsymbol{X}; \boldsymbol{\theta}))| \leq E|-\tilde{\boldsymbol{X}}Y| + E|-\tilde{\boldsymbol{X}}Y| \leq 2E|\tilde{\boldsymbol{X}}| < \infty$. Then, dominated convergence theorem implies that $\nabla_{\boldsymbol{\theta}} \mathcal{R}_\gamma(\boldsymbol{\theta}) = E[\nabla_{\boldsymbol{\theta}} L_\gamma(Yf(\boldsymbol{X}; \boldsymbol{\theta}))]$. Hence, the Hessian matrix equals $\nabla_{\boldsymbol{\theta}} E[\nabla_{\boldsymbol{\theta}} L_\gamma(Yf(\boldsymbol{X}; \boldsymbol{\theta}))]$. We next

derive the form of $E[\nabla_{\boldsymbol{\theta}} L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}))]$. Note that

$$\begin{aligned}
E[\nabla_{\boldsymbol{\theta}} L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}))] &= E\Big[-\tilde{\boldsymbol{X}} Y I_{\{Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<\gamma\}} - \frac{(1-\gamma)^2\tilde{\boldsymbol{X}}Y}{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}-2\gamma+1)^2} I_{\{Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}\geq\gamma\}}\Big] \\
&= E\Big\{ I_{\{Y=1\}}\Big[-\tilde{\boldsymbol{X}} I_{\{\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<\gamma\}} - \frac{(1-\gamma)^2\tilde{\boldsymbol{X}}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}-2\gamma+1)^2} I_{\{\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}\geq\gamma\}}\Big] \\
&\qquad + I_{\{Y=-1\}}\Big[\tilde{\boldsymbol{X}} I_{\{-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<\gamma\}} + \frac{(1-\gamma)^2\tilde{\boldsymbol{X}}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}+2\gamma-1)^2} I_{\{-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}\geq\gamma\}}\Big]\Big\} \\
&= E\Big\{ p(\boldsymbol{X})\Big[-\tilde{\boldsymbol{X}} I_{\{\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<\gamma\}} - \frac{(1-\gamma)^2\tilde{\boldsymbol{X}}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}-2\gamma+1)^2} I_{\{\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}\geq\gamma\}}\Big]\Big\} \\
&\qquad + E\Big\{ (1-p(\boldsymbol{X}))\Big[\tilde{\boldsymbol{X}} I_{\{-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<\gamma\}} + \frac{(1-\gamma)^2\tilde{\boldsymbol{X}}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}+2\gamma-1)^2} I_{\{-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}\geq\gamma\}}\Big]\Big\} \\
&= E_1(\boldsymbol{\theta}) + E_2(\boldsymbol{\theta}).
\end{aligned}$$

After tedious algebra, we can show

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} E_1(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0\gamma}} &= E\Big\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T p(\boldsymbol{X}) C(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma})\Big\}, \\
\nabla_{\boldsymbol{\theta}} E_2(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0\gamma}} &= E\Big\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T (1-p(\boldsymbol{X})) D(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma})\Big\},
\end{aligned}$$

where

$$C(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma}) = \delta(\gamma-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}) - \frac{(1-\gamma)^2\delta(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-\gamma)}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-2\gamma+1)^2} + \frac{2(1-\gamma)^2 I_{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\geq\gamma)}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-2\gamma+1)^3}, \quad \text{(S.8)}$$

$$D(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma}) = \delta(\gamma+\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}) - \frac{(1-\gamma)^2\delta(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+\gamma)}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+2\gamma-1)^2} - \frac{2(1-\gamma)^2 I_{(-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\geq\gamma)}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+2\gamma-1)^3}, \quad \text{(S.9)}$$

and $\delta(\cdot)$ is the Dirac delta function. Hence, we can write the Hessian matrix as

$$H(\boldsymbol{\theta}_{0\gamma}) = E\Big\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\Big[p(\boldsymbol{X})C(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma}) + (1-p(\boldsymbol{X}))D(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma})\Big]\Big\}. \quad \text{(S.10)}$$

Finally, we establish the positive definiteness of $H(\boldsymbol{\theta}_{0\gamma})$. We write $H(\boldsymbol{\theta}_{0\gamma}) = R_1(\boldsymbol{\theta}_{0\gamma}) +$

$R_2(\boldsymbol{\theta}_{0\gamma})$ with

$$R_1(\boldsymbol{\theta}_{0\gamma}) = E\Big\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\Big[p(\boldsymbol{X})\delta(\gamma-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma})+(1-p(\boldsymbol{X}))\delta(\gamma+\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma})\Big]\Big\},$$

$$R_2(\boldsymbol{\theta}_{0\gamma}) = (1-\gamma)^2 E\Big\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\Big[(1-p(\boldsymbol{X}))\Big(\frac{\delta(\gamma+\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma})}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+2\gamma-1)^2}-\frac{2I_{(-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\geq\gamma)}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+2\gamma-1)^3}\Big)$$
$$-p(\boldsymbol{X})\Big(\frac{\delta(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma})-\gamma}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-2\gamma+1)^2}-\frac{2I_{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\leq\gamma)}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-2\gamma+1)^3}\Big)\Big]\Big\}.$$

Next we show the positive definiteness of $R_1(\boldsymbol{\theta}_{0\gamma})$. Let $f_{\boldsymbol{x}}$ be the density of $\tilde{\boldsymbol{x}}^T\boldsymbol{\theta}_{0\gamma}$. According to Lemma 9 in Rocha et al. (2009), Assumption $(L1)$ implies that $f_{\boldsymbol{x}}(\gamma)>0$, $f_{\boldsymbol{x}}(-\gamma)>0$, $P(Y=1|\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=\gamma)>0$, and $P(Y=-1|\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=-\gamma)>0$. Note that $R_1(\boldsymbol{\theta}_{0\gamma})$ can be rewritten as

$$\begin{aligned}R_1(\boldsymbol{\theta}_{0\gamma}) &= E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y=1,\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=\gamma\Big]P(Y=1|\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=\gamma)f_{\boldsymbol{X}}(\gamma)\\ &\quad+E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y=-1,\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=-\gamma\Big]P(Y=-1|\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=-\gamma)f_{\boldsymbol{X}}(-\gamma).\end{aligned}$$

In order to show $R_1(\boldsymbol{\theta}_{0\gamma})$ is positive definite, it remains to show that $E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y=1,\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=\gamma\Big]$ or $E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y=-1,\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=-\gamma\Big]$ is strictly positive definite. Rocha et al. (2009) showed that

$$\begin{aligned}E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y,\tilde{\boldsymbol{X}}^T\theta_{0\gamma}=\gamma\Big] &= E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y,\boldsymbol{X}^T v_{\boldsymbol{w}_{0\gamma}}=\frac{\gamma-b_{0\gamma}}{\|\boldsymbol{w}_{0\gamma}\|}\Big]\\ &\succeq \Big(\frac{\gamma-b_{0\gamma}}{\|\boldsymbol{w}_{0\gamma}\|}\Big)^2(v_{\boldsymbol{w}_{0\gamma}}v_{\boldsymbol{w}_{0\gamma}}^T)+Var\Big(\boldsymbol{X}|Y,\boldsymbol{X}^T v_{\boldsymbol{w}_{0\gamma}}=\frac{\gamma-b_{0\gamma}}{\|\boldsymbol{w}_{0\gamma}\|}\Big)\end{aligned}\quad(S.11)$$

where $S_1\succeq S_2$ means S_1-S_2 is positive semi-definite, and $v_{\boldsymbol{w}_{0\gamma}}=\frac{\boldsymbol{w}_{0\gamma}}{\|\boldsymbol{w}_{0\gamma}\|}$. By assumption $(A1)$, $Var(\boldsymbol{X}|Y)$ is non-singular, and hence $Var\Big(\boldsymbol{X}|Y,\boldsymbol{X}^T v_{\boldsymbol{w}_{0\gamma}}=\frac{\gamma-b_{0\gamma}}{\|\boldsymbol{w}_{0\gamma}\|}\Big)$ has rank $(d-1)$. Therefore, the right hand side of (S.11) is strictly positive definite when $\gamma\neq b_{0\gamma}$. Similarly, $E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y,\tilde{\boldsymbol{X}}^T\theta_{0\gamma}=-\gamma\Big]$ is strictly positive definite when $\gamma\neq -b_{0\gamma}$. Therefore, either $E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y=1,\boldsymbol{X}^T\boldsymbol{w}_{0\gamma}+b_{0\gamma}=\gamma\Big]$ or $E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y=-1,\boldsymbol{X}^T\boldsymbol{w}_{0\gamma}+b_{0\gamma}=-\gamma\Big]$ will be

strictly positive definite at $\boldsymbol{\theta}_{0\gamma}$. This leads to the positive definiteness of $R_1(\boldsymbol{\theta}_{0\gamma})$.

In addition, similar argument implies that $R_2(\boldsymbol{\theta}_{0\gamma})$ is positive definite at $\boldsymbol{\theta}_{0\gamma}$. This is due to the fact that $(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+2\gamma-1)^3<0$ when $\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+\gamma\leq 0$, and $(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-2\gamma+1)^3>0$ when $\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-\gamma\geq 0$. Therefore, the Hessian matrix $H(\boldsymbol{\theta}_{0\gamma})$ is strictly positive definite for any $\gamma\in[0,1]$. This concludes the proof of Corollary 1. ■

S.6. Proof of Corollary 2

Following the proof of Theorem 1, we only need to show that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_\gamma-\boldsymbol{\theta}_{0\gamma})=-n^{-1/2}H(\boldsymbol{\theta}_{0\gamma})^{-1}\sum_{i=1}^{n}M_i(\boldsymbol{\theta}_{0\gamma})+o_P(1),$$

where

$$M_i(\boldsymbol{\theta}_{0\gamma})=-Y_i\tilde{\boldsymbol{X}}_i I_{\{Y_if(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})<\gamma\}}-\frac{(1-\gamma)^2Y_i\tilde{\boldsymbol{X}}_i I_{\{Y_if(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})\geq\gamma\}}}{\Big(Y_if(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})-2\gamma+1\Big)^2}.$$

Similarly, we denote $Z=(\boldsymbol{X}^T,Y)$ and $t=(b_t,\boldsymbol{w}_t^T)^T$, and write

$$\begin{aligned}
&L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}_{0\gamma}+t))-L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}_{0\gamma}))\\
&=(1-Y\tilde{\boldsymbol{X}}^T(\boldsymbol{\theta}_{0\gamma}+t))I_{\{Y\tilde{\boldsymbol{X}}^T(\boldsymbol{\theta}_{0\gamma}+t)<\gamma\}}+\frac{(1-\gamma)^2}{Y\tilde{\boldsymbol{X}}^T(\boldsymbol{\theta}_{0\gamma}+t)-2\gamma+1}I_{\{Y\tilde{\boldsymbol{X}}^T(\boldsymbol{\theta}_{0\gamma}+t)\geq\gamma\}}\\
&\quad-(1-Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma})I_{\{Y\widetilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}<\gamma\}}-\frac{(1-\gamma)^2}{Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-2\gamma+1}I_{\{Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\geq\gamma\}}\\
&=M(\boldsymbol{\theta}_{0\gamma})^Tt+R(Z,t),
\end{aligned}$$

where

$$
\begin{aligned}
M(\boldsymbol{\theta}_{0\gamma}) &= -Y\tilde{\boldsymbol{X}}^T I_{\{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma})<\gamma\}} - \frac{(1-\gamma)^2 Y\tilde{\boldsymbol{X}}^T}{(Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma})-2\gamma+1)^2} I_{\{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma})\geq\gamma\}};\\
R(Z,t) &= \Big(1-Yf(\boldsymbol{X};\boldsymbol{\theta}_{0\gamma}+t)\Big)\Big[I_{\{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma}+t)<\gamma\}} - I_{\{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma})<\gamma\}}\Big] + \frac{(1-\gamma)^2 I_{\{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma}+t)\geq\gamma\}}}{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma}+t)-2\gamma+1}\\
&\quad - \left[\frac{(1-\gamma)^2}{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma})-2\gamma+1} - \frac{(1-\gamma)^2 Yf(\boldsymbol{X},t)}{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma})-2\gamma+1}\right] I_{\{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma})\geq\gamma\}}.
\end{aligned}
$$

It is easy to check that $E(M(\boldsymbol{\theta}_{0\gamma})) = \nabla_{\boldsymbol{\theta}}\mathcal{R}_\gamma(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0\gamma}}$,

$$
E[R(Z,t)] = \frac{1}{2}t^T H(\boldsymbol{\theta}_{0\gamma})t + o(\|t\|^2) \quad \text{and} \quad E[R^2(Z,t)] = O(\|t\|^3).
$$

The remaining arguments follow exactly from the proof of Theorem 1. ■

S.7. Proof of Lemma 1

In the proof of Corollary 2, we showed that for any $\gamma \in [0,1]$,

$$
\sqrt{n}(\widehat{\boldsymbol{\theta}}_\gamma - \boldsymbol{\theta}_{0\gamma}) = -n^{-1/2}H(\boldsymbol{\theta}_{0\gamma})^{-1}\sum_{i=1}^n M_i(\boldsymbol{\theta}_{0\gamma}) + o_P(1); \tag{S.12}
$$

$$
\sqrt{n}(\widehat{\mathcal{D}}_\gamma - D_{0\gamma}) = n^{-1/2}\sum_{i=1}^n \psi_{i\gamma} + o_P(1), \tag{S.13}
$$

where $\psi_{i\gamma} = \frac{1}{2}|Y_i - \text{sign}\{f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})\}| - D_{0\gamma} - \dot{d}(\boldsymbol{\theta}_{0\gamma})^T H(\boldsymbol{\theta}_{0\gamma})^{-1}M_i(\boldsymbol{\theta}_{0\gamma})$. In addition, (S.12) and (S.13) converge to normal distributions.

Next, we show that the right hand sides of (S.12) and (S.13) are uniformly bounded over

$\gamma \in [0,1]$. Denoting the L_1 norm as $\|\cdot\|_1$, we have

$$\begin{aligned}
&\sup_{\gamma\in[0,1]} \left\| M_i(\boldsymbol{\theta}_{0\gamma}) \right\|_1 \\
\leq\ & \sup_{\gamma\in[0,1]} \left\| -Y_i \tilde{\boldsymbol{X}}_i I_{(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})<\gamma)} \right\|_1 + \sup_{\gamma\in[0,1]} \left\| \frac{(1-\gamma)^2 Y_i \tilde{\boldsymbol{X}}_i I_{(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})\geq\gamma)}}{\Big(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma}) - 2\gamma + 1\Big)^2} \right\|_1 \\
\leq\ & 2\left\| \tilde{\boldsymbol{X}}_i \right\|_1. \qquad \text{(S.14)}
\end{aligned}$$

In addition, $\lambda_{\max}(H(\boldsymbol{\theta}_{0\gamma})) \leq c_2$ in Assumption (B1) implies that each component of the Hessian matrix is uniformly bounded since $\|H(\boldsymbol{\theta}_{0\gamma})\|_{\max} \leq \|H(\boldsymbol{\theta}_{0\gamma})\|_2 = \lambda_{\max}(H(\boldsymbol{\theta}_{0\gamma}))$. This combining with (S.14) and Central Limit Theorem leads to

$$\sup_{\gamma\in[0,1]} \left\| \sqrt{n}(\widehat{\boldsymbol{\theta}}_\gamma - \boldsymbol{\theta}_{0\gamma}) \right\|_1 = O_P(1). \qquad \text{(S.15)}$$

Similarly,

$$\begin{aligned}
&\sup_{\gamma\in[0,1]} \left| \psi_{i\gamma} \right| \\
\leq\ & \sup_{\gamma\in[0,1]} \frac{1}{2} |Y_i - \operatorname{sign}(\tilde{\boldsymbol{X}}_i^T \boldsymbol{\theta}_{0\gamma})| + \sup_{\gamma\in[0,1]} |D_{0\gamma}| + \sup_{\gamma\in[0,1]} \left| \dot{d}(\boldsymbol{\theta}_{0\gamma})^T H(\boldsymbol{\theta}_{0\gamma})^{-1} M_i(\boldsymbol{\theta}_{0\gamma}) \right| \\
\leq\ & 1 + 1 + \sup_{\gamma\in[0,1]} \left\| \dot{d}(\boldsymbol{\theta}_{0\gamma}) \right\|_1 \sup_{\gamma\in[0,1]} \left\| H(\boldsymbol{\theta}_{0\gamma})^{-1} \right\|_{\max} \sup_{\gamma\in[0,1]} \left\| M_i(\boldsymbol{\theta}_{0\gamma}) \right\|_1 \\
\leq\ & 2 + c_3 \|\tilde{\boldsymbol{X}}_i\|_1, \qquad \text{(S.16)}
\end{aligned}$$

where c_3 in (S.16) is a constant according to $\|H(\boldsymbol{\theta}_{0\gamma})^{-1}\|_{\max} \leq \|H(\boldsymbol{\theta}_{0\gamma})^{-1}\|_2 = 1/\lambda_{\min}(H(\boldsymbol{\theta}_{0\gamma})) \leq 1/c_1$ from Assumption (B1), and

$$\|\dot{d}(\boldsymbol{\theta}_{0\gamma})\|_1 \leq 4\left\| \triangledown E\Big(I_{(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})<0)}\Big) \right\|_1 \leq 4\delta(-Y_i \boldsymbol{\theta}_{0\gamma}^T \tilde{\boldsymbol{X}}_i) \|\tilde{\boldsymbol{X}}_i\|_1 = 0 \ \text{ a.s.}$$

with $\delta(z) = 0$ for $z \neq 0$ and ∞ at $z = 0$. So (S.16) leads to

$$\sup_{\gamma\in[0,1]} \sqrt{n}\Big|\widehat{\mathcal{D}}_\gamma - D_{0\gamma}\Big| = O_P(1). \tag{S.17}$$

In the end, the definitions of γ_0^* and $\widehat{\gamma}_0^*$ imply that

$$D_{0\gamma_0^*} - D_{0\widehat{\gamma}_0^*} \leq 0 \;\; and \;\; \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} - \widehat{\mathcal{D}}_{\gamma_0^*} \leq 0. \tag{S.18}$$

Therefore, we have $D_{0\gamma_0^*} - \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} = D_{0\gamma_0^*} - D_{0\widehat{\gamma}_0^*} + D_{0\widehat{\gamma}_0^*} - \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} \leq D_{0\widehat{\gamma}_0^*} - \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} = O_P(n^{-1/2})$ based on (S.17) and (S.18). Using similar arguments, we have $\widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} - D_{0\gamma_0^*} \leq O_P(n^{-1/2})$. The above discussions imply that $\Big|\widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} - D_{0\gamma_0^*}\Big| = O_P(n^{-1/2})$. This concludes the proof of Lemma 1. ■

S.8. Lemma 3

The following Lemma will be used in the proof of Lemma 4.

Lemma 3 *The generalization error* $D_{0\gamma} = \frac{1}{2}E|Y_0 - sign\{\tilde{\boldsymbol{X}}_0^T\widehat{\boldsymbol{\theta}}_\gamma\}|$ *is continuous w.r.t.* γ *a.s.*

Proof of Lemma 3: The discontinuity of sign function happens only at $\tilde{\boldsymbol{X}}_0^T\widehat{\boldsymbol{\theta}}_\gamma = 0$, which is assumed to have probability zero. Hence, it is sufficient to show $\widehat{\boldsymbol{\theta}}_\gamma$ is continuous in γ by dominated convergence theorem. Recall that $\widehat{\boldsymbol{\theta}}_\gamma = \arg\min_{\boldsymbol{\theta}\in R^{d+1}} O_{n\gamma}(\boldsymbol{\theta})$ with

$$O_{n\gamma}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} L_\gamma\Big(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\Big) + \frac{\lambda_n \boldsymbol{w}^T\boldsymbol{w}}{2}.$$

Note that $O_{n\gamma}(\boldsymbol{\theta})$ is continuous w.r.t. γ due to the continuity of $L_\gamma(u)$ w.r.t. γ. Then, for any sequence $\gamma_n \to \gamma_{00}$ with $\gamma_{00} \in [0,1]$, continuous mapping theorem implies that $|O_{n\gamma_n}(\boldsymbol{\theta}) - O_{n\gamma_{00}}(\boldsymbol{\theta})| < \delta$ for any $\delta > 0$ when n is sufficiently large. Denote $\widehat{\boldsymbol{\theta}}_{\gamma_{00}} = \arg\min_{\boldsymbol{\theta}} O_{n\gamma_{00}}(\boldsymbol{\theta})$ and $\mathcal{G} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\gamma_{00}}\| \leq \epsilon\}$. For each fixed ϵ, we construct

$$\delta = \frac{\min_{\boldsymbol{\theta}\in R^{d+1}\setminus\mathcal{G}} O_{n\gamma_{00}}(\boldsymbol{\theta}) - O_{n\gamma_{00}}(\widehat{\boldsymbol{\theta}}_{\gamma_{00}})}{2}.$$

Then we have

$$\begin{aligned} O_{n\gamma_{00}}(\widehat{\boldsymbol{\theta}}_{\gamma_{00}}) &= \min_{\boldsymbol{\theta}\in R^{d+1}\setminus\mathcal{G}} O_{n\gamma_{00}}(\boldsymbol{\theta}) - 2\delta \\ &< \min_{\boldsymbol{\theta}\in R^{d+1}\setminus\mathcal{G}} O_{n\gamma_{00}}(\boldsymbol{\theta}) + O_{n\gamma_n}(\boldsymbol{\theta}) - O_{n\gamma_{00}}(\boldsymbol{\theta}) - \delta \\ &\leq O_{n\gamma_n}(\boldsymbol{\theta}) - \delta, \end{aligned}$$

which is true for any $\boldsymbol{\theta} \in R^{d+1}$. Therefore,

$$O_{n\gamma_{00}}(\widehat{\boldsymbol{\theta}}_{\gamma_{00}}) < \min_{\boldsymbol{\theta}\in R^{d+1}\setminus\mathcal{G}} O_{n\gamma_n}(\boldsymbol{\theta}) - \delta. \tag{S.19}$$

On the other hand, $|O_{n\gamma_n}(\boldsymbol{\theta}) - O_{n\gamma_{00}}(\boldsymbol{\theta})| < \delta$ implies that $O_{n\gamma_n}(\widehat{\boldsymbol{\theta}}_{\gamma_{00}}) - O_{n\gamma_{00}}(\widehat{\boldsymbol{\theta}}_{\gamma_{00}}) < \delta$ and hence $\min_{\boldsymbol{\theta}\in R^{d+1}} O_{n\gamma_n}(\boldsymbol{\theta}) < O_{n\gamma_{00}}(\widehat{\boldsymbol{\theta}}_{\gamma_{00}}) + \delta$. This combining with (S.19) leads to

$$\min_{\boldsymbol{\theta}\in R^{d+1}} O_{n\gamma_n}(\boldsymbol{\theta}) < \min_{\boldsymbol{\theta}\in R^{d+1}\setminus\mathcal{G}} O_{n\gamma_n}(\boldsymbol{\theta}).$$

Therefore, $\arg\min_{\boldsymbol{\theta}\in R^{d+1}} O_{n\gamma_n}(\boldsymbol{\theta}) \in \mathcal{G}$, and hence $\widehat{\boldsymbol{\theta}}_{\gamma}$ is continuous at γ_{00}. Note that ϵ can be made arbitrarily small and γ_{00} is an arbitrary element within $[0, 1]$. This concludes Lemma 3. ■

S.9. Lemma 4

Lemma 4 shows the (element-wise) asymptotic equivalence between Λ_0 and $\widehat{\Lambda}_0$. It will be used in the proof of Theorem 3.

Lemma 4 *Suppose that the assumptions in Lemma 1 hold. We have, as $n \to \infty$, (i) for any $\widehat{\gamma} \in \widehat{\Lambda}_0$, there exists a $\gamma \in \Lambda_0$ such that $\widehat{\gamma} \xrightarrow{P} \gamma$; (ii) for any $\gamma \in \Lambda_0$, there exists a $\widehat{\gamma} \in \widehat{\Lambda}_0$ satisfying $\widehat{\gamma} \xrightarrow{P} \gamma$.*

Proof of Lemma 4: Our proof consists of two steps. In the first step, for any $\widehat{\gamma} \in \widehat{\Lambda}_0$ with

$\widehat{\gamma} \xrightarrow{P} \gamma$, we have

$$\begin{aligned} D_{0\gamma} - D_{0\gamma_0^*} &= (D_{0\gamma} - D_{0\widehat{\gamma}}) + (D_{0\widehat{\gamma}} - \widehat{\mathcal{D}}_{\widehat{\gamma}}) + (\widehat{\mathcal{D}}_{\widehat{\gamma}} - \widehat{\mathcal{D}}_{\gamma_0^*}) + (\widehat{\mathcal{D}}_{\gamma_0^*} - D_{0\gamma_0^*}) \\ &= I + II + III + IV. \end{aligned}$$

Obviously, we have $I = o_P(1)$ according to continuous mapping theorem and Lemma 3, and $II, IV = o_P(1)$ due to (S.17). As for III, we have $III \leq \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} - \widehat{\mathcal{D}}_{\gamma_0^*} + n^{-1/2}\phi_{\widehat{\gamma},\widehat{\gamma}_0^*;\alpha/2} \leq o_P(1)$ since $\widehat{\gamma} \in \widehat{\Lambda}_0$ defined in (15). The above discussions lead to the conclusion that $D_{0\gamma} - D_{0\gamma_0^*} \leq o_P(1)$. Therefore, we have $P(\gamma \in \Lambda_0) \geq P(D_{0\gamma} - D_{0\gamma_0^*} \leq 0) \to 1$.

In the second step, we apply the contradiction argument. Assume there exists some $\gamma \in \Lambda_0$ such that $\widehat{\gamma} \notin \widehat{\Lambda}_0$ for any $\widehat{\gamma} \xrightarrow{P} \gamma$. The above assumption directly implies that $\widehat{\mathcal{D}}_{\widehat{\gamma}} - \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} > o_P(1)$. The analysis in the first step further implies that there exists some $\gamma^* \in \Lambda_0$, i.e., $D_{0\gamma^*} = D_{0\gamma_0^*}$, with probability tending to one such that $\widehat{\gamma}_0^* \xrightarrow{P} \gamma^*$. Then, we have

$$\begin{aligned} D_{0\gamma} - D_{0\gamma^*} &= (D_{0\gamma} - D_{0\widehat{\gamma}}) + (D_{0\widehat{\gamma}} - \widehat{\mathcal{D}}_{\widehat{\gamma}}) + (\widehat{\mathcal{D}}_{\widehat{\gamma}} - \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*}) + (\widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} - D_{0\gamma^*}) \\ &= I + II + III' + IV'. \end{aligned}$$

Recall that $I, II = o_P(1)$ and $III' > o_P(1)$ as shown in the above. We also have $IV' = o_P(1)$ due to (S.17) and the fact that $\widehat{\gamma}_0^* \xrightarrow{P} \gamma^*$. In summary, we have $D_{0\gamma} - D_{0\gamma^*} > o_P(1)$, which contradicts the definition of γ. This concludes the proof of Lemma 4. ■

S.10. Proof of Theorem 3

The proof consists of two major steps. In the first step, we show that

$$\sup_{\gamma \in [0,1]} n \left| \widehat{DBI}(S(\boldsymbol{X}; \widehat{\boldsymbol{\theta}}_\gamma)) - DBI(S(\boldsymbol{X}; \widehat{\boldsymbol{\theta}}_\gamma)) \right| \to 0. \tag{S.20}$$

Denote $\overline{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) = \frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{x}}_{i(-d)}^{\dagger T} Var(\widehat{\boldsymbol{\eta}}_\gamma^\dagger)\widetilde{\boldsymbol{x}}_{i(-d)}^\dagger$, where $\widetilde{\boldsymbol{x}}_{i(-d)}^\dagger = (1, (R_\gamma \boldsymbol{x}_i)_{(-d)}^T)^T$ and R_γ is the transformation matrix associated with the loss function L_γ. Then we have

$$\begin{aligned} & \sup_{\gamma\in[0,1]} n\left|\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\right| \\ \leq & \sup_{\gamma\in[0,1]} n\left|\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - \overline{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\right| + \sup_{\gamma\in[0,1]} n\left|\overline{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\right|. \end{aligned} \tag{S.21}$$

Next we show each summand in (S.21) converges to 0. For the first summand, we have

$$\begin{aligned} & \sup_{\gamma\in[0,1]} n\left|\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - \overline{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\right| \\ = & \sup_{\gamma\in[0,1]} n\left|\frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{x}}_{i(-d)}^{\dagger T}\widehat{Var}(\widehat{\boldsymbol{\eta}}_\gamma^\dagger)\widetilde{\boldsymbol{x}}_{i(-d)}^\dagger - \frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{x}}_{i(-d)}^{\dagger T} Var(\widehat{\boldsymbol{\eta}}_\gamma^\dagger)\widetilde{\boldsymbol{x}}_{i(-d)}^\dagger\right| \\ = & \sup_{\gamma\in[0,1]} \left|\sum_{i=1}^n \widetilde{\boldsymbol{x}}_{i(-d)}^{\dagger T}[(\widehat{Var}(\widehat{\boldsymbol{\eta}}_\gamma^\dagger) - Var(\widehat{\boldsymbol{\eta}}_\gamma^\dagger))]\widetilde{\boldsymbol{x}}_{i(-d)}^\dagger\right|, \end{aligned} \tag{S.22}$$

where

$$Var(\widehat{\boldsymbol{\eta}}_\gamma^\dagger) = \frac{\Sigma_{0\gamma,(-d)}^\dagger}{n(w_{\gamma,d}^\dagger)^2} \text{ and } \widehat{Var}(\widehat{\boldsymbol{\eta}}_\gamma^\dagger) = \frac{\widehat{\Sigma}_{\gamma,(-d)}^\dagger}{n(\widehat{w}_{\gamma,d}^\dagger)^2}.$$

Here, $\widehat{w}_{\gamma,d}^\dagger$ is the last dimension of $\widehat{\boldsymbol{\theta}}_\gamma^*$. Since $\widehat{w}_{\gamma,d}^\dagger$ follows the normal distribution with mean $w_{\gamma,d}^\dagger$ and variance converging to 0, we have $\widehat{w}_{\gamma,d}^\dagger = w_{\gamma,d}^\dagger + o_P(1)$, and hence $(\widehat{w}_{\gamma,d}^\dagger)^2 = (w_{\gamma,d}^\dagger)^2 + o_P(1)$ due to the boundedness of $w_{\gamma,d}^\dagger$. In addition, uniform law of large numbers implies that each component of $\widehat{\Sigma}_\gamma^\dagger - \Sigma_{0\gamma}^\dagger$ uniformly converges to 0 w.r.t. γ, because each element of $\widehat{\Sigma}_\gamma^\dagger$ is continuous w.r.t. γ (by similar arguments as in Lemma 3). Therefore, we have

$$\begin{aligned} n\left[\widehat{Var}(\widehat{\boldsymbol{\eta}}_\gamma^\dagger) - Var(\widehat{\boldsymbol{\eta}}_\gamma^\dagger)\right] &= \frac{\widehat{\Sigma}_{\gamma,(-d)}^\dagger}{(\widehat{w}_{\gamma,d}^\dagger)^2} - \frac{\Sigma_{0\gamma,(-d)}^\dagger}{(w_{\gamma,d}^\dagger)^2} \\ &= \frac{\widehat{\Sigma}_{\gamma,(-d)}^\dagger - \Sigma_{0\gamma,(-d)}^\dagger}{(w_{\gamma,d}^\dagger)^2 + o_P(1)} - \frac{\Sigma_{0\gamma,(-d)}^\dagger o_P(1)}{(w_{\gamma,d}^\dagger)^2[(w_{\gamma,d}^\dagger)^2 + o_P(1)]}, \end{aligned} \tag{S.23}$$

where the second term in (S.23) uniformly converges to 0 due to Assumption (B1) and the boundedness of $w^{\dagger}_{\gamma,d}$. Therefore, each element of (S.23) uniformly converges to 0, which implies that (S.22) converges to 0.

As for the second summand of (S.21), we again apply uniform law of large numbers to show

$$\sup_{\gamma\in[0,1]} n\left|\overline{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma}))\right| \to 0.$$

Note that $\tilde{\boldsymbol{X}}^{\dagger T}_{(-d)} Var(\widehat{\boldsymbol{\eta}}_{\gamma})^{\dagger} \tilde{\boldsymbol{X}}^{\dagger}_{(-d)}$ is continuous w.r.t. γ by similar arguments as in Lemma 3, and

$$n\left|\tilde{\boldsymbol{X}}^{\dagger T}_{(-d)} Var(\widehat{\boldsymbol{\eta}}^{\dagger}_{\gamma}) \tilde{\boldsymbol{X}}^{\dagger}_{(-d)}\right| = \left|(1,(R_{\gamma}\boldsymbol{x})^{T}_{(-d)})^{T} n Var(\widehat{\boldsymbol{\eta}}^{\dagger}_{\gamma})(1,(R_{\gamma}\boldsymbol{x})^{T}_{(-d)})\right| \le c_4\left|1+\boldsymbol{x}^{T}_{(-d)}\boldsymbol{x}_{(-d)}\right| \le c_5,$$

where the first inequality holds because each component of $nVar(\widehat{\boldsymbol{\eta}}^{\dagger}_{\gamma})$ is uniformly bounded due to the boundedness of $w^{\dagger}_{\gamma,d}$ and Assumption (B1). Then the uniform law of large number implies

$$\begin{aligned} & \sup_{\gamma\in[0,1]} n\left|\overline{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma}))\right| \\ = & \sup_{\gamma\in[0,1]} \left|\frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{x}}^{\dagger T}_{i(-d)} (w^{\dagger}_{\gamma,d})^{-2} \Sigma^{\dagger}_{0\gamma,(-d)}(\widehat{\boldsymbol{\eta}}^{\dagger}_{\gamma}) \widetilde{\boldsymbol{x}}^{\dagger}_{i(-d)} - E\left(\tilde{\boldsymbol{X}}^{\dagger T}_{(-d)} (w^{\dagger}_{\gamma,d})^{-2} \Sigma^{\dagger}_{0\gamma,(-d)} \tilde{\boldsymbol{X}}^{\dagger}_{(-d)}\right)\right| \\ \to & \ 0. \end{aligned} \tag{S.24}$$

Combining (S.22) and (S.24) leads to (S.20).

In the second step of the proof, we show $n(\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0}))) \le o_P(1)$ and $n(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0}))) \le o_P(1)$, from which the desirable result (20) follows.

Firstly, we prove

$$n\left(\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0}))\right) \le o_P(1).$$

Denote $\widehat{\gamma}_0^\sharp = \arg\min_{\gamma\in\widehat{\Lambda}_0} DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))$. For γ_0 defined in (19), Theorem 4 implies that there exists a $\widehat{\gamma}_0^\triangle \in \widehat{\Lambda}_0$ such that $\widehat{\gamma}_0^\triangle \xrightarrow{P} \gamma_0$, then we have

$$\begin{aligned}
& n\Big(\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0}))\Big) \\
= \;& n\Big(\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp}))\Big) + n\Big(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\triangle}))\Big) \\
& + n\Big(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\triangle})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0}))\Big) \\
\le \;& n\Big(\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp}))\Big) + n\Big(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\triangle}))\Big) \\
& + n\Big(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\triangle})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0}))\Big) \\
\le \;& \sup_{\gamma\in\widehat{\Lambda}_0} n\Big|\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\Big| + o_P(1) \\
\le \;& o_P(1), \qquad\qquad (S.25)
\end{aligned}$$

where $\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})) \le \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp}))$ according to (18), $DBI(S(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp})) \le DBI(S(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\triangle}))$ due to $\widehat{\gamma}_0^\sharp \in \widehat{\Lambda}_0$, $DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\triangle})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0})) = o_P(n^{-1})$ according to $\widehat{\gamma}_0^\triangle \xrightarrow{P} \gamma_0$ and continuous mapping theorem. All these together with (S.20) lead to (S.25).

Secondly, we prove

$$n(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0}))) \le o_P(1).$$

Denote $\widetilde{\gamma}_0 = \arg\min_{\gamma\in\Lambda_0} \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))$. For $\widehat{\gamma}_0$ defined in (18), Lemma 4 implies that there exists $\widetilde{\gamma}_0^\sharp \in \Lambda_0$ such that $\widehat{\gamma}_0 \xrightarrow{P} \widetilde{\gamma}_0^\sharp$, then we have

$$\begin{aligned}
& n\Big(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0}))\Big) \\
\le \;& n\Big(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0}))\Big) + n\Big(\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0^\sharp}))\Big) \\
& + n\Big(\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0^\sharp})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0}))\Big) \\
\le \;& \sup_{\gamma\in\Lambda_0} n\Big|\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\Big| + o_P(1) \le o_P(1), \qquad\qquad (S.26)
\end{aligned}$$

where $DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0})) \le DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0}))$ by the definition of γ_0, $\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0})) \le$

$\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0^\sharp}))$ due to the definition of $\widetilde{\gamma}_0$, and $\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0^\sharp})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})) = o_P(n^{-1})$ according to $\widehat{\gamma}_0 \xrightarrow{P} \widetilde{\gamma}_0^\sharp$ and continuous mapping theorem.

Consequently, combining (*S*.25) and (*S*.26) leads to $n\left|\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0}))-DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0}))\right| \to 0$, which concludes the proof of Theorem 3. ■

S.11. Notation Table S1

References

Hinkley, D.V. (1969), On the Ratio of Two Correlated Normal Random Variables, *Biometrika*, **56**, 635-639.

Hjort, N. and Pollard, D. (1993), Asymptotics for Minimisers of Convex Processes, *Statistical Research Report, Department of Mathematics, University of Oslo.*

Pollard, D. (1991), Asymptotics for Least Absolute Deviation Regression Estimatros, *Econometric Theory*, **7**, 186-199.

van der Vaart, A. W. (1998), Asymptotic Statistics, *Cambridge University Press.*

van der Vaart, A. W. and Wellner, J. A. (1996), Weak Convergence and Empirical Processes, *Springer-Velag, New York.*

Table S1: Important notation, its meaning, and where it first appears.

Notation	Meaning	Section No.
$\boldsymbol{x}$	input variable	2
$\tilde{\boldsymbol{x}}$	$\tilde{\boldsymbol{x}} = (1, \boldsymbol{x}^T)^T$	2
$p(\boldsymbol{x})$	conditional probability $P(Y = 1 \vert \boldsymbol{X} = \boldsymbol{x})$	2
b,$\boldsymbol{w}$, $\boldsymbol{\theta}$	intercept, coefficient and parameter $\boldsymbol{\theta} = (b, \boldsymbol{w}^T)^T$	2
$S(\boldsymbol{x}; \boldsymbol{\theta})$	the decision boundary induced from $\boldsymbol{\theta}$	2
$\mathcal{P}(\boldsymbol{X}, Y)$	joint distribution of $(\boldsymbol{X}, Y)$	2
$\mathcal{R}_L$	risk of loss function L	2
b_{0L}, $\boldsymbol{w}_{0L}$, $\boldsymbol{\theta}_{0L}$	true intercept, coefficient, and parameter	2
$(\boldsymbol{x}_i, y_i)$, $\mathcal{D}_n$	training data $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$	2
O_{nL}	empirical risk of loss function L	2
$\widehat{b}_L$, $\widehat{\boldsymbol{w}}_L$, $\widehat{\boldsymbol{\theta}}_L$	estimated intercept, coefficient, and parameter	2
D_{0L}	GE from loss function L	3.1
$\widehat{D}_L$, $\widehat{D}(\widehat{\boldsymbol{\theta}}_L)$	empirical generalization error from loss function L	3.1
$\widehat{\mathcal{D}}_L$	K-CV error from loss function L	3.1
$G(\boldsymbol{\theta}_{0L})$, $H(\boldsymbol{\theta}_{0L})$	the gradient matrix and Hessian matrix	3.1
$\mathcal{W}_L$	$= n^{1/2}(\widehat{\mathcal{D}}_L - D_{0L})$	3.1
$D_0(\boldsymbol{\theta})$	$= \frac{1}{2}E\vert y_0 - \text{sign}\{f(\boldsymbol{x}_0; \boldsymbol{\theta})\}\vert$	3.1
$\dot{d}(\boldsymbol{\theta})$	$= \nabla_{\boldsymbol{\theta}} E(\widehat{D}(\boldsymbol{\theta}))$	3.1
D_{0j}, $\widehat{D}_j$, $\widehat{\mathcal{D}}_j$	GE, empirical GE, and K-CV error w.r.t L_j	3.1
Δ_{12}, $\widehat{\Delta}_{12}$	$\Delta_{12} = D_{02} - D_{01}$; $\widehat{\Delta}_{12} = \widehat{\mathcal{D}}_2 - \widehat{\mathcal{D}}_1$	3.1
$\mathcal{W}_j$	$= n^{1/2}(\widehat{\mathcal{D}}_j - D_{0j})$	3.1
$\mathcal{W}_{\Delta_{12}}$	$= \mathcal{W}_2 - \mathcal{W}_1$	3.1
G_i	the random variable generated from Exp(1)	3.1
$\widehat{\boldsymbol{\theta}}_j^*$, W_j^*, $W_{\Delta_{12}}^*$	the perturbed version of the corresponding terms	3.1
$\widehat{\boldsymbol{\theta}}_j^{*(r)}$, $W_j^{*(r)}$, $W_{\Delta_{12}}^{*(r)}$	the corresponding terms in the rth replication	3.1
$\phi_{1,2;\alpha}$	the αth upper percentile of the sequence $W_{\Delta_{12}}^{*(r)}$	3.1
$\mathcal{X}_1, \ldots, \mathcal{X}_d$	the original axes	3.2
R_L	the transformation matrix induced from loss L	3.2
$\widehat{b}_L^\dagger$, $\widehat{\boldsymbol{w}}_L^\dagger$, $\widehat{\boldsymbol{\theta}}_L^\dagger$	transformed estimates of parameters	3.2
$\Sigma_{0L}^\dagger$, $\widehat{\Sigma}_L^\dagger$	the covariance matrix and its transformed estimator	3.2
$\Sigma_{0L,(-d)}^\dagger$, $\widehat{\Sigma}_{L,(-d)}^\dagger$	removing last row and last column of $\Sigma_{0L}^\dagger$ and $\widehat{\Sigma}_L^\dagger$	3.2
L_γ	the LUM loss function indexed by γ	4
$\boldsymbol{\theta}_{0\gamma}$, $\widehat{\boldsymbol{\theta}}_\gamma$	true and estimated parameter from L_γ	4
$\mathcal{R}_\gamma$	true risk from L_γ	4
$D_{0\gamma}$, $\widehat{\mathcal{D}}_\gamma$	GE and CV error from L_γ	4
γ_0^*, $\widehat{\gamma}_0^*$	LUM index of minimal GE, minimal K-CV error	4
Λ_0, $\widehat{\Lambda}_0$	true and estimated set of potentially good classifiers	4
γ_0, $\widehat{\gamma}_0$	optimal index and its estimate	4