

SUPPLEMENTAL MATERIALS TO
“ROBUST CLASSIFICATION WITH FEATURE SELECTION
USING NON-GAUSSIAN DATA”

Xiuchuan Liu and Xianzheng Huang

University of South Carolina

Throughout the study, we assume that the observed data consist of n independent observations $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, drawn from the joint distribution of $(\mathbf{X}, Y)$. Let $n_0 = \sum_{i=1}^n \mathbb{1}(Y_i = 0)$ and $n_1 = n - n_0$. Condition (A3) requires $\beta_0^\circ = \log\{P(Y = 1)/P(Y = 0)\}$ to be finite, so that $P(Y = k) \in (0, 1)$ for $k = 0, 1$. By the strong law of large numbers, n_k/n converges to $P(Y = k)$ almost surely, and in particular $n_k \asymp n$ as $n \rightarrow \infty$. We use this whenever a rate stated in n_k is restated in n .

Appendix A: The Proof of Theorem 1

Theorem 1 concerns the proposed S-RoLLR classifier $\hat{g}_n(\mathbf{X})$ under the correct model specification. Define the vector of regression coefficients $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\top)^\top$, and denote by $\boldsymbol{\Theta}$ the corresponding parameter space that is compact under Condition (A3). More specifically, for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\|\boldsymbol{\theta}\|_1 \leq M$

for a constant $M > |\beta_0^\circ| + s$.

Theorem 1. *Under Conditions (A1)-(A3), the classification risk of the S-RoLLR classifier converges to the optimal Bayes risk as $n \rightarrow \infty$, that is, $\lim_{n \rightarrow \infty} R(\hat{g}_n) = R^*$.*

Proof. First, let $\mathbf{x} = (x_1, \dots, x_p) \in \mathbb{R}^p$ be a point in the original feature space, we establish the consistency of the transformed feature $\hat{\mathbf{T}}(\mathbf{x})$ as an estimator for the oracle feature $\mathbf{T}^*(\mathbf{x})$. Here, $\hat{\mathbf{T}}(\mathbf{x}) = (\hat{T}_1(x_1), \dots, \hat{T}_p(x_p))^\top$, where $\hat{T}_j(x_j) = \log(f_{\text{TPSC}}(x_j|\hat{\Psi}_{j1})/f_{\text{TPSC}}(x_j|\hat{\Psi}_{j0}))$, in which $\hat{\Psi}_{jk}$ is the maximum likelihood estimator (MLE) for Ψ_{jk} , and $\mathbf{T}^*(\mathbf{x}) = (T_1^*(x_1), \dots, T_p^*(x_p))^\top$, where $T_j^*(x_j) = \log(f_{\text{TPSC}}(x_j|\Psi_{j1}^*)/f_{\text{TPSC}}(x_j|\Psi_{j0}^*))$, in which Ψ_{jk}^* is the truth of Ψ_{jk} , for $j \in [p], k = 0, 1$. Under regularity conditions guaranteed by (A1), $\hat{\Psi}_{jk} \xrightarrow{p} \Psi_{jk}^*$, as $n_k \rightarrow \infty$. Consider the mapping $h(\Psi_{j0}, \Psi_{j1}) = \log f_{\text{TPSC}}(\cdot | \Psi_{j1}) - \log f_{\text{TPSC}}(\cdot | \Psi_{j0})$, which is a continuous mapping by the construction of the TPSC density. It follows that $h(\hat{\Psi}_{j0}, \hat{\Psi}_{j1}) \xrightarrow{p} h(\Psi_{j0}^*, \Psi_{j1}^*)$, as $n_0, n_1 \rightarrow \infty$, by the continuous mapping theorem. This is equivalent to $\hat{T}_j(\cdot) \xrightarrow{p} T_j^*(\cdot)$, as $n_0, n_1 \rightarrow \infty$, for $j \in [p]$. Hence, $\hat{\mathbf{T}}(\mathbf{x}) \xrightarrow{p} \mathbf{T}^*(\mathbf{x})$, as $n \rightarrow \infty$.

Second, let $\pi^*(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$ denote the class membership probability, and we establish the consistency of class membership estimation using the S-RoLLR classifier. Under Condition (A2) that imposes the

conditional independence of p features, the true log-odds (with parameters evaluated at their truth) is

$$\begin{aligned}
 \eta^*(\mathbf{x}) &= \log \frac{P(Y = 1 \mid \mathbf{X} = \mathbf{x})}{P(Y = 0 \mid \mathbf{X} = \mathbf{x})} \\
 &= \log \frac{P(Y = 1)}{P(Y = 0)} + \sum_{j=1}^p \log \frac{f(x_j \mid Y = 1)}{f(x_j \mid Y = 0)} \\
 &= \beta_0^\circ + \sum_{j=1}^p T_j^*(x_j), \text{ with } \beta_0^\circ = \log \frac{P(Y = 1)}{P(Y = 0)}. \tag{A.1}
 \end{aligned}$$

It follows that

$$\pi^*(\mathbf{x}) = P(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \frac{1}{1 + \exp(-\eta^*(\mathbf{x}))} = \mathcal{G}(\eta^*(\mathbf{x})),$$

where $\mathcal{G}(t) = 1/\{1 + \exp(-t)\}$ is the sigmoid function. This suggests that, when regressing the class label Y on the oracle features $\mathbf{T}^*(\mathbf{X})$, the logistic regression model with a linear predictor $\eta^*(\mathbf{X})$ (linear in $\mathbf{T}^*(\mathbf{X})$) is a correct model. Certainly, the oracle features are not available in a given application due to their dependence on the TPSC parameters. The proposed classifier S-RoLLR involves an L_1 -penalized logistic regression of Y on the transformed features in $\hat{\mathbf{T}}(\mathbf{X})$, minimizing $\mathcal{L}_n(\boldsymbol{\theta}) = Q_n(\boldsymbol{\theta}) + \lambda_n \|\boldsymbol{\beta}\|_1$ over $\boldsymbol{\Theta}$, where $Q_n(\boldsymbol{\theta})$ is the unpenalized negative log-likelihood (times $1/n$) given below,

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \left[-Y_i \left\{ \beta_0 + \hat{\mathbf{T}}(\mathbf{X}_i)^\top \boldsymbol{\beta} \right\} + \log \left(1 + \exp \left(\beta_0 + \hat{\mathbf{T}}(\mathbf{X}_i)^\top \boldsymbol{\beta} \right) \right) \right].$$

Denote by $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}^\top)^\top$ a minimizer of $\mathcal{L}_n(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}$. Under Condition

(A3), the penalty is asymptotically negligible uniformly over the parameter space, because

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\mathcal{L}_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta})| = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \lambda_n \|\boldsymbol{\beta}\|_1 \leq \lambda_n M \longrightarrow 0.$$

Consequently, once the uniform convergence $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \xrightarrow{P} 0$ is established below, the same holds with Q_n replaced by $\mathcal{L}_n$, and every argmin argument applied to Q_n applies verbatim to $\mathcal{L}_n$. We therefore work with Q_n in what follows. The class membership probability $\pi^*(\mathbf{x})$ is estimated by $\hat{\pi}(\mathbf{x}) = \mathcal{G}(\hat{\eta}(\mathbf{x}))$, with $\hat{\eta}(\mathbf{x}) = \hat{\beta}_0 + \hat{\mathbf{T}}(\mathbf{x})^\top \hat{\boldsymbol{\beta}}$.

The logistic regression of Y on $\mathbf{T}^*(\mathbf{X})$ has the counterpart negative log-likelihood converging to, as $n \rightarrow \infty$,

$$Q_0(\boldsymbol{\theta}) = E_{\mathbf{X}, Y} [-Y \{\beta_0 + \mathbf{T}^*(\mathbf{X})^\top \boldsymbol{\beta}\} + \log(1 + \exp(\beta_0 + \mathbf{T}^*(\mathbf{X})^\top \boldsymbol{\beta}))].$$

According to (A.1), define the canonical sparse coefficient vector

$$\boldsymbol{\theta}^\circ = (\beta_0^\circ, \boldsymbol{\beta}^{\circ\top})^\top, \text{ with } \beta_0^\circ = \log \frac{P(Y=1)}{P(Y=0)} \text{ and } \beta_j^\circ = \mathbb{1}(j \in \mathcal{I}), \text{ for } j \in [p].$$

Then $\eta^*(\mathbf{X}) = \beta_0^\circ + \mathbf{T}^*(\mathbf{X})^\top \boldsymbol{\beta}^\circ$. Condition (A3) requires $M > |\beta_0^\circ| + |\mathcal{I}| = \|\boldsymbol{\theta}^\circ\|_1$, so that $\boldsymbol{\theta}^\circ$ lies in the interior of the compact parameter space $\boldsymbol{\Theta}$.

For a generic $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\top)^\top \in \boldsymbol{\Theta}$, write $\eta_{\boldsymbol{\theta}}(\mathbf{X}) = \beta_0 + \mathbf{T}^*(\mathbf{X})^\top \boldsymbol{\beta}$, and define the set of population minimizers $\boldsymbol{\Theta}_0 = \arg \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q_0(\boldsymbol{\theta})$. The difference between the population logistic risks at $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\circ$ can be written as

$$Q_0(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta}^\circ) = E_{\mathbf{X}} [D_{\text{KL}} \{ \text{Bern}(\pi^*(\mathbf{X})) \parallel \text{Bern}(\mathcal{G}\{\eta_{\boldsymbol{\theta}}(\mathbf{X})\}) \}] \quad (\text{A.2})$$

The integrand in (A.2) is nonnegative and is zero if and only if its two Bernoulli success probabilities agree. Since $\mathcal{G}$ is strictly increasing, it follows that $\Theta_0 = \{\boldsymbol{\theta} \in \Theta : \eta_{\boldsymbol{\theta}}(\mathbf{X}) = \eta^*(\mathbf{X}) \text{ almost surely}\}$. In particular, $\boldsymbol{\theta}^\circ \in \Theta_0$. Moreover, because $T_j^*(X_j) \equiv 0$ for $j \in \mathcal{I}_N$, changing the noise coefficients of $\boldsymbol{\theta}^\circ$ without leaving Θ does not change the oracle linear predictor. Thus, the coefficient vector need not be identifiable, although the population linear predictor is identifiable.

Next, we establish consistency for the set Θ_0 , i.e.,

$$d(\hat{\boldsymbol{\theta}}, \Theta_0) := \inf_{\boldsymbol{\theta}_0 \in \Theta_0} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \xrightarrow{P} 0.$$

Because Θ_0 is a set rather than a single point, we use the set-valued form of the argmin consistency argument (van der Vaart, 2000, Theorem 5.7).

The set Θ_0 is well separated in the sense that

$$\inf_{\boldsymbol{\theta} \in \Theta: d(\boldsymbol{\theta}, \Theta_0) \geq \epsilon} Q_0(\boldsymbol{\theta}) > \inf_{\boldsymbol{\theta} \in \Theta} Q_0(\boldsymbol{\theta}) \quad \text{for every } \epsilon > 0.$$

This is because, if this failed for some $\epsilon > 0$, compactness of Θ and continuity of Q_0 would produce a convergent sequence $\{\boldsymbol{\theta}^{(m)}\}$ with $d(\boldsymbol{\theta}^{(m)}, \Theta_0) \geq \epsilon$ and $Q_0(\boldsymbol{\theta}^{(m)}) \rightarrow \inf_{\Theta} Q_0$, whose limit would belong to Θ_0 , a contradiction. Given this separation, it suffices to show that $Q_n(\boldsymbol{\theta})$ converges uniformly to $Q_0(\boldsymbol{\theta})$ over the compact parameter space Θ .

To achieve this, we define the negative log-likelihood evaluated at one

data point,

$$\ell_i(\boldsymbol{\theta}; Y_i, \mathbf{T}(\mathbf{X}_i)) = [-Y_i \{\beta_0 + \mathbf{T}(\mathbf{X}_i)^\top \boldsymbol{\beta}\} + \log(1 + \exp(\beta_0 + \mathbf{T}(\mathbf{X}_i)^\top \boldsymbol{\beta}))].$$

By the triangle inequality, $|Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})|$ is no larger than

$$\left| Q_n(\boldsymbol{\theta}) - \frac{1}{n} \sum_{i=1}^n \ell_i(\boldsymbol{\theta}; Y_i, \mathbf{T}^*(\mathbf{X}_i)) \right| + \left| \frac{1}{n} \sum_{i=1}^n \ell_i(\boldsymbol{\theta}; Y_i, \mathbf{T}^*(\mathbf{X}_i)) - Q_0(\boldsymbol{\theta}) \right|. \quad (\text{A.3})$$

Consider the second difference in (A.3). Set $\mathbf{Z} = (\mathbf{X}^\top, Y)^\top \in \mathbb{R}^{p+1}$ and, for

$\mathbf{z} = (\mathbf{x}^\top, y)^\top \in \mathbb{R}^{p+1}$, define

$$a(\mathbf{z}, \boldsymbol{\theta}) = -y \{\beta_0 + \mathbf{T}^*(\mathbf{x})^\top \boldsymbol{\beta}\} + \log(1 + \exp\{\beta_0 + \mathbf{T}^*(\mathbf{x})^\top \boldsymbol{\beta}\}),$$

so that $\ell_i(\boldsymbol{\theta}; Y_i, \mathbf{T}^*(\mathbf{X}_i)) = a(\mathbf{Z}_i, \boldsymbol{\theta})$ and $Q_0(\boldsymbol{\theta}) = E\{a(\mathbf{Z}_1, \boldsymbol{\theta})\}$. The vectors

$\mathbf{Z}_1, \dots, \mathbf{Z}_n$ form a random sample from the distribution of $\mathbf{Z}$ on $\mathbb{R}^{p+1}$, be-

cause $\mathbf{T}^*(\cdot)$ is a fixed function that does not depend on the sample. The

function $a(\mathbf{z}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for each $\mathbf{z}$ and measurable in $\mathbf{z}$ for each

$\boldsymbol{\theta}$, and $\boldsymbol{\Theta}$ is compact. Moreover, since $\log(1 + e^u) \leq \log 2 + |u|$ and, under

Condition (A3), $|\beta_0 + \mathbf{T}^*(\mathbf{x})^\top \boldsymbol{\beta}| \leq M\{1 + \|\mathbf{T}^*(\mathbf{x})\|_\infty\}$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we

have the envelope

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |a(\mathbf{z}, \boldsymbol{\theta})| \leq h(\mathbf{z}) := \log 2 + (1 + |y|) M \{1 + \|\mathbf{T}^*(\mathbf{x})\|_\infty\} \quad \text{for every } \mathbf{z} \in \mathbb{R}^{p+1},$$

with $E\{h(\mathbf{Z}_1)\} < \infty$ because $|Y| \leq 1$ almost surely and $E\|\mathbf{T}^*(\mathbf{X}_1)\|_\infty < \infty$,

the latter following from $|T_j^*(x)| = O(\log |x|)$ and $E\{(\log |X_{1j}|)^3\} < \infty$,

both established in the first part of the proof of Lemma 3 in Appendix D.

The uniform law of large numbers (Jennrich, 1969, Theorem 2) therefore gives

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{i=1}^n \ell_i(\boldsymbol{\theta}; Y_i, \mathbf{T}^*(\mathbf{X}_i)) - Q_0(\boldsymbol{\theta}) \right| \longrightarrow 0 \quad \text{almost surely,}$$

so that the second difference in (A.3) converges to zero in probability uniformly over $\boldsymbol{\Theta}$. The first absolute difference in (A.3) is equal to

$$\left| \frac{1}{n} \sum_{i=1}^n \left\{ \ell_i(\boldsymbol{\theta}; Y_i, \hat{\mathbf{T}}(\mathbf{X}_i)) - \ell_i(\boldsymbol{\theta}; Y_i, \mathbf{T}^*(\mathbf{X}_i)) \right\} \right|.$$

By the Lagrange mean value theorem, since $(\partial/\partial \mathbf{t}) \ell_i(\boldsymbol{\theta}; Y_i, \mathbf{t}) = \{\mathcal{G}(\beta_0 + \mathbf{t}^\top \boldsymbol{\beta}) - Y_i\} \boldsymbol{\beta}$ and $|\mathcal{G}(\cdot) - Y_i| \leq 1$,

$$\left| \ell_i(\boldsymbol{\theta}; Y_i, \hat{\mathbf{T}}(\mathbf{X}_i)) - \ell_i(\boldsymbol{\theta}; Y_i, \mathbf{T}^*(\mathbf{X}_i)) \right| \leq \left| \sum_{j=1}^p \beta_j \left\{ \hat{T}_j(X_{ij}) - T_j^*(X_{ij}) \right\} \right|.$$

Moreover,

$$\begin{aligned} \left| \sum_{j=1}^p \beta_j \left\{ \hat{T}_j(X_{ij}) - T_j^*(X_{ij}) \right\} \right| &\leq \max_{j \in [p]} \left| \hat{T}_j(X_{ij}) - T_j^*(X_{ij}) \right| \cdot \sum_{j=1}^p |\beta_j| \\ &\leq M \max_{i \in [n]} \max_{j \in [p]} \left| \hat{T}_j(X_{ij}) - T_j^*(X_{ij}) \right|, \end{aligned}$$

where M is a constant satisfying $\sum_{j=1}^p |\beta_j| \leq M$ under Condition (A3).

The rightmost bound depends on neither i nor $\boldsymbol{\theta}$, and it is $o_p(1)$ by (D.8) of Appendix D. The pointwise consistency $\hat{T}_j(x) \xrightarrow{p} T_j^*(x)$ at a fixed x , established in the first part of this proof, does not suffice at this step, because

the arguments X_{ij} are random and their number grows with n ; it is invoked later, where the conclusion drawn is itself pointwise in $\mathbf{x}$. It follows that

$$\begin{aligned} & \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \ell_i(\boldsymbol{\theta}; Y_i, \hat{\mathbf{T}}(\mathbf{X}_i)) - \ell_i(\boldsymbol{\theta}; Y_i, \mathbf{T}^*(\mathbf{X}_i)) \right\} \right| \\ & \leq M \max_{i \in [n]} \max_{j \in [p]} \left| \hat{T}_j(X_{ij}) - T_j^*(X_{ij}) \right| \xrightarrow{p} 0, \end{aligned}$$

and thus, returning to (A.3), we have $\sup_{\boldsymbol{\theta}} |Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \xrightarrow{p} 0$. By the display in the paragraph introducing $\mathcal{L}_n$, we also have $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\mathcal{L}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \xrightarrow{p} 0$. Combined with the separation property of $\boldsymbol{\Theta}_0$ established above, this gives $d(\hat{\boldsymbol{\theta}}, \boldsymbol{\Theta}_0) \xrightarrow{p} 0$.

It remains to convert this set consistency into convergence of the fitted linear predictor. Pick $\boldsymbol{\theta}_n = (\beta_{0,n}, \boldsymbol{\beta}_n^\top)^\top \in \boldsymbol{\Theta}_0$ with $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n\|_1 \xrightarrow{p} 0$. Such a choice exists because $\boldsymbol{\Theta}_0$ is closed and p is fixed, so all norms on $\mathbb{R}^{p+1}$ are equivalent. Since $T_j^*(\cdot) \equiv 0$ identically for every $j \in \mathcal{I}_N$, and since every element of $\boldsymbol{\Theta}_0$ reproduces η^* , we have $\eta_{\boldsymbol{\theta}_n}(\mathbf{x}) = \eta^*(\mathbf{x})$ for $P_{\mathbf{X}}$ -almost every $\mathbf{x}$. For such an $\mathbf{x}$,

$$\hat{\eta}(\mathbf{x}) - \eta^*(\mathbf{x}) = (\hat{\beta}_0 - \beta_{0,n}) + \left\{ \hat{\mathbf{T}}(\mathbf{x}) - \mathbf{T}^*(\mathbf{x}) \right\}^\top \hat{\boldsymbol{\beta}} + \mathbf{T}^*(\mathbf{x})^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n).$$

The first and third terms are bounded by $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n\|_1 \max\{1, \|\mathbf{T}^*(\mathbf{x})\|_\infty\} = o_p(1)$, and the second is bounded by $\|\hat{\mathbf{T}}(\mathbf{x}) - \mathbf{T}^*(\mathbf{x})\|_\infty \|\hat{\boldsymbol{\beta}}\|_1 \leq M \|\hat{\mathbf{T}}(\mathbf{x}) - \mathbf{T}^*(\mathbf{x})\|_\infty = o_p(1)$ by the first part of this proof and Condition (A3). Hence $\hat{\eta}(\mathbf{x}) \xrightarrow{p} \eta^*(\mathbf{x})$, and consequently $\hat{\pi}(\mathbf{x}) \xrightarrow{p} \pi^*(\mathbf{x})$ by the continuous mapping

theorem.

Finally, we are ready to derive the limit of the classification risk of the S-RoLLR classifier, $\hat{g}_n(\mathbf{x}) = \mathbb{1}(\hat{\pi}(\mathbf{x}) > 0.5)$. The optimal Bayes classifier is $g^*(\mathbf{x}) = \mathbb{1}(\pi^*(\mathbf{x}) > 0.5)$, with the risk $R^* = E_{\mathbf{X}}\{\min(\pi^*(\mathbf{X}), 1 - \pi^*(\mathbf{X}))\}$. Because the discontinuity set of the indicator function has measure zero with respect to the distribution of $\mathbf{X}$, the continuous mapping theorem applies and allows one to conclude $\hat{g}_n(\mathbf{x}) \xrightarrow{p} g^*(\mathbf{x})$ following the second part of the proof above leading to $\hat{\pi}(\mathbf{x}) \xrightarrow{p} \pi^*(\mathbf{x})$. Using the law of iterated expectations, we elaborate the risk function of $\hat{g}_n(\mathbf{x})$ as follows,

$$\begin{aligned}
 R(\hat{g}_n) &= E_{\mathbf{X}, Y} \{ \mathbb{1}(\hat{g}_n(\mathbf{X}) \neq Y) \} \\
 &= E_{\mathbf{X}} \{ P(\hat{g}_n(\mathbf{X}) \neq Y \mid \mathbf{X}) \} \\
 &= E_{\mathbf{X}} \{ P(Y = 1 \mid \mathbf{X}) \mathbb{1}(\hat{g}_n(\mathbf{X}) = 0) + P(Y = 0 \mid \mathbf{X}) \mathbb{1}(\hat{g}_n(\mathbf{X}) = 1) \} \\
 &= E_{\mathbf{X}} [\pi^*(\mathbf{X}) \mathbb{1}(\hat{g}_n(\mathbf{X}) = 0) + \{1 - \pi^*(\mathbf{X})\} \mathbb{1}(\hat{g}_n(\mathbf{X}) = 1)].
 \end{aligned}$$

Because the expression inside $E_{\mathbf{X}}[\cdot]$ above is bounded, by the dominated convergence theorem,

$$\begin{aligned}
 \lim_{n \rightarrow \infty} R(\hat{g}_n) &= E_{\mathbf{X}} \left\{ \lim_{n \rightarrow \infty} [\pi^*(\mathbf{X}) \mathbb{1}(\hat{g}_n(\mathbf{X}) = 0) + \{1 - \pi^*(\mathbf{X})\} \mathbb{1}(\hat{g}_n(\mathbf{X}) = 1)] \right\} \\
 &= E_{\mathbf{X}} [\pi^*(\mathbf{X}) \mathbb{1}(g^*(\mathbf{X}) = 0) + \{1 - \pi^*(\mathbf{X})\} \mathbb{1}(g^*(\mathbf{X}) = 1)] \\
 &= E_{\mathbf{X}} \{ \min(\pi^*(\mathbf{X}), 1 - \pi^*(\mathbf{X})) \} \\
 &= R^*,
 \end{aligned}$$

This completes the proof. $\square$

Appendix B: The Proof of Theorem 2

Throughout this appendix, T_j^* denotes the log ratio of the two true class-specific densities of X_j , as defined in Section 3.2 of the main article.

B.1 The Proof of Lemma 1

Before proving Theorem 2, we establish Lemma 1 first. For $j \in [p]$ and $k \in \{0, 1\}$, denote by $\Psi_{jk}^\dagger$ the model parameters in the Kullback–Leibler projection of the true class-conditional feature density $\tilde{f}_j(\cdot \mid Y = k)$ onto the TPSC family, and define $e_{jk}(t) = \log \tilde{f}_j(t \mid Y = k) - \log f_{\text{TPSC}}(t \mid \Psi_{jk}^\dagger)$. Condition (A5) stated in the main article assumes that $E\{e_{jk}^2(X_j)\} \leq \delta_f$ for $j \in \mathcal{I}$ and $k \in \{0, 1\}$. For $j \in \mathcal{I}_N$ the two class-conditional densities coincide, $\tilde{f}_j(\cdot \mid Y = 0) = \tilde{f}_j(\cdot \mid Y = 1)$. They therefore share a Kullback–Leibler projection onto the TPSC family, $\Psi_{j1}^\dagger = \Psi_{j0}^\dagger$, and the two fitting errors coincide as well, $e_{j1}(X_j) \equiv e_{j0}(X_j)$. It follows that, for $j \in \mathcal{I}_N$, $E[\{e_{j1}(X_j) - e_{j0}(X_j)\}^2] = 0$. Lemma 1 concerns this mean squared difference for $j \in \mathcal{I}$.

Lemma 1. *Under Condition (A5), $j \in \mathcal{I}$,*

$$E \left[\{e_{j1}(X_j) - e_{j0}(X_j)\}^2 \right] \leq \begin{cases} 2\delta_f(1 - \rho_j), & \text{if } \rho_j \leq 0.5, \\ \delta_f, & \text{if } \rho_j > 0.5, \end{cases} \quad (\text{B.1})$$

where

$$\rho_j = \frac{E \{e_{j1}(X_j)e_{j0}(X_j)\}}{\sqrt{E \{e_{j1}^2(X_j)\} E \{e_{j0}^2(X_j)\}}}.$$

If either second moment in the denominator is zero, we set $\rho_j = 1/2$ by convention.

Proof. Let $a = E\{e_{j1}^2(X_j)\}$ and $b = E\{e_{j0}^2(X_j)\}$. If either a or b is zero, the left-hand side of (B.1) is $a + b \leq \delta_f$, and the result follows from the convention $\rho_j = 1/2$. It therefore remains to consider $a, b > 0$. By the definition of ρ_j , we have

$$\begin{aligned} & E \left[\{e_{j1}(X_j) - e_{j0}(X_j)\}^2 \right] \\ &= E \{e_{j1}^2(X_j)\} + E \{e_{j0}^2(X_j)\} - 2\rho_j \sqrt{E \{e_{j1}^2(X_j)\} E \{e_{j0}^2(X_j)\}}. \end{aligned} \quad (\text{B.2})$$

Under Condition (A5), $a, b \in (0, \delta_f]$. Then (B.2) can be written as a non-negative function of a and b , $f(a, b) = a + b - 2\rho_j \sqrt{ab}$, defined in the domain $\mathcal{D} = [0, \delta_f] \times [0, \delta_f]$. Next, we find the maximum of $f(a, b)$ over $\mathcal{D}$.

We first focus on the interior of $\mathcal{D}$ and consider interior stationary points of $f(a, b)$. Setting the gradient of $f(a, b)$ to zero gives the following

equations that an interior stationary point satisfies,

$$\frac{\partial}{\partial a} f(a, b) = 1 - \rho_j \sqrt{\frac{b}{a}} = 0, \quad (\text{B.3})$$

$$\frac{\partial}{\partial b} f(a, b) = 1 - \rho_j \sqrt{\frac{a}{b}} = 0, \quad (\text{B.4})$$

where (B.3) is equivalent to $\sqrt{b/a} = 1/\rho_j$, i.e., $\sqrt{a/b} = \rho_j (> 0)$. Substituting $\sqrt{a/b}$ in (B.4) with ρ_j gives $\rho_j^2 = 1$, and thus $\rho_j = 1$ by the positivity constraint on ρ_j according (B.3). With $\rho_j = 1$, (B.3) yields $a = b$, suggesting that interior stationary points of the non-negative $f(a, b)$ lie on the line in $\mathcal{D}$, $\{(a, b) \in \mathcal{D} : 0 < a = b < \delta_f\}$, where $f(a, b)$ reaches its minimum of zero since, with $\rho_j = 1$, $f(a, a) = 0$ for any $a > 0$. Therefore, the maximum of $f(a, b)$ cannot be reached in the interior of $\mathcal{D}$.

We next turn to the boundary of $\mathcal{D}$ to seek the maximum of $f(a, b)$, $\mathcal{D}_0 \cup \mathcal{D}_1$, where $\mathcal{D}_0 = \{(a, b) \in \mathcal{D} : a = 0 \text{ or } b = 0\}$ and $\mathcal{D}_1 = \{(a, b) \in \mathcal{D} : a = \delta_f \text{ or } b = \delta_f\}$. On $\mathcal{D}_0$, the function reduces to $f(0, b) = b \leq \delta_f$ or $f(a, 0) = a \leq \delta_f$, and thus $\max_{\mathcal{D}_0} f(a, b) = \delta_f$. On $\mathcal{D}_1$, take the half of $\mathcal{D}_1$ given by $\{(a, b) \in \mathcal{D} : a = \delta_f\}$ as an example, where $f(a, b) = f(\delta_f, b) = \delta_f + b - 2\rho_j \sqrt{\delta_f b}$ reduces to a univariate function of b . Clearly, if $\rho_j \leq 0$, $f(\delta_f, b)$ is an increasing function of b , and thus its maximum is attained at $b = \delta_f$. On the other hand, if $\rho_j > 0$, the second derivative of this univariate function is $\rho_j \sqrt{\delta_f} / (2b^{3/2}) > 0$. This indicates that $f(\delta_f, b)$ is a convex function (of b) when $\rho_j > 0$, and thus its maximum can only

be reached at $b = 0$ or δ_f as two endpoints of the plausible range of b in $\mathcal{D}$. By symmetry of the other half of $\mathcal{D}_1$ given by $\{(a, b) \in \mathcal{D} : b = \delta_f\}$, $f(a, \delta_f) \leq f(\delta_f, \delta_f)$ if $\rho_j \leq 0$. Otherwise, the maximum of $f(a, \delta_f)$ as a function of a can only be attained at the two endpoints $a = 0$ or δ_f .

In summary, the global maximum of $f(a, b)$ can only occur at the corner points of $\mathcal{D}$: $(0, 0)$, $(\delta_f, 0)$, $(0, \delta_f)$, and (δ_f, δ_f) . Clearly, $f(0, 0) = 0$ is the minimum of $f(a, b)$. Evaluating $f(a, b)$ at the latter three points gives $f(\delta_f, 0) = f(0, \delta_f) = \delta_f$ and $f(\delta_f, \delta_f) = 2\delta_f(1 - \rho_j)$. By the Cauchy–Schwarz inequality, $\rho_j \in [-1, 1]$. Comparing δ_f and $2\delta_f(1 - \rho_j)$ reveals that $\delta_f \leq 2\delta_f(1 - \rho_j)$ is equivalent to $\rho_j \leq 0.5$. Hence, the maximum of $f(a, b)$ is $2\delta_f(1 - \rho_j)$ if $\rho_j \in [-1, 0.5]$, reached at $(a, b) = (\delta_f, \delta_f)$. Otherwise, the maximum of $f(a, b)$ is δ_f , attained at $(a, b) = (\delta_f, 0), (0, \delta_f)$. This proves (B.1). $\square$

B.2 Proving Theorem 2

Now, we are ready to prove Theorem 2 in the main article, reiterated below for completeness.

Theorem 2. *Under Conditions (A3)–(A7), the S-RoLLR classification risk $R(\hat{g}_n)$ converges to $R^* + \Delta R$ as $n \rightarrow \infty$, where the asymptotic excess risk*

ΔR compared with the optimal Bayes classifier satisfies

$$\Delta R \leq \sqrt{\frac{1}{2} [\delta_{lo} + \{2|J_1|(1 - \bar{\rho}_1) + |J_2|\} M^2 \delta_f]}.$$

Here, M and δ_f are given in (A3) and (A5), respectively, the index sets

$$J_1 = \{j \in \mathcal{I} : -1 \leq \rho_j \leq 0.5\} \text{ and } J_2 = \{j \in \mathcal{I} : 0.5 < \rho_j \leq 1\}$$

partition the signal set $\mathcal{I}$, $|J_\ell|$ denotes the cardinality of J_ℓ , for $\ell = 1, 2$,

and $\bar{\rho}_1 = |J_1|^{-1} \sum_{j \in J_1} \rho_j$, with $\bar{\rho}_1 = 0$ when $J_1 = \emptyset$, in which

$$\rho_j = \frac{E\{e_{j1}(X_j)e_{j0}(X_j)\}}{\sqrt{E\{e_{j1}^2(X_j)\}E\{e_{j0}^2(X_j)\}}}, \quad j \in \mathcal{I}. \quad (\text{B.5})$$

If either second moment in the denominator of (B.5) is zero, we set $\rho_j = 1/2$ by convention.

Proof. Corresponding to the pseudo truths of the TPSC parameters, $\Psi_{jk}^\dagger$,

define pseudo oracle features $T_j^\dagger(X_j) = \log(f_{\text{TPSC}}(X_j|\Psi_{j1}^\dagger)/f_{\text{TPSC}}(X_j|\Psi_{j0}^\dagger))$,

for $j \in [p]$, and the corresponding linear predictor $\eta^\dagger(\mathbf{X}) = \beta_0^\dagger + \mathbf{T}^\dagger(\mathbf{X})^\top \boldsymbol{\beta}^\dagger$,

where $\mathbf{T}^\dagger(\mathbf{X}) = (T_1^\dagger(X_1), \dots, T_p^\dagger(X_p))^\top$, and

$$(\beta_0^\dagger, \boldsymbol{\beta}^\dagger) = \arg \min_{(\beta_0, \boldsymbol{\beta}^\top)^\top \in \boldsymbol{\Theta}} E_{\mathbf{X}, Y} [-Y \{\beta_0 + \mathbf{T}^\dagger(\mathbf{X})^\top \boldsymbol{\beta}\} + \log(1 + \exp(\beta_0 + \mathbf{T}^\dagger(\mathbf{X})^\top \boldsymbol{\beta}))]$$

are the pseudo truths of the regression coefficients when regressing Y on

$\mathbf{T}^\dagger(\mathbf{X})$. For a point $\mathbf{x}$ in the original feature space, let $\pi^\dagger(\mathbf{x}) = \mathcal{G}(\eta^\dagger(\mathbf{x})) =$

$1/\{1 + \exp(-\eta^\dagger(\mathbf{x}))\}$, and $g^\dagger(\mathbf{x}) = \mathbb{1}(\pi^\dagger(\mathbf{x}) > 0.5)$. As in Appendix A, the

minimizing pair $(\beta_0^\dagger, \boldsymbol{\beta}^\dagger)$ need not be unique: for $j \in \mathcal{I}_N$ the two Kullback–

Leibler projection problems coincide, so $\Psi_{j1}^\dagger = \Psi_{j0}^\dagger$ and hence $T_j^\dagger(\cdot) \equiv 0$,

and the coefficients attached to noise features are unconstrained. However, the map $\eta \mapsto E_{\mathbf{X},Y}\{-Y\eta(\mathbf{X}) + \log(1 + e^{\eta(\mathbf{X})})\}$ is strictly convex in η pointwise, so any two minimizers over the convex class $\mathcal{H}_2$ would have their midpoint attaining a strictly smaller value unless they agree almost surely. Consequently the pseudo-true linear predictor $\eta^\dagger$, and therefore $\pi^\dagger$, $g^\dagger$ and $R(g^\dagger)$, are well defined, even though $\beta^\dagger$ is not.

According to White (1982), $\hat{\Psi}_{jk} \xrightarrow{p} \Psi_{jk}^\dagger$, for $j \in [p], k = 0, 1$, following which one can show that $\lim_{n \rightarrow \infty} R(\hat{g}_n) = R(g^\dagger)$, by repeating the set-valued argument of Appendix A with Θ_0 replaced by the set of minimizers defining $(\beta_0^\dagger, \beta^\dagger)$. In what follows, we derive an upper bound for $\Delta R = R(g^\dagger) - R^*$ in three steps.

In the first step, we form an upper bound for ΔR involving membership probabilities. In particular, by Corollary 6.2 in Devroye et al. (2013),

$$\Delta R \leq 2\sqrt{E_{\mathbf{X}} \left[\{\pi^\dagger(\mathbf{X}) - \pi^*(\mathbf{X})\}^2 \right]}. \quad (\text{B.6})$$

Zooming in on the squared difference in (B.6), by Lemma 11.6.1 in Cover and Thomas (2006), we have

$$\{\pi^\dagger(\cdot) - \pi^*(\cdot)\}^2 \leq \frac{1}{2} D_{\text{KL}}(\text{Bernoulli}(\pi^*(\cdot)) \parallel \text{Bernoulli}(\pi^\dagger(\cdot))). \quad (\text{B.7})$$

To elaborate the Kullback–Leibler divergence in (B.7), we define the nega-

tive log-likelihood of an observation Y from $\text{Bernoulli}(\pi(\mathbf{X}))$,

$$Q(Y, \pi(\mathbf{X})) = -Y \log(\pi(\mathbf{X})) - (1 - Y) \log(1 - \pi(\mathbf{X})). \quad (\text{B.8})$$

Then

$$D_{\text{KL}}(\text{Bernoulli}(\pi^*(\mathbf{X})) \parallel \text{Bernoulli}(\pi^\dagger(\mathbf{X}))) = E_{Y|\mathbf{X}} \{Q(Y, \pi^\dagger(\mathbf{X})) - Q(Y, \pi^*(\mathbf{X}))\}.$$

and $E_{\mathbf{X}}[D_{\text{KL}}(\text{Bernoulli}(\pi^*(\mathbf{X})) \parallel \text{Bernoulli}(\pi^\dagger(\mathbf{X})))] = \mathcal{L}(\pi^\dagger) - \mathcal{L}(\pi^*)$, where

$\mathcal{L}(\pi) = E_{\mathbf{X}, Y} \{Q(Y, \pi(\mathbf{X}))\}$. Now (B.7) implies

$$E_{\mathbf{X}} \left[\{\pi^\dagger(\mathbf{X}) - \pi^*(\mathbf{X})\}^2 \right] \leq \frac{1}{2} \{ \mathcal{L}(\pi^\dagger) - \mathcal{L}(\pi^*) \}.$$

Combining with (B.6), we have

$$\Delta R \leq 2 \sqrt{E_{\mathbf{X}} \left[\{\pi^\dagger(\mathbf{X}) - \pi^*(\mathbf{X})\}^2 \right]} \leq \sqrt{2 \{ \mathcal{L}(\pi^\dagger) - \mathcal{L}(\pi^*) \}}. \quad (\text{B.9})$$

In the second step, we find an upper bound for the difference, $\mathcal{L}(\pi^\dagger) - \mathcal{L}(\pi^*)$, in (B.9). To achieve this, we define two classes of linear predictors.

One is the class of linear predictors that are linear in the oracle features

$\mathbf{T}^*(\mathbf{X})$, $\mathcal{H}_1 = \{a + \mathbf{T}^*(\mathbf{X})^\top \mathbf{b} : (a, \mathbf{b}^\top)^\top \in \Theta\}$. The other one is the class

of linear predictors that are linear in the pseudo oracle features $\mathbf{T}^\dagger(\mathbf{X})$,

$\mathcal{H}_2 = \{a + \mathbf{T}^\dagger(\mathbf{X})^\top \mathbf{b} : (a, \mathbf{b}^\top)^\top \in \Theta\}$. Define

$$\begin{aligned} \eta_{\text{proj}}(\mathbf{X}) &= a^* + \mathbf{T}^*(\mathbf{X})^\top \mathbf{b}^* \\ &= \arg \min_{\eta \in \mathcal{H}_1} E_{\mathbf{X}} \left[\{\eta^*(\mathbf{X}) - \eta(\mathbf{X})\}^2 \right], \text{ and} \\ \eta_{\text{proj}}^\dagger(\mathbf{X}) &= \arg \min_{\eta \in \mathcal{H}_2} E_{\mathbf{X}} \left[\{\eta_{\text{proj}}(\mathbf{X}) - \eta(\mathbf{X})\}^2 \right]. \end{aligned}$$

Let $\pi_{\text{proj}}(\mathbf{X}) = \mathcal{G}(\eta_{\text{proj}}(\mathbf{X}))$ and $\pi_{\text{proj}}^\dagger(\mathbf{X}) = \mathcal{G}(\eta_{\text{proj}}^\dagger(\mathbf{X}))$.

Because $\pi^\dagger(\mathbf{X})$ is based on the pseudo truths of the regression coefficients (that minimizes the mean of the negative log-likelihood of Y given $\mathbf{X}$), we have $\mathcal{L}(\pi^\dagger) \leq \mathcal{L}(\pi_{\text{proj}}^\dagger)$, and thus $\mathcal{L}(\pi^\dagger) - \mathcal{L}(\pi^*) \leq \mathcal{L}(\pi_{\text{proj}}^\dagger) - \mathcal{L}(\pi^*)$. With $\mathcal{L}(\cdot)$ defined via the negative log-likelihood $Q(Y, \pi(\mathbf{X}))$ in (B.8), we now view $Q(Y, \pi(\mathbf{X}))$ as a function of the linear predictor η (with its dependence on $\mathbf{X}$ suppressed), $J(\eta) = -Y \log(\mathcal{G}(\eta)) - (1 - Y) \log(1 - \mathcal{G}(\eta))$. A second-order Taylor's expansion of $J(\eta_{\text{proj}}^\dagger)$ around η^* gives

$$\begin{aligned} J(\eta_{\text{proj}}^\dagger) &= J(\eta^*) + J'(\eta^*) \left(\eta_{\text{proj}}^\dagger - \eta^* \right) + \frac{1}{2} J''(\tilde{\eta}) \left(\eta_{\text{proj}}^\dagger - \eta^* \right)^2 \\ &= J(\eta^*) + \{ \mathcal{G}(\eta^*) - Y \} \left(\eta_{\text{proj}}^\dagger - \eta^* \right) + \frac{1}{2} \mathcal{G}(\tilde{\eta}) \{ 1 - \mathcal{G}(\tilde{\eta}) \} \left(\eta_{\text{proj}}^\dagger - \eta^* \right)^2, \end{aligned}$$

where $\tilde{\eta}$ lies between $\eta_{\text{proj}}^\dagger$ and η^* . Returning to the function $Q(Y, \pi(\mathbf{X}))$, the preceding Taylor expansion is equivalent to

$$\begin{aligned} &Q(Y, \pi_{\text{proj}}^\dagger(\mathbf{X})) - Q(Y, \pi^*(\mathbf{X})) \\ &= \{ \mathcal{G}(\eta^*(\mathbf{X})) - Y \} \left\{ \eta_{\text{proj}}^\dagger(\mathbf{X}) - \eta^*(\mathbf{X}) \right\} + \\ &\quad \frac{1}{2} \mathcal{G}(\tilde{\eta}(\mathbf{X})) \{ 1 - \mathcal{G}(\tilde{\eta}(\mathbf{X})) \} \left\{ \eta_{\text{proj}}^\dagger(\mathbf{X}) - \eta^*(\mathbf{X}) \right\}^2, \end{aligned}$$

following which we have

$$\begin{aligned}
 \mathcal{L}(\pi_{\text{proj}}^\dagger) - \mathcal{L}(\pi^*) &= E_{\mathbf{X}, Y} \left\{ Q(Y, \pi_{\text{proj}}^\dagger(\mathbf{X})) - Q(Y, \pi^*(\mathbf{X})) \right\} \\
 &= E_{\mathbf{X}, Y} \left[\{ \mathcal{G}(\eta^*(\mathbf{X})) - Y \} \left\{ \eta_{\text{proj}}^\dagger(\mathbf{X}) - \eta^*(\mathbf{X}) \right\} \right] + \\
 &\quad \frac{1}{2} E_{\mathbf{X}, Y} \left[\mathcal{G}(\tilde{\eta}(\mathbf{X})) \{ 1 - \mathcal{G}(\tilde{\eta}(\mathbf{X})) \} \left\{ \eta_{\text{proj}}^\dagger(\mathbf{X}) - \eta^*(\mathbf{X}) \right\}^2 \right] \\
 &= \frac{1}{2} E_{\mathbf{X}, Y} \left[\mathcal{G}(\tilde{\eta}(\mathbf{X})) \{ 1 - \mathcal{G}(\tilde{\eta}(\mathbf{X})) \} \left\{ \eta_{\text{proj}}^\dagger(\mathbf{X}) - \eta^*(\mathbf{X}) \right\}^2 \right] \\
 &\leq \frac{1}{8} E_{\mathbf{X}} \left[\left\{ \eta_{\text{proj}}^\dagger(\mathbf{X}) - \eta^*(\mathbf{X}) \right\}^2 \right], \tag{B.10}
 \end{aligned}$$

where the second to last equation is because $E(Y|\mathbf{X}) = \mathcal{G}(\eta^*(\mathbf{X}))$ and thus $E_{\mathbf{X}, Y}[\{ \mathcal{G}(\eta^*(\mathbf{X})) - Y \} \{ \eta_{\text{proj}}^\dagger(\mathbf{X}) - \eta^*(\mathbf{X}) \}] = 0$ by the law of iterated expectation, and the last inequality is because $\mathcal{G}(t)\{1 - \mathcal{G}(t)\} \in [0, 1/4]$.

In the third step, we find an upper bound for the expectation in (B.10).

Using $\eta_{\text{proj}}(\mathbf{X})$ to bridge the gap between the two linear predictors in (B.10),

we have

$$\begin{aligned}
 E_{\mathbf{X}} \left[\left\{ \eta_{\text{proj}}^\dagger(\mathbf{X}) - \eta^*(\mathbf{X}) \right\}^2 \right] &= E_{\mathbf{X}} \left[\left\{ \eta_{\text{proj}}^\dagger(\mathbf{X}) - \eta_{\text{proj}}(\mathbf{X}) + \eta_{\text{proj}}(\mathbf{X}) - \eta^*(\mathbf{X}) \right\}^2 \right] \\
 &\leq 2E_{\mathbf{X}} \left[\left\{ \eta_{\text{proj}}^\dagger(\mathbf{X}) - \eta_{\text{proj}}(\mathbf{X}) \right\}^2 \right] + 2E_{\mathbf{X}} \left[\left\{ \eta_{\text{proj}}(\mathbf{X}) - \eta^*(\mathbf{X}) \right\}^2 \right], \tag{B.11}
 \end{aligned}$$

where the last inequality is due to $(s + t)^2 \leq 2s^2 + 2t^2$, and the second expectation in (B.11) is bounded from above by δ_{l_0} under Conditions (A6) and (A7). To bound the first expectation in (B.11) from above, we introduce $h(\mathbf{X}) = a^* + \mathbf{T}^\dagger(\mathbf{X})^\top \mathbf{b}^* \in \mathcal{H}_2$. By the construction of $\eta_{\text{proj}}^\dagger(\mathbf{X})$,

$$\begin{aligned}
 & E_{\mathbf{X}} \left[\left\{ \eta_{\text{proj}}^{\dagger}(\mathbf{X}) - \eta_{\text{proj}}(\mathbf{X}) \right\}^2 \right] \\
 & \leq E_{\mathbf{X}} \left[\{h(\mathbf{X}) - \eta_{\text{proj}}(\mathbf{X})\}^2 \right] \\
 & = E_{\mathbf{X}} \left[\{a^* + \mathbf{T}^{\dagger}(\mathbf{X})^{\top} \mathbf{b}^* - a^* - \mathbf{T}^*(\mathbf{X})^{\top} \mathbf{b}^*\}^2 \right] \\
 & = E_{\mathbf{X}} \left(\left[\{\mathbf{T}^{\dagger}(\mathbf{X}) - \mathbf{T}^*(\mathbf{X})\}^{\top} \mathbf{b}^* \right]^2 \right) \\
 & \leq \|\mathbf{b}^*\|_2^2 E_{\mathbf{X}} \left\{ \|\mathbf{T}^{\dagger}(\mathbf{X}) - \mathbf{T}^*(\mathbf{X})\|_2^2 \right\}, \text{ by the Cauchy-Schwarz inequality,} \\
 & \leq \|\mathbf{b}^*\|_1^2 \sum_{j=1}^p E_{X_j} \left[\left\{ T_j^{\dagger}(X_j) - T_j^*(X_j) \right\}^2 \right] \\
 & \leq M^2 \sum_{j=1}^p E_{X_j} \left[\left\{ \log f_{\text{TPSC}}(X_j | \boldsymbol{\Psi}_{j1}^{\dagger}) - \log f_{\text{TPSC}}(X_j | \boldsymbol{\Psi}_{j0}^{\dagger}) \right. \right. \\
 & \quad \left. \left. - \log \tilde{f}_j(X_j | Y=1) + \log \tilde{f}_j(X_j | Y=0) \right\}^2 \right] \\
 & = M^2 \sum_{j=1}^p E_{X_j} \left[\{e_{j1}(X_j) - e_{j0}(X_j)\}^2 \right] \\
 & = M^2 \sum_{j \in \mathcal{I}} E_{X_j} \left[\{e_{j1}(X_j) - e_{j0}(X_j)\}^2 \right], \tag{B.12}
 \end{aligned}$$

where the last equation is due to the remarks above Lemma 1, and the expectation in (B.12) is the expectation that Lemma 1 concerns. Using the quantity ρ_j defined in Lemma 1, we partition the index set $\mathcal{I}$ into $J_1 = \{j \in \mathcal{I} : \rho_j \leq 0.5\}$ and $J_2 = \{j \in \mathcal{I} : \rho_j > 0.5\}$. Then, by Lemma 1,

(B.12) is bounded from above by M^2 times

$$\sum_{j \in J_1} 2\delta_f(1 - \rho_j) + \sum_{j \in J_2} \delta_f = 2\delta_f|J_1|(1 - \bar{\rho}_1) + \delta_f|J_2|,$$

where $\bar{\rho}_1 = |J_1|^{-1} \sum_{j \in J_1} \rho_j$, with $\bar{\rho}_1 = 0$ when $J_1 = \emptyset$.

Returning to (B.11), we have

$$E_{\mathbf{X}} \left[\left\{ \eta_{\text{proj}}^\dagger(\mathbf{X}) - \eta^*(\mathbf{X}) \right\}^2 \right] \leq 2M^2\delta_f (2|J_1|(1 - \bar{\rho}_1) + |J_2|) + 2\delta_{lo},$$

which gives, according to (B.10),

$$\begin{aligned} \mathcal{L}(\pi_{\text{proj}}^\dagger) - \mathcal{L}(\pi^*) &\leq \frac{1}{8} \{ 2M^2\delta_f (2|J_1|(1 - \bar{\rho}_1) + |J_2|) + 2\delta_{lo} \} \\ &= \frac{M^2\delta_f}{4} (2|J_1|(1 - \bar{\rho}_1) + |J_2|) + \frac{1}{4}\delta_{lo}, \end{aligned}$$

which is also an upper bound for $\mathcal{L}(\pi^\dagger) - \mathcal{L}(\pi^*)$.

Finally, by (B.9),

$$\begin{aligned} \Delta R &\leq \sqrt{2 \left\{ \frac{M^2\delta_f}{4} (2|J_1|(1 - \bar{\rho}_1) + |J_2|) + \frac{1}{4}\delta_{lo} \right\}} \\ &= \sqrt{\frac{M^2\delta_f}{2} (2|J_1|(1 - \bar{\rho}_1) + |J_2|) + \frac{1}{2}\delta_{lo}}. \end{aligned} \tag{B.13}$$

This completes the proof. $\square$

The two terms in (B.13) have distinct sources. The first measures the increased risk from modeling each within-class feature distribution by a TPSC density, when the true distribution need not belong to that family. The second measures the increased risk from taking the log-odds function

to be linear in the oracle features, when the true log-odds function need not be linear in them.

Appendix C: The Proof of Theorem 3

To prepare for proving Theorem 3 in the main article, we establish Lemma 2 concerning a noise feature X_j , i.e., $j \in \mathcal{I}_N = \{j \in [p] : X_j \perp Y\}$. By construction, for $j \in \mathcal{I}_N$, the model parameters in the distribution of X_j given $Y = 0$ agree with those given $Y = 1$. That is, in the true density functions of X_j under the two classes, $f(X_j|Y = 0; \Psi_{j0}^*)$ and $f(X_j|Y = 1; \Psi_{j1}^*)$, we have $\Psi_{j0}^* = \Psi_{j1}^*$, for which we use Ψ_j^* to denote the common truth, and use $f(X_j; \Psi)$ to refer to the common density with parameters Ψ .

Lemma 2. *Under Condition (A1), for $j \in \mathcal{I}_N$, the transformed feature $\hat{T}_j(X_j) = \log f(X_j; \hat{\Psi}_{j1}) - \log f(X_j; \hat{\Psi}_{j0})$, can be decomposed into*

$$\hat{T}_j(X_j) = (\hat{\Psi}_{j1} - \hat{\Psi}_{j0})^\top \mathbf{S}_j(X_j; \Psi_j^*) + O_p(n^{-1}), \quad (\text{C.1})$$

where $\hat{\Psi}_{jk}$ is the maximum likelihood estimator of Ψ_{jk} based on data $\{(\mathbf{X}_i, Y_i) : i \in [n] \text{ and } Y_i = k\}$, for $k = 0, 1$, and $\mathbf{S}_j(X_j; \Psi) = (\partial/\partial\Psi) \log f(X_j; \Psi)$ is the score function.

Proof. A second-order Taylor expansion of $\log f(X_j; \Psi)$ around Ψ_j^* gives

$$\begin{aligned} \log f(X_j; \Psi) = & \log f(X_j; \Psi_j^*) + (\Psi - \Psi_j^*)^\top (\partial/\partial\Psi) \log f(X_j; \Psi)|_{\Psi=\Psi_j^*} + \\ & \frac{1}{2}(\Psi - \Psi_j^*)^\top (\partial^2/\partial\Psi\partial\Psi^\top) \log f(X_j; \Psi)|_{\Psi=\tilde{\Psi}_j} (\Psi - \Psi_j^*), \end{aligned} \quad (\text{C.2})$$

where $\tilde{\Psi}_j$ lies between Ψ and Ψ_j^* . Evaluating Ψ at $\hat{\Psi}_{j1}$ and $\hat{\Psi}_{j0}$ yields the corresponding Taylor expansions of $\log f(X_j; \hat{\Psi}_{j1})$ and $\log f(X_j; \hat{\Psi}_{j0})$, respectively, the difference between which gives the transformed feature

$$\begin{aligned} & \hat{T}_j(X_j) \\ = & \left(\hat{\Psi}_{j1} - \hat{\Psi}_{j0} \right)^\top \mathbf{S}_j(X_j; \Psi_j^*) + \frac{1}{2} \left\{ \left(\hat{\Psi}_{j1} - \Psi_j^* \right)^\top \mathbf{H}_j(X_j; \tilde{\Psi}_{j1}) \left(\hat{\Psi}_{j1} - \Psi_j^* \right) \right. \\ & \left. - \left(\hat{\Psi}_{j0} - \Psi_j^* \right)^\top \mathbf{H}_j(X_j; \tilde{\Psi}_{j0}) \left(\hat{\Psi}_{j0} - \Psi_j^* \right) \right\}, \end{aligned}$$

where $\mathbf{H}_j(X_j; \Psi) = (\partial^2/\partial\Psi\partial\Psi^\top) \log f(X_j; \Psi)$ is the Hessian matrix, $\tilde{\Psi}_{jk}$ lies between $\hat{\Psi}_{jk}$ and Ψ_j^* , for $k = 0, 1$. Denote by R_j the expression inside the curly brackets above. We zoom in on R_j next.

Under Condition (A1) that guarantees regularity conditions for maximum likelihood inference, the Hessian matrices, $\mathbf{H}_j(X_j; \tilde{\Psi}_{j1})$ and $\mathbf{H}_j(X_j; \tilde{\Psi}_{j0})$, have eigenvalues uniformly bounded in probability. Let M_H be an upper bound of these eigenvalues in absolute value in probability. By the Rayleigh

quotient theorem (Horn and Johnson, 2013, Section 4.2),

$$\begin{aligned}
 & |R_j| \\
 &= \left| \left(\hat{\Psi}_{j1} - \Psi_j^* \right)^\top \mathbf{H}_j(X_j; \tilde{\Psi}_{j1}) \left(\hat{\Psi}_{j1} - \Psi_j^* \right) - \left(\hat{\Psi}_{j0} - \Psi_j^* \right)^\top \mathbf{H}_j(X_j; \tilde{\Psi}_{j0}) \left(\hat{\Psi}_{j0} - \Psi_j^* \right) \right| \\
 &\leq \left| \left(\hat{\Psi}_{j1} - \Psi_j^* \right)^\top \mathbf{H}_j(X_j; \tilde{\Psi}_{j1}) \left(\hat{\Psi}_{j1} - \Psi_j^* \right) \right| + \left| \left(\hat{\Psi}_{j0} - \Psi_j^* \right)^\top \mathbf{H}_j(X_j; \tilde{\Psi}_{j0}) \left(\hat{\Psi}_{j0} - \Psi_j^* \right) \right| \\
 &\leq M_H \left(\left\| \hat{\Psi}_{j1} - \Psi_j^* \right\|_2^2 + \left\| \hat{\Psi}_{j0} - \Psi_j^* \right\|_2^2 \right). \tag{C.3}
 \end{aligned}$$

Under Condition (A1), $\hat{\Psi}_{jk}$ is $\sqrt{n_k}$ -consistent. That is, $\|\hat{\Psi}_{jk} - \Psi_j^*\|_2 = O_p(n_k^{-1/2}) = O_p(n^{-1/2})$, for $k = 0, 1$. Therefore, $R_j = O_p(n^{-1})$ according to (C.3). This completes the proof. $\square$

Now we are ready to prove Theorem 3 that concerns the gradient of the objective function that S-RoLLR classifier minimizes with the LASSO penalty set at zero, that is,

$$Q_n(\beta_0, \boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^n \left\{ Y_i(\beta_0 + \hat{\mathbf{T}}_i^\top \boldsymbol{\beta}) - \log(1 + \exp(\beta_0 + \hat{\mathbf{T}}_i^\top \boldsymbol{\beta})) \right\}, \tag{C.4}$$

where $\hat{\mathbf{T}}_i = (\hat{T}_1(X_{i1}), \dots, \hat{T}_p(X_{ip}))^\top$, for $i = 1, \dots, n$, and the gradient function of interest is $(\partial/\partial \boldsymbol{\beta})Q_n(\beta_0, \boldsymbol{\beta})$, equal to

$$G(\beta_0, \boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^n \left\{ Y_i - \mathcal{G}(\beta_0 + \hat{\mathbf{T}}_i^\top \boldsymbol{\beta}) \right\} \hat{\mathbf{T}}_i. \tag{C.5}$$

Theorem 3. *Under Conditions (A1)–(A3), the gradient function in (C.5),*

$G(\beta_0, \boldsymbol{\beta}) = \{G_1(\beta_0, \boldsymbol{\beta}), \dots, G_p(\beta_0, \boldsymbol{\beta})\}^\top$, *evaluated at the canonical sparse*

representative $\boldsymbol{\theta}^\circ = (\beta_0^\circ, \boldsymbol{\beta}^{\circ\top})^\top \in \boldsymbol{\Theta}_0$, satisfies $\max_{j \in \mathcal{I}} |G_j(\boldsymbol{\theta}^\circ)| = O_p(n^{-1/2})$

and $\max_{j \in \mathcal{I}_N} |G_j(\boldsymbol{\theta}^\circ)| = O_p(n^{-1})$.

Proof. By (C.5), the j -th entry of $G_j(\beta_0^\circ, \boldsymbol{\beta}^\circ)$ is, for $j \in [p]$,

$$\begin{aligned} & G_j(\beta_0^\circ, \boldsymbol{\beta}^\circ) \\ &= -\frac{1}{n} \sum_{i=1}^n \hat{T}_j(X_{ij}) \left\{ Y_i - \pi(\hat{\mathbf{T}}_i; \beta_0^\circ, \boldsymbol{\beta}^\circ) \right\} \\ &= -\frac{1}{n} \sum_{i=1}^n \hat{T}_j(X_{ij}) \left\{ Y_i - \pi(\mathbf{T}_i^*; \beta_0^\circ, \boldsymbol{\beta}^\circ) + \pi(\mathbf{T}_i^*; \beta_0^\circ, \boldsymbol{\beta}^\circ) - \pi(\hat{\mathbf{T}}_i; \beta_0^\circ, \boldsymbol{\beta}^\circ) \right\} \\ &= -\frac{1}{n} \sum_{i=1}^n \hat{T}_j(X_{ij}) \{ Y_i - \pi(\mathbf{T}_i^*; \beta_0^\circ, \boldsymbol{\beta}^\circ) \} \end{aligned} \quad (\text{C.6})$$

$$+ \frac{1}{n} \sum_{i=1}^n \hat{T}_j(X_{ij}) \left\{ \pi(\hat{\mathbf{T}}_i; \beta_0^\circ, \boldsymbol{\beta}^\circ) - \pi(\mathbf{T}_i^*; \beta_0^\circ, \boldsymbol{\beta}^\circ) \right\}. \quad (\text{C.7})$$

For the average in (C.6), observing that the summands in (C.6) are not independent due to their dependence on $\hat{\boldsymbol{\Psi}}_{j0}$ and $\hat{\boldsymbol{\Psi}}_{j1}$ via $\hat{T}_j(X_{ij})$, we re-express (C.6) as $A_n + B_n$, where

$$\begin{aligned} A_n &= -\frac{1}{n} \sum_{i=1}^n T_j^*(X_{ij}) \{ Y_i - \pi(\mathbf{T}_i^*; \beta_0^\circ, \boldsymbol{\beta}^\circ) \}, \\ B_n &= -\frac{1}{n} \sum_{i=1}^n \left\{ \hat{T}_j(X_{ij}) - T_j^*(X_{ij}) \right\} \{ Y_i - \pi(\mathbf{T}_i^*; \beta_0^\circ, \boldsymbol{\beta}^\circ) \}. \end{aligned}$$

The summands of A_n are independent and identically distributed (i.i.d.) with zero mean and finite variance (bounded by $E\{T_j^{*2}(X_{1j})\}/4 < \infty$), thus $A_n = O_p(n^{-1/2})$ by the central limit theorem.

To study the order of B_n , define $\Delta\boldsymbol{\Psi}_{jk} = \hat{\boldsymbol{\Psi}}_{jk} - \boldsymbol{\Psi}_{jk}^*$, for $k = 0, 1$.

Under Condition (A1), $\|\Delta\boldsymbol{\Psi}_{jk}\|_2 = O_p(n^{-1/2})$. Using the Taylor expansion

in (C.2), one has

$$\hat{T}_j(X_{ij}) - T_j^*(X_{ij}) = \Delta \Psi_{j1}^\top \mathbf{S}_j(X_{ij}; \Psi_{j1}^*) - \Delta \Psi_{j0}^\top \mathbf{S}_j(X_{ij}; \Psi_{j0}^*) + r_{ij1} - r_{ij0}, \quad (\text{C.8})$$

where $r_{ijk} = (1/2)\Delta \Psi_{jk}^\top \mathbf{H}_j(X_{ij}; \tilde{\Psi}_{ijk})\Delta \Psi_{jk}$, and $\tilde{\Psi}_{ijk}$ lies between $\hat{\Psi}_{jk}$ and Ψ_{jk}^* . Under Condition (A1), together with the weak law of large numbers, one has $n^{-1} \sum_{i=1}^n \|\mathbf{S}_j(X_{ij}; \Psi_{jk}^*)\|_2^2 = O_p(1)$ and $n^{-1} \sum_{i=1}^n \|\mathbf{H}_j(X_{ij}; \tilde{\Psi}_{ijk})\|_{\text{op}}^2 = O_p(1)$, for $k = 0, 1$. Consequently,

$$\frac{1}{n} \sum_{i=1}^n \{\Delta \Psi_{jk}^\top \mathbf{S}_j(X_{ij}; \Psi_{jk}^*)\}^2 \leq \|\Delta \Psi_{jk}\|_2^2 \frac{1}{n} \sum_{i=1}^n \|\mathbf{S}_j(X_{ij}; \Psi_{jk}^*)\|_2^2 = O_p(n^{-1}),$$

whereas

$$\frac{1}{n} \sum_{i=1}^n r_{ijk}^2 \leq \frac{1}{4} \|\Delta \Psi_{jk}\|_2^4 \frac{1}{n} \sum_{i=1}^n \|\mathbf{H}_j(X_{ij}; \tilde{\Psi}_{ijk})\|_{\text{op}}^2 = O_p(n^{-2}),$$

which imply, by (C.8), $n^{-1} \sum_{i=1}^n \{\hat{T}_j(X_{ij}) - T_j^*(X_{ij})\}^2 = O_p(n^{-1})$ because $(a_1 + \dots + a_4)^2 \leq 4(a_1^2 + \dots + a_4^2)$. Finally, with $|Y_i - \pi(\mathbf{T}_i^*; \beta_0^\circ, \boldsymbol{\beta}^\circ)| \leq 1$ and by the Cauchy–Schwarz inequality,

$$\begin{aligned} & |B_n| \\ & \leq \left[\frac{1}{n} \sum_{i=1}^n \{\hat{T}_j(X_{ij}) - T_j^*(X_{ij})\}^2 \right]^{1/2} \times \left[\frac{1}{n} \sum_{i=1}^n \{Y_i - \pi(\mathbf{T}_i^*; \beta_0^\circ, \boldsymbol{\beta}^\circ)\}^2 \right]^{1/2} \\ & = O_p(n^{-1/2}). \end{aligned}$$

Combining this result with $A_n = O_p(n^{-1/2})$ shows that (C.6) is $O_p(n^{-1/2})$.

For the average in (C.7), we elaborate the difference as follows, where we define the linear predictors $\eta_i^\circ = \beta_0^\circ + \mathbf{T}_i^{*\top} \boldsymbol{\beta}^\circ$ and $\hat{\eta}_i^\circ = \beta_0^\circ + \hat{\mathbf{T}}_i^\top \boldsymbol{\beta}^\circ$,

$$\begin{aligned} \pi(\hat{\mathbf{T}}_i; \beta_0^\circ, \boldsymbol{\beta}^\circ) - \pi(\mathbf{T}_i^*; \beta_0^\circ, \boldsymbol{\beta}^\circ) &= \mathcal{G}(\hat{\eta}_i^\circ) - \mathcal{G}(\eta_i^\circ) \\ &= \mathcal{G}'(\tilde{\eta}_i^\circ)(\hat{\eta}_i^\circ - \eta_i^\circ), \text{ by the mean value theorem,} \\ &= \mathcal{G}'(\tilde{\eta}_i^\circ) \sum_{\ell=1}^p \beta_\ell^\circ \left\{ \hat{T}_\ell(X_{i\ell}) - T_\ell^*(X_{i\ell}) \right\} \\ &= \mathcal{G}'(\tilde{\eta}_i^\circ) \sum_{\ell \in \mathcal{I}} \left\{ \hat{T}_\ell(X_{i\ell}) - T_\ell^*(X_{i\ell}) \right\}, \end{aligned}$$

where $\tilde{\eta}_i^\circ$ lies between $\hat{\eta}_i^\circ$ and η_i° , and $\mathcal{G}'(t) = (d/dt)\mathcal{G}(t) = \mathcal{G}(t)\{1 - \mathcal{G}(t)\}$.

It follows that (C.7) is equal to

$$\frac{1}{n} \sum_{i=1}^n \hat{T}_j(X_{ij}) \mathcal{G}'(\tilde{\eta}_i^\circ) \sum_{\ell \in \mathcal{I}} \left\{ \hat{T}_\ell(X_{i\ell}) - T_\ell^*(X_{i\ell}) \right\}, \quad (\text{C.9})$$

where $\hat{T}_j(X_{ij}) = O_p(1)$ because $\hat{T}_j(X_{ij}) \xrightarrow{p} T_j^*(X_{ij})$, $\mathcal{G}'(\tilde{\eta}_i^\circ) = O_p(1)$ because $\mathcal{G}'(t) \in (0, 0.25]$ for all $t \in \mathbb{R}$, and $\hat{T}_j(X_{ij}) - T_j^*(X_{ij}) = O_p(n^{-1/2})$. Given a fixed p , (C.7) is thus $O_p(n^{-1/2})$.

With (C.6) and (C.7) both of order $O_p(n^{-1/2})$, we have $G_j(\beta_0^\circ, \boldsymbol{\beta}^\circ) = O_p(n^{-1/2})$, for $j \in [p]$. This proves $G_j(\beta_0^\circ, \boldsymbol{\beta}^\circ) = O_p(n^{-1/2})$, for $j \in \mathcal{I} \subseteq [p]$.

For $j \in \mathcal{I}_N$, the rate of $O_p(n^{-1/2})$ for both (C.6) and (C.7) can be further refined because, now $\boldsymbol{\Psi}_{j1}^* = \boldsymbol{\Psi}_{j0}^* = \boldsymbol{\Psi}_j^*$, $T_j^*(X_{ij}) = \log f(X_{ij}; \boldsymbol{\Psi}_{j1}^*) - \log f(X_{ij}; \boldsymbol{\Psi}_{j0}^*) = 0$. In other words, for a non-informative feature X_j , its transformed counterpart $\hat{T}_j(X_{ij})$ converges in probability to zero, unlike for an informative feature $X_{j'}$ (i.e., $j' \in \mathcal{I}$) that has its transformed counter-

part $\hat{T}_{j'}(X_{ij'})$ converging to a non-degenerate random variable $T_{j'}^*(X_{ij'})$ in probability.

In particular, using (C.1) from Lemma 2 established for non-informative features, (C.6) is dominated by

$$\begin{aligned} & -\frac{1}{n} \sum_{i=1}^n \left(\hat{\Psi}_{j1} - \hat{\Psi}_{j0} \right)^\top \mathbf{S}_j(X_{ij}; \Psi_j^*) \{Y_i - \pi(\mathbf{T}_i^*; \beta_0^\circ, \beta^\circ)\} \\ & = \left(\hat{\Psi}_{j0} - \hat{\Psi}_{j1} \right)^\top \times \frac{1}{n} \sum_{i=1}^n \mathbf{S}_j(X_{ij}; \Psi_j^*) \{Y_i - \pi(\mathbf{T}_i^*; \beta_0^\circ, \beta^\circ)\}, \end{aligned}$$

where the first term has $\hat{\Psi}_{j0} - \hat{\Psi}_{j1} = O_p(n^{-1/2})$, and the second term is an average of i.i.d. random quantities, each of which has a mean of zero and a finite variance, and thus, by the central limit theorem, is $O_p(n^{-1/2})$. This indicates that (C.6) is $O_p(n^{-1})$.

Using (C.1) in (C.9), we have that (C.7) is dominated by

$$\left(\hat{\Psi}_{j1} - \hat{\Psi}_{j0} \right)^\top \times \frac{1}{n} \sum_{i=1}^n \mathbf{S}_j(X_{ij}; \Psi_j^*) \mathcal{G}'(\tilde{\eta}_i^\circ) \sum_{\ell \in \mathcal{I}} \left\{ \hat{T}_\ell(X_{i\ell}) - T_\ell^*(X_{i\ell}) \right\},$$

where, similar to the arguments for (C.9), the average above is $O_p(n^{-1/2})$, and thus, with the difference of order $O_p(n^{-1/2})$ multiplying in front, the above expression is $O_p(n^{-1})$. With both (C.6) and (C.7) of order $O_p(n^{-1})$ here, we have $G_j(\beta_0^\circ, \beta^\circ) = O_p(n^{-1})$ for $j \in \mathcal{I}_N$. Since p is fixed, the pointwise orders above hold uniformly over the finitely many indices within each block, which gives the two maxima in the statement. This completes the proof. $\square$

Appendix D: The Proof of Theorem 4

Condition (A8) constrains only the signal block $\mathbf{I}_{\mathcal{S}\mathcal{S}}^*$, not the full matrix $\mathbf{I}^*$. For a noise feature $j \in \mathcal{I}_N$ one has $\Psi_{j1}^* = \Psi_{j0}^*$, so the oracle transform satisfies $T_j^*(X_j) \equiv 0$. The j -th row and column of $\mathbf{I}^*$ therefore vanish, and $\mathbf{I}^*$ is singular whenever $\mathcal{I}_N \neq \emptyset$. Requiring the full $\mathbf{I}^*$ to be positive definite would be unsatisfiable in the regime that Theorem 4 concerns. The signal block, by contrast, is positive definite whenever the signal transforms $\{T_j^*\}_{j \in \mathcal{I}}$ are non-degenerate and, together with the intercept, not affinely dependent. It is all that the subsequent analysis requires, because the noise features are controlled through the score separation of Theorem 3 and the L_1 penalty rather than through curvature; see the analysis below.

To signify the dependence of the objective function in (C.4) on the transformed features $\{\hat{\mathbf{T}}_i\}_{i=1}^n$, we write $Q_n(\boldsymbol{\theta}; \hat{\mathbf{T}})$ for the objective in (C.4), and $Q_n(\boldsymbol{\theta}; \mathbf{T}^*)$ for its oracle counterpart obtained by replacing $\{\hat{\mathbf{T}}_i\}_{i=1}^n$ with $\{\mathbf{T}_i^*\}_{i=1}^n$. Throughout this appendix, the population target is the canonical sparse representative $\boldsymbol{\theta}^\circ = (\beta_0^\circ, \boldsymbol{\beta}^{\circ\top})^\top$, where $\boldsymbol{\beta}_{\mathcal{I}_N}^\circ = \mathbf{0}$.

Besides Theorem 3, the proof of Theorem 4 uses Lemma 3, Corollary 1, and Lemma 4, established below under Condition (A8). Define the oracle

Fisher information matrix by

$$\mathbf{I}^* = E \left\{ \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} Q_n(\boldsymbol{\theta}; \mathbf{T}^*) \middle|_{\boldsymbol{\theta} = \boldsymbol{\theta}^\circ} \right\} = E \left[\mathcal{G}(\tilde{\mathbf{T}}_i^{*\top} \boldsymbol{\theta}^\circ) \left\{ 1 - \mathcal{G}(\tilde{\mathbf{T}}_i^{*\top} \boldsymbol{\theta}^\circ) \right\} \tilde{\mathbf{T}}_i^* \tilde{\mathbf{T}}_i^{*\top} \right],$$

where $\tilde{\mathbf{T}}_i^* = (1, \mathbf{T}_i^{*\top})^\top$.

This definition does not depend on the choice of a representative from $\boldsymbol{\Theta}_0$, because every element of $\boldsymbol{\Theta}_0$ induces the same oracle linear predictor almost surely. Moreover, Condition (A8) identifies the signal block of the coefficient vector. To see this, let $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. By the characterization of $\boldsymbol{\Theta}_0$ in Appendix A, $\tilde{\mathbf{T}}^{*\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) = 0$ almost surely. Since $T_j^*(\cdot) \equiv 0$ for every $j \in \mathcal{I}_N$, the noise features contribute nothing to this inner product, so that $\tilde{\mathbf{T}}_S^{*\top}(\boldsymbol{\theta}_S - \boldsymbol{\theta}_S^\circ) = 0$ almost surely. Squaring, multiplying by the bounded weight $\mathcal{G}(\eta^*(\mathbf{X}))\{1 - \mathcal{G}(\eta^*(\mathbf{X}))\}$ and taking expectations gives

$$0 = E \left[\mathcal{G}(\eta^*(\mathbf{X}))\{1 - \mathcal{G}(\eta^*(\mathbf{X}))\} \left\{ \tilde{\mathbf{T}}_S^{*\top}(\boldsymbol{\theta}_S - \boldsymbol{\theta}_S^\circ) \right\}^2 \right] = (\boldsymbol{\theta}_S - \boldsymbol{\theta}_S^\circ)^\top \mathbf{I}_{SS}^* (\boldsymbol{\theta}_S - \boldsymbol{\theta}_S^\circ).$$

Condition (A8) therefore implies $\boldsymbol{\theta}_S = \boldsymbol{\theta}_S^\circ$. Thus, under Condition (A8), the population non-identifiability is confined to the noise coefficients.

Since $T_j^*(\cdot) \equiv 0$ for every $j \in \mathcal{I}_N$, the rows and columns of $\mathbf{I}^*$ indexed by $\mathcal{I}_N$ are zero. Condition (A8) consequently places its lower eigenvalue requirement on the signal block $\mathbf{I}_{SS}^*$, where $\mathcal{S} = \{0\} \cup \mathcal{I}$: $\lambda_{\min}(\mathbf{I}_{SS}^*) \geq 2\kappa$. For $\boldsymbol{\nu} \in \mathbb{R}^{p+1}$, let $\boldsymbol{\nu}_S$ and $\boldsymbol{\nu}_{\mathcal{I}_N}$ denote the corresponding subvectors. For

$\mathbf{B} \in \mathbb{R}^{(p+1) \times (p+1)}$, $\mathbf{B}_{\mathcal{S}\mathcal{S}}$ is the $|\mathcal{S}| \times |\mathcal{S}|$ principal submatrix indexed by $\mathcal{S}$, and $\mathbf{B}_{\mathcal{S}\mathcal{I}_N}$ is the $|\mathcal{S}| \times |\mathcal{I}_N|$ submatrix with row indices $\mathcal{S}$ and column indices $\mathcal{I}_N$. This shows that every $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ satisfies $\boldsymbol{\theta}_{\mathcal{S}} = \boldsymbol{\theta}_{\mathcal{S}}^\circ$. Conversely, any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\boldsymbol{\theta}_{\mathcal{S}} = \boldsymbol{\theta}_{\mathcal{S}}^\circ$ has $\eta_{\boldsymbol{\theta}}(\mathbf{X}) = \eta^*(\mathbf{X})$ almost surely, again because $T_j^*(\cdot) \equiv 0$ on $\mathcal{I}_N$, and hence lies in $\boldsymbol{\Theta}_0$. Under Condition (A8) the set of population minimizers therefore takes the explicit form

$$\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \boldsymbol{\theta}_{\mathcal{S}} = \boldsymbol{\theta}_{\mathcal{S}}^\circ\}, \quad (\text{D.1})$$

so that, by (D.1), $\|\boldsymbol{\beta}\|_1 = \|\boldsymbol{\beta}_{\mathcal{I}}^\circ\|_1 + \|\boldsymbol{\beta}_{\mathcal{I}_N}\|_1$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Therefore, the canonical sparse representative is the unique minimizer of $\|\boldsymbol{\beta}\|_1$ over $\boldsymbol{\Theta}_0$, $\boldsymbol{\theta}^\circ = \arg \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \|\boldsymbol{\beta}\|_1$. This is the population-level reason why the L_1 penalty singles out $\boldsymbol{\theta}^\circ$ among all population minimizers, and it is what Lemma 3 establishes at the sample level.

Lemma 3 shows that the L_1 penalty selects the canonical sparse representative from the set of population minimizers.

Lemma 3. *Under Conditions (A1)–(A3) and (A8), suppose that $\lambda_n \asymp n^{-1/2}$. Then any minimizer $\hat{\boldsymbol{\theta}}$ of (D.9) satisfies*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\|_1 = O_p(n^{-1/2}).$$

Proof. The set-consistency result in Appendix A does not identify the noise coefficients. The purpose of this lemma is to show that the L_1 penalty local-

izes the penalized estimator around the canonical sparse representative $\boldsymbol{\theta}^\circ$.

Firstly, we show that the signal block of $\hat{\mathbf{I}}(\boldsymbol{\theta}) = (\partial^2/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top)Q_n(\boldsymbol{\theta};\hat{\mathbf{T}})$, as an empirical version of $\mathbf{I}^*$, satisfies a local restricted strong convexity condition with high probability. Equivalently, we will show that, for any $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}^\circ$, $\mathcal{B}_n(W) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|_2 \leq W/\sqrt{n}\}$, the smallest eigenvalue of the signal block $\hat{\mathbf{I}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\theta})$, $\lambda_{\min}(\hat{\mathbf{I}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\theta}))$, is no less than κ with probability tending to one as $n \rightarrow \infty$. Since $\boldsymbol{\theta}^\circ \in \text{int}(\boldsymbol{\Theta})$, for every fixed $W < \infty$ one has $\mathcal{B}_n(W) \subset \text{int}(\boldsymbol{\Theta})$ for all sufficiently large n .

With both $\hat{\mathbf{I}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\theta})$ and $\mathbf{I}_{\mathcal{S}\mathcal{S}}^*$ symmetric, by Weyl's eigenvalue perturbation inequality, we have $|\lambda_{\min}(\hat{\mathbf{I}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\theta})) - \lambda_{\min}(\mathbf{I}_{\mathcal{S}\mathcal{S}}^*)| \leq \|\hat{\mathbf{I}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\theta}) - \mathbf{I}_{\mathcal{S}\mathcal{S}}^*\|_{\text{op}}$, where $\|\cdot\|_{\text{op}}$ is the operator norm, and thus $\lambda_{\min}(\hat{\mathbf{I}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\theta})) \geq \lambda_{\min}(\mathbf{I}_{\mathcal{S}\mathcal{S}}^*) - \|\hat{\mathbf{I}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\theta}) - \mathbf{I}_{\mathcal{S}\mathcal{S}}^*\|_{\text{op}}$. Under Condition (A8) stating that $\lambda_{\min}(\mathbf{I}_{\mathcal{S}\mathcal{S}}^*) \geq 2\kappa$, to prove $\lambda_{\min}(\hat{\mathbf{I}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\theta})) \geq \kappa$ with probability tending to one as $n \rightarrow \infty$ for $\boldsymbol{\theta} \in \mathcal{B}_n(W)$, it suffices to establish $\|\hat{\mathbf{I}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\theta}) - \mathbf{I}_{\mathcal{S}\mathcal{S}}^*\|_{\text{op}} \leq \kappa$ for $\boldsymbol{\theta} \in \mathcal{B}_n(W)$ with probability tending to one. Since the operator norm of a principal submatrix never exceeds that of the full matrix, $\|\hat{\mathbf{I}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\theta}) - \mathbf{I}_{\mathcal{S}\mathcal{S}}^*\|_{\text{op}} \leq \|\hat{\mathbf{I}}(\boldsymbol{\theta}) - \mathbf{I}^*\|_{\text{op}}$, so it suffices to bound the latter from above. With this goal in mind, we inspect three operator norms arising from applying the triangle inequality

on $\|\hat{\mathbf{I}}(\boldsymbol{\theta}) - \mathbf{I}^*\|_{\text{op}}$ as follows,

$$\|\hat{\mathbf{I}}(\boldsymbol{\theta}) - \mathbf{I}^*\|_{\text{op}} \leq \underbrace{\|\hat{\mathbf{I}}(\boldsymbol{\theta}) - \hat{\mathbf{I}}(\boldsymbol{\theta}^\circ)\|_{\text{op}}}_{(a)} + \underbrace{\|\hat{\mathbf{I}}(\boldsymbol{\theta}^\circ) - \tilde{\mathbf{I}}(\boldsymbol{\theta}^\circ)\|_{\text{op}}}_{(b)} + \underbrace{\|\tilde{\mathbf{I}}(\boldsymbol{\theta}^\circ) - \mathbf{I}^*\|_{\text{op}}}_{(c)}, \quad (\text{D.2})$$

where $\tilde{\mathbf{I}}(\boldsymbol{\theta}^\circ) = (\partial^2 / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top) Q_n(\boldsymbol{\theta}; \mathbf{T}^*)|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\circ}$, and thus $\mathbf{I}^* = E\{\tilde{\mathbf{I}}(\boldsymbol{\theta}^\circ)\}$.

We first bound the norm (a) in (D.2). Differentiating (C.4) with respect to $\boldsymbol{\theta}$ twice gives, with $\tilde{\mathbf{T}}_i = (1, \hat{\mathbf{T}}_i^\top)^\top$,

$$\hat{\mathbf{I}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathcal{G}(\tilde{\mathbf{T}}_i^\top \boldsymbol{\theta}) \left\{ 1 - \mathcal{G}(\tilde{\mathbf{T}}_i^\top \boldsymbol{\theta}) \right\} \tilde{\mathbf{T}}_i \tilde{\mathbf{T}}_i^\top, \quad (\text{D.3})$$

which is Lipschitz continuous in $\boldsymbol{\theta}$ with random Lipschitz constant $L_a = n^{-1} \sum_{i=1}^n \|\tilde{\mathbf{T}}_i\|_2^3$. In what follows, we show that $L_a = O_p(1)$, which requires a closer look at the logarithm of the TPSC density as the building block of $\{\hat{\mathbf{T}}_i\}_{i=1}^n$, and the derivative of the log-density with respect to $\boldsymbol{\Psi} = (\tau, w, \sigma, \delta)^\top$. This log-density is

$$\begin{aligned} & \log f_{\text{TPSC}}(x | \boldsymbol{\Psi}) \\ &= \log 2 - \log \sigma + \mathbb{1}(x < \tau) \left\{ \log w - \frac{1}{2} \log \frac{w}{1-w} + \log g(z_1; \delta) \right\} \\ & \quad + \mathbb{1}(x \geq \tau) \left\{ \log(1-w) + \frac{1}{2} \log \frac{w}{1-w} + \log g(z_2; \delta) \right\}, \end{aligned}$$

where $z_1 = (x - \tau)/\sigma_1$, $z_2 = (x - \tau)/\sigma_2$, and $g(\cdot; \delta)$ is the density of Student's t with δ degrees of freedom, t_δ . By construction, the tail decay rate of $f_{\text{TPSC}}(x|\boldsymbol{\Psi})$ is of the same order as that of $g(\cdot; \delta)$, hence $f_{\text{TPSC}}(x|\boldsymbol{\Psi}) =$

$O(|x|^{-(\delta+1)})$, and thus

$$|\log f_{\text{TPSC}}(x|\mathbf{\Psi})| = O(\log |x|). \quad (\text{D.4})$$

As for $(\partial/\partial\mathbf{\Psi}) \log f_{\text{TPSC}}(x|\mathbf{\Psi})$, which can be derived analytically (omitted here), only one entry, $(\partial/\partial\delta) \log f_{\text{TPSC}}(x|\mathbf{\Psi})$, diverges as x diverges, which is dominated by $0.5\{\mathbb{1}(x < \tau) \log(1+z_1^2/\delta) + \mathbb{1}(x \geq \tau) \log(1+z_2^2/\delta)\}$. Therefore, for any $x \in \mathbb{R}$, $|(\partial/\partial\delta) \log f_{\text{TPSC}}(x|\mathbf{\Psi})| \leq C_1 + C_2 \log(1+x^2)$ for some positive constants C_1 and C_2 that depend on $\mathbf{\Psi}$, and thus

$$\left\| \frac{\partial}{\partial\mathbf{\Psi}} \log f_{\text{TPSC}}(x|\mathbf{\Psi}) \right\|_2 \leq C_3 + C_4 \log(1+x^2), \quad (\text{D.5})$$

for some positive constants C_3 and C_4 that depend on $\mathbf{\Psi}$, for any $x \in \mathbb{R}$.

Zooming in on the summand in $L_a = n^{-1} \sum_{i=1}^n \|\tilde{\mathbf{T}}_i\|^3$, we apply Hölder's inequality to have

$$\begin{aligned} \|\tilde{\mathbf{T}}_i\|^3 &\leq \sqrt{p+1} \left\{ 1 + \sum_{j=1}^p \left| \hat{T}_j(X_{ij}) \right|^3 \right\}, \\ &= \sqrt{p+1} \left\{ 1 + \sum_{j=1}^p \left| \hat{T}_j(X_{ij}) - T_j^*(X_{ij}) + T_j^*(X_{ij}) \right|^3 \right\} \\ &\leq \sqrt{p+1} \left[1 + \sum_{j=1}^p \left\{ \left| \hat{T}_j(X_{ij}) - T_j^*(X_{ij}) \right| + |T_j^*(X_{ij})| \right\}^3 \right] \\ &\leq \sqrt{p+1} \left[1 + 4 \sum_{j=1}^p \left\{ \left| \hat{T}_j(X_{ij}) - T_j^*(X_{ij}) \right|^3 + |T_j^*(X_{ij})|^3 \right\} \right]. \end{aligned}$$

Hence,

$$L_a \leq 4\sqrt{p+1} \left\{ \frac{1}{4} + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p \left| \hat{T}_j(X_{ij}) - T_j^*(X_{ij}) \right|^3 + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p |T_j^*(X_{ij})|^3 \right\}. \quad (\text{D.6})$$

By (D.4), $|T_j^*(X_{1j})|^3 = O_p((\log |X_{1j}|)^3)$. Since the logarithm grows slower than any polynomial decay rate of the TPSC density's tail, $E\{(\log |X_{1j}|)^3\} < \infty$ under Condition (A1), and hence $E\{\sum_{j=1}^p |T_j^*(X_{1j})|^3\} < \infty$. With $\{\sum_{j=1}^p |T_j^*(X_{ij})|^3\}_{i=1}^n$ i.i.d., $n^{-1} \sum_{i=1}^n \sum_{j=1}^p |T_j^*(X_{ij})|^3$ converges in probability to its expectation, by Khintchine's weak law of large numbers, implying that this average is $O_p(1)$. We next show that the first average in (D.6) is $o_p(1)$. To achieve this, we study $|\hat{T}_j(X_{ij}) - T_j^*(X_{ij})|$, for $j \in [p]$.

By the definition of the transformed features and oracle features,

$$\begin{aligned} & \left| \hat{T}_j(X_{ij}) - T_j^*(X_{ij}) \right| \\ &= \left| \log f_{\text{TPSC}}(X_{ij} | \hat{\Psi}_{j1}) - \log f_{\text{TPSC}}(X_{ij} | \hat{\Psi}_{j0}) - \log f_{\text{TPSC}}(X_{ij} | \Psi_{j1}^*) + \log f_{\text{TPSC}}(X_{ij} | \Psi_{j0}^*) \right| \\ &\leq \sum_{k=0}^1 \left| \log f_{\text{TPSC}}(X_{ij} | \hat{\Psi}_{jk}) - \log f_{\text{TPSC}}(X_{ij} | \Psi_{jk}^*) \right| \\ &\leq \sum_{k=0}^1 \left\| \frac{\partial}{\partial \Psi} \log f_{\text{TPSC}}(X_{ij} | \Psi) \right\|_{\Psi = \tilde{\Psi}_k} \left\| \hat{\Psi}_{jk} - \Psi_{jk}^* \right\|_2, \end{aligned} \quad (\text{D.7})$$

where $\tilde{\Psi}_k$ lies between $\hat{\Psi}_{jk}$ and Ψ_{jk}^* by the mean value theorem, for $k = 0, 1$. With the regularity conditions guaranteed by Conditions (A1)–(A3), $\|\hat{\Psi}_{jk} - \Psi_{jk}^*\|_2 = O_p(n^{-1/2})$. Let $\mathcal{N}(\Psi_{jk}^*)$ be a compact neighborhood of Ψ_{jk}^* which $\hat{\Psi}_{jk}$ falls in with probability tending to one as $n \rightarrow \infty$, then $\tilde{\Psi}_{jk} \in$

$\mathcal{N}(\Psi_{jk}^*)$ with probability tending to one. By (D.5), for $j \in [p]$ and $k = 0, 1$,

$$\max_{i \in [n]} \left\| \frac{\partial}{\partial \Psi} \log f_{\text{TPSC}}(X_{ij} \mid \Psi) \right\|_2 \leq C_3 + C_4 \max_{i \in [n]} \log(1 + X_{ij}^2).$$

As for $\max_{i \in [n]} \log(1 + X_{ij}^2)$, because the tail properties of the TPSC density inherits those of t_δ , $P(|X_{ij}| > s) = O(s^{-\delta})$ as $s \rightarrow +\infty$, and thus $\max_{i \in [n]} X_{ij} = O_p(n^{1/\delta})$ (Resnick, 2007), suggesting $\max_{i \in [n]} \log(1 + X_{ij}^2) = O_p(\log n)$. It follows that $\sup_{\Psi \in \mathcal{N}(\Psi_{jk}^*)} \max_{i \in [n]} \|(\partial/\partial \Psi) \log f_{\text{TPSC}}(X_{ij} \mid \Psi)\|_2 = O_p(\log n)$. Using this in (D.7), and taking the maximum over the finitely many indices $j \in [p]$, we conclude

$$\max_{i \in [n]} \max_{j \in [p]} \left| \hat{T}_j(X_{ij}) - T_j^*(X_{ij}) \right| = O_p(n^{-1/2} \log n) = o_p(1). \quad (\text{D.8})$$

The derivation of (D.8) relies only on Condition (A1) and on the tail behaviour of the TPSC density; it does not draw on any result established in Appendices A–C, and may therefore be cited from those appendices without circularity. In particular, (D.8) shows that the first average in (D.6) is $o_p(1)$, and hence $L_a = O_p(1)$. Returning to the norm (a) in (D.2), we now have, for $\boldsymbol{\theta} \in \mathcal{B}_n(W)$,

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_n(W)} \left\| \hat{\mathbf{I}}(\boldsymbol{\theta}) - \hat{\mathbf{I}}(\boldsymbol{\theta}^\circ) \right\|_{\text{op}} \leq L_a \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|_2 \leq L_a \cdot W n^{-1/2} = o_p(1).$$

We turn to the norm (b) in (D.2), which relates to the sensitivity of the

Hessian to feature perturbations. In particular, this norm is equal to

$$\begin{aligned}
 & \left\| \hat{\mathbf{I}}(\boldsymbol{\theta}^\circ) - \tilde{\mathbf{I}}(\boldsymbol{\theta}^\circ) \right\|_{\text{op}} \\
 &= \left\| \frac{1}{n} \sum_{i=1}^n \mathcal{G}(\tilde{\mathbf{T}}_i^\top \boldsymbol{\theta}^\circ) \left\{ 1 - \mathcal{G}(\tilde{\mathbf{T}}_i^\top \boldsymbol{\theta}^\circ) \right\} \tilde{\mathbf{T}}_i \tilde{\mathbf{T}}_i^\top - \right. \\
 & \quad \left. \frac{1}{n} \sum_{i=1}^n \mathcal{G}(\tilde{\mathbf{T}}_i^{*\top} \boldsymbol{\theta}^\circ) \left\{ 1 - \mathcal{G}(\tilde{\mathbf{T}}_i^{*\top} \boldsymbol{\theta}^\circ) \right\} \tilde{\mathbf{T}}_i^* \tilde{\mathbf{T}}_i^{*\top} \right\|_{\text{op}}, \text{ by (D.3),} \\
 &= \left\| \frac{1}{n} \sum_{i=1}^n \left\{ H(\hat{\mathbf{T}}_i; \boldsymbol{\theta}^\circ) - H(\mathbf{T}_i^*; \boldsymbol{\theta}^\circ) \right\} \right\|_{\text{op}} \\
 &\leq \frac{1}{n} \sum_{i=1}^n \left\| H(\hat{\mathbf{T}}_i; \boldsymbol{\theta}^\circ) - H(\mathbf{T}_i^*; \boldsymbol{\theta}^\circ) \right\|_{\text{op}}, \text{ by the triangle inequality,}
 \end{aligned}$$

where $\tilde{\mathbf{T}}_i^* = (1, \mathbf{T}_i^{*\top})^\top$ in contrast to $\tilde{\mathbf{T}}_i = (1, \hat{\mathbf{T}}_i^\top)^\top$, for $i = 1, \dots, n$, and $H(\mathbf{t}; \boldsymbol{\theta}) = \mathcal{G}(\tilde{\mathbf{t}}^\top \boldsymbol{\theta}) \{1 - \mathcal{G}(\tilde{\mathbf{t}}^\top \boldsymbol{\theta})\} \tilde{\mathbf{t}} \tilde{\mathbf{t}}^\top$, with $\tilde{\mathbf{t}} = (1, \mathbf{t}^\top)^\top$. Since $|T_j^*(x)| = O(\log |x|)$ by (D.4) and $\max_{i \in [n]} \log(1 + X_{ij}^2) = O_p(\log n)$, the event

$$\mathcal{E}_n = \left\{ \max_{i \in [n]} \|\mathbf{T}_i^*\|_2 \leq C \log n \text{ and } \max_{i \in [n]} \|\hat{\mathbf{T}}_i\|_2 \leq C \log n \right\}$$

satisfies $P(\mathcal{E}_n) \rightarrow 1$ for a sufficiently large constant $C > 0$. On $\mathcal{E}_n$, $H(\mathbf{t}; \boldsymbol{\theta}^\circ)$ is continuously differentiable on $\{\|\mathbf{t}\|_2 \leq C \log n\}$ with gradient of order $O((\log n)^2)$, hence Lipschitz there with a non-random constant $L_b = O((\log n)^q)$ for some fixed $q < \infty$. Therefore, on $\mathcal{E}_n$,

$$\frac{1}{n} \sum_{i=1}^n \left\| H(\hat{\mathbf{T}}_i; \boldsymbol{\theta}^\circ) - H(\mathbf{T}_i^*; \boldsymbol{\theta}^\circ) \right\|_{\text{op}} \leq L_b \cdot \max_{i \in [n]} \left\| \hat{\mathbf{T}}_i - \mathbf{T}_i^* \right\|_2,$$

where, as previously established, $\max_{i \in [n]} \|\hat{\mathbf{T}}_i - \mathbf{T}_i^*\|_2 = O_p(\log n / \sqrt{n})$. This implies that the norm (b) in (D.2) is $o_p(1)$.

It remains to bound the norm (c) in (D.2), which is

$$\begin{aligned} \left\| \tilde{\mathbf{I}}(\boldsymbol{\theta}^\circ) - \mathbf{I}^* \right\|_{\text{op}} &= \left\| \tilde{\mathbf{I}}(\boldsymbol{\theta}^\circ) - E \left\{ \tilde{\mathbf{I}}(\boldsymbol{\theta}^\circ) \right\} \right\|_{\text{op}} \\ &= \left\| \frac{1}{n} \sum_{i=1}^n H(\mathbf{T}_i^*; \boldsymbol{\theta}^\circ) - E \left\{ H(\mathbf{T}_1^*; \boldsymbol{\theta}^\circ) \right\} \right\|_{\text{op}}, \end{aligned}$$

since $\{(Y_i, \mathbf{T}_i^*)\}_{i=1}^n$ are i.i.d., $\tilde{\mathbf{I}}(\boldsymbol{\theta}^\circ)$ converges in probability to its expectation by the weak law of large numbers. It follows that

$$\left\| \tilde{\mathbf{I}}(\boldsymbol{\theta}^\circ) - \mathbf{I}^* \right\|_{\text{op}} \leq \left\| \tilde{\mathbf{I}}(\boldsymbol{\theta}^\circ) - \mathbf{I}^* \right\|_F \leq (p+1) \left\| \tilde{\mathbf{I}}(\boldsymbol{\theta}^\circ) - \mathbf{I}^* \right\|_\infty = o_p(1).$$

Combining results established for the three norms (a)–(c) in (D.2) yields $\sup_{\boldsymbol{\theta} \in \mathcal{B}_n(W)} \left\| \hat{\mathbf{I}}(\boldsymbol{\theta}) - \mathbf{I}^* \right\|_{\text{op}} = o_p(1)$, and thus $\sup_{\boldsymbol{\theta} \in \mathcal{B}_n(W)} \left\| \hat{\mathbf{I}}_{SS}(\boldsymbol{\theta}) - \mathbf{I}_{SS}^* \right\|_{\text{op}} = o_p(1)$. Therefore, with probability tending to one, for all $\boldsymbol{\theta} \in \mathcal{B}_n(W)$, $\lambda_{\min}(\hat{\mathbf{I}}_{SS}(\boldsymbol{\theta})) \geq \kappa$. Equivalently, $\mathbf{u}^\top \hat{\mathbf{I}}_{SS}(\boldsymbol{\theta}) \mathbf{u} \geq \kappa \|\mathbf{u}\|_2^2$ for every $\mathbf{u} \in \mathbb{R}^{|\mathcal{S}|}$, which establishes local restricted strong convexity of $Q_n(\boldsymbol{\theta}; \hat{\mathbf{T}})$ on the signal block.

Secondly, we study the penalized MLE $\hat{\boldsymbol{\theta}}$ that minimizes the LASSO-penalized objective function,

$$\mathcal{L}_n(\boldsymbol{\theta}) = Q_n(\boldsymbol{\theta}; \hat{\mathbf{T}}) + \lambda_n \|\boldsymbol{\beta}\|_1, \quad (\text{D.9})$$

where the intercept β_0 in $\boldsymbol{\theta}$ is not penalized. In particular, we will show that, for a sufficiently large constant W , $\hat{\boldsymbol{\theta}} \in \mathcal{B}_n(W)$ with probability tending to one as $n \rightarrow \infty$.

Let $\boldsymbol{\theta} = \boldsymbol{\theta}^\circ + \boldsymbol{\nu}$, where $\boldsymbol{\nu} = (\nu_0, \boldsymbol{\nu}_\beta^\top)^\top$ satisfies $\|\boldsymbol{\nu}\|_2 = Wn^{-1/2}$. Consider

$$\begin{aligned}
 & D_n(\boldsymbol{\nu}) \\
 &= \mathcal{L}_n(\boldsymbol{\theta}^\circ + \boldsymbol{\nu}) - \mathcal{L}_n(\boldsymbol{\theta}^\circ) \\
 &= Q_n(\boldsymbol{\theta}^\circ + \boldsymbol{\nu}; \hat{\mathbf{T}}) - Q_n(\boldsymbol{\theta}^\circ; \hat{\mathbf{T}}) + \lambda_n(\|\boldsymbol{\beta}^\circ + \boldsymbol{\nu}_\beta\|_1 - \|\boldsymbol{\beta}^\circ\|_1) \\
 &= \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} Q_n(\boldsymbol{\theta}; \hat{\mathbf{T}}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\circ} \right\}^\top \boldsymbol{\nu} + \frac{1}{2} \boldsymbol{\nu}^\top \left\{ \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} Q_n(\boldsymbol{\theta}; \hat{\mathbf{T}}) \Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \right\} \boldsymbol{\nu} \quad (\text{D.10}) \\
 &\quad + \lambda_n(\|\boldsymbol{\beta}^\circ + \boldsymbol{\nu}_\beta\|_1 - \|\boldsymbol{\beta}^\circ\|_1),
 \end{aligned}$$

where we use a second-order Taylor expansion of $Q_n(\boldsymbol{\theta}^\circ + \boldsymbol{\nu}; \hat{\mathbf{T}})$ around $\boldsymbol{\theta}^\circ$ to reach the last step, bringing in $\tilde{\boldsymbol{\theta}}$ that lies on the line segment between $\boldsymbol{\theta}^\circ$ and $\boldsymbol{\theta}^\circ + \boldsymbol{\nu}$, and thus $\tilde{\boldsymbol{\theta}} \in \mathcal{B}_n(W)$. Next, we bound the above three terms that $D_n(\boldsymbol{\nu})$ decomposes into separately, exploiting the block structure linking to the partition of $\boldsymbol{\nu}$ into two subvectors: $\boldsymbol{\nu}_S$ and $\boldsymbol{\nu}_{\mathcal{I}_N}$. Equivalently, there is the partition of the subvector of $\boldsymbol{\nu}_\beta$ into $\boldsymbol{\nu}_{\mathcal{I}}$ and $\boldsymbol{\nu}_{\mathcal{I}_N}$, with $\|\boldsymbol{\nu}_\beta\|_1 = \|\boldsymbol{\nu}_{\mathcal{I}}\|_1 + \|\boldsymbol{\nu}_{\mathcal{I}_N}\|_1$.

We first look into the curvature term, i.e., the second term in (D.10) without the multiplier of 1/2 that is equal to

$$\boldsymbol{\nu}^\top \hat{\mathbf{I}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\nu} = \boldsymbol{\nu}_S^\top \hat{\mathbf{I}}_{SS}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\nu}_S + 2\boldsymbol{\nu}_S^\top \hat{\mathbf{I}}_{S\mathcal{I}_N}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\nu}_{\mathcal{I}_N} + \boldsymbol{\nu}_{\mathcal{I}_N}^\top \hat{\mathbf{I}}_{\mathcal{I}_N\mathcal{I}_N}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\nu}_{\mathcal{I}_N}. \quad (\text{D.11})$$

By the restricted strong convexity established above, $\boldsymbol{\nu}_S^\top \hat{\mathbf{I}}_{SS}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\nu}_S \geq \kappa \|\boldsymbol{\nu}_S\|_2^2$ on the event $E_{1,n}$ that this bound holds uniformly on $\mathcal{B}_n(W)$. Turn to the non-signal blocks. For $j \in \mathcal{I}_N$ the oracle transform vanishes, $T_j^* \equiv 0$, and

the first part of this proof, following (D.7), gives $\max_{i \in [n], j \in [p]} |\hat{T}_j(X_{ij}) - T_j^*(X_{ij})| = O_p(\log n / \sqrt{n})$. The noise entries of $\tilde{\mathbf{T}}_i$ are therefore uniformly $O_p(\log n / \sqrt{n})$. Combined with $n^{-1} \sum_{i=1}^n \|(\tilde{\mathbf{T}}_i)_S\|_2 = O_p(1)$, which holds by the same logarithmic-moment bound in (D.4) that yields $L_a = O_p(1)$, this gives $\|\hat{\mathbf{I}}_{S\mathcal{I}_N}(\tilde{\boldsymbol{\theta}})\|_{\text{op}} = O_p(\log n / \sqrt{n})$ and $\|\hat{\mathbf{I}}_{\mathcal{I}_N\mathcal{I}_N}(\tilde{\boldsymbol{\theta}})\|_{\text{op}} = O_p(\log^2 n / n)$. It follows that

$$\boldsymbol{\nu}^\top \hat{\mathbf{I}}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\nu} \geq \kappa \|\boldsymbol{\nu}_S\|_2^2 - 2 \|\hat{\mathbf{I}}_{S\mathcal{I}_N}(\tilde{\boldsymbol{\theta}})\|_{\text{op}} \|\boldsymbol{\nu}_S\|_2 \|\boldsymbol{\nu}_{\mathcal{I}_N}\|_2 \geq \kappa \|\boldsymbol{\nu}_S\|_2^2 + o_p(n^{-1}),$$

where the first inequality results from dropping $\boldsymbol{\nu}_{\mathcal{I}_N}^\top \hat{\mathbf{I}}_{\mathcal{I}_N\mathcal{I}_N}(\tilde{\boldsymbol{\theta}}) \boldsymbol{\nu}_{\mathcal{I}_N} \geq 0$ from (D.11) (since $\hat{\mathbf{I}}(\tilde{\boldsymbol{\theta}}) \succeq 0$), and we use $\|\boldsymbol{\nu}\|_2^2 = W^2 n^{-1}$ to reach the step. Hence the second term in (D.10) is at least $(\kappa/2) \|\boldsymbol{\nu}_S\|_2^2 + o_p(n^{-1})$.

Consider first the score term, the first term in (D.10). The score at $\boldsymbol{\theta}^\circ$ separates across blocks. By Theorem 3, the signal block is of order $O_p(n^{-1/2})$ and the noise block of order $O_p(n^{-1})$. Theorem 3 does not cover the intercept score, which is $O_p(n^{-1/2})$ by the central limit theorem. The argument is the one used for the signal scores, with the constant multiplier one in place of $\hat{T}_j(X_{ij})$. Thus, for any $\varepsilon > 0$ there are finite constants C_ε, C_g and an integer n_ε such that, for all $n > n_\varepsilon$, on an event $E_{2,n}$ with $P(E_{2,n}) \geq 1 - \varepsilon$,

$$\left\| \frac{\partial}{\partial \boldsymbol{\theta}_S} Q_n(\boldsymbol{\theta}; \hat{\mathbf{T}}) \right\|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\circ} \leq C_\varepsilon n^{-1/2} \text{ and } \max_{j \in \mathcal{I}_N} \left| \frac{\partial}{\partial \beta_j} Q_n(\boldsymbol{\theta}; \hat{\mathbf{T}}) \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\circ} \leq C_g n^{-1}.$$

On $E_{2,n}$, by the Cauchy–Schwarz and Hölder inequalities,

$$\{(\partial/\partial\boldsymbol{\theta})Q_n(\boldsymbol{\theta}^\circ; \hat{\mathbf{T}})\}^\top \boldsymbol{\nu} \geq -C_\varepsilon n^{-1/2} \|\boldsymbol{\nu}_S\|_2 - C_g n^{-1} \|\boldsymbol{\nu}_{\mathcal{I}_N}\|_1.$$

For the penalty term following (D.10), since $\boldsymbol{\beta}_{\mathcal{I}_N}^\circ = \mathbf{0}$ and the intercept is unpenalized, the reverse triangle inequality gives

$$\lambda_n(\|\boldsymbol{\beta}^\circ + \boldsymbol{\nu}_\beta\|_1 - \|\boldsymbol{\beta}^\circ\|_1) \geq \lambda_n(\|\boldsymbol{\nu}_{\mathcal{I}_N}\|_1 - \|\boldsymbol{\nu}_T\|_1) \geq \lambda_n \|\boldsymbol{\nu}_{\mathcal{I}_N}\|_1 - C_\lambda n^{-1/2} \sqrt{p} \|\boldsymbol{\nu}_S\|_2,$$

where $\lambda_n \asymp n^{-1/2}$ provides a finite C_λ with $\lambda_n \leq C_\lambda n^{-1/2}$, and $\|\boldsymbol{\nu}_T\|_1 \leq \sqrt{p} \|\boldsymbol{\nu}_S\|_2$.

Combining the three bounds on $E_{1,n} \cap E_{2,n}$ and grouping the noise contribution,

$$D_n(\boldsymbol{\nu}) \geq \frac{\kappa}{2} \|\boldsymbol{\nu}_S\|_2^2 - (C_\varepsilon + C_\lambda \sqrt{p}) n^{-1/2} \|\boldsymbol{\nu}_S\|_2 + (\lambda_n - C_g n^{-1}) \|\boldsymbol{\nu}_{\mathcal{I}_N}\|_1 + o_p(n^{-1}).$$

Because $\lambda_n \asymp n^{-1/2} \gg n^{-1}$, there is $c_\lambda > 0$ with $\lambda_n - C_g n^{-1} \geq c_\lambda n^{-1/2}$ for large n . Using $\|\boldsymbol{\nu}_{\mathcal{I}_N}\|_1 \geq \|\boldsymbol{\nu}_{\mathcal{I}_N}\|_2$ and writing $a = \|\boldsymbol{\nu}_S\|_2$, $b = \|\boldsymbol{\nu}_{\mathcal{I}_N}\|_2$ with $a^2 + b^2 = W^2 n^{-1}$,

$$D_n(\boldsymbol{\nu}) \geq \frac{\kappa}{2} a^2 - (C_\varepsilon + C_\lambda \sqrt{p}) n^{-1/2} a + c_\lambda n^{-1/2} b + o_p(n^{-1}).$$

If $a \geq W n^{-1/2} / \sqrt{2}$, the first two terms are bounded below by $n^{-1} \left\{ \frac{\kappa}{4} W^2 - (C_\varepsilon + C_\lambda \sqrt{p}) W \right\}$.

If $b \geq W n^{-1/2} / \sqrt{2}$, minimizing the first two terms over $a \geq 0$ gives the lower bound $-\frac{(C_\varepsilon + C_\lambda \sqrt{p})^2}{2\kappa} n^{-1}$, whereas $c_\lambda n^{-1/2} b \geq \frac{c_\lambda W}{\sqrt{2}} n^{-1}$. Thus, for every $\varepsilon > 0$, one can choose a finite constant W_ε sufficiently large so that

$\inf_{\|\boldsymbol{\nu}\|_2=W_\varepsilon n^{-1/2}} D_n(\boldsymbol{\nu}) > 0$ on an event whose probability has limit inferior at least $1 - \varepsilon$.

To see that this localizes every minimizer, suppose that a minimizer $\hat{\boldsymbol{\theta}}$ lies outside $\mathcal{B}_n(W_\varepsilon)$. The line segment joining $\boldsymbol{\theta}^\circ$ and $\hat{\boldsymbol{\theta}}$ then intersects the boundary of $\mathcal{B}_n(W_\varepsilon)$ at some point $\boldsymbol{\theta}_t$. By convexity,

$$\mathcal{L}_n(\boldsymbol{\theta}_t) \leq (1-t)\mathcal{L}_n(\boldsymbol{\theta}^\circ) + t\mathcal{L}_n(\hat{\boldsymbol{\theta}}) \leq \mathcal{L}_n(\boldsymbol{\theta}^\circ),$$

contradicting the strict positivity of $\mathcal{L}_n(\boldsymbol{\theta}_t) - \mathcal{L}_n(\boldsymbol{\theta}^\circ)$ on that boundary. Consequently,

$$\limsup_{n \rightarrow \infty} P \left\{ \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\|_2 > W_\varepsilon n^{-1/2} \right\} \leq \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\|_2 = O_p(n^{-1/2})$. Because p is fixed, equivalence of norms in finite-dimensional spaces then gives $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\|_1 = O_p(n^{-1/2})$.

This completes the proof. $\square$

The Karush–Kuhn–Tucker condition for the L_1 -penalized logistic objective in (D.9) (Park and Hastie, 2007) gives the following characterization at an interior solution.

Corollary 1. *Let $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}^\top)^\top \in \text{int}(\boldsymbol{\Theta})$. Then $\hat{\boldsymbol{\theta}}$ minimizes (D.9) if and only if*

$$\left. \frac{\partial}{\partial \beta_0} Q_n(\boldsymbol{\theta}; \hat{\mathbf{T}}) \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$$

and, for every $j \in [p]$,

$$\begin{cases} G_j(\hat{\boldsymbol{\theta}}) = -\lambda_n \text{sign}(\hat{\beta}_j), & \hat{\beta}_j \neq 0, \\ |G_j(\hat{\boldsymbol{\theta}})| \leq \lambda_n, & \hat{\beta}_j = 0. \end{cases}$$

Here,

$$G_j(\hat{\boldsymbol{\theta}}) = -\frac{1}{n} \sum_{i=1}^n \left\{ Y_i - \mathcal{G} \left(\hat{\beta}_0 + \hat{\mathbf{T}}_i^\top \hat{\boldsymbol{\beta}} \right) \right\} \hat{T}_j(X_{ij})$$

is the j -th component of the gradient in (C.5).

Lemma 4. Assume Conditions (A1)–(A3) and (A8). For a finite constant

$C > 0$, let

$$\mathcal{N}_n(C) = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{p+1} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|_1 \leq Cn^{-1/2} \right\}.$$

Then the gradient function in (C.5) satisfies

$$\max_{j \in \mathcal{I}} \sup_{\boldsymbol{\theta} \in \mathcal{N}_n(C)} |G_j(\boldsymbol{\theta}) - G_j(\boldsymbol{\theta}^\circ)| = O_p(n^{-1/2}) \text{ and } \max_{j \in \mathcal{I}_N} \sup_{\boldsymbol{\theta} \in \mathcal{N}_n(C)} |G_j(\boldsymbol{\theta}) - G_j(\boldsymbol{\theta}^\circ)| = O_p(n^{-1}).$$

Proof. Fix $C < \infty$ and let $\boldsymbol{\theta} \in \mathcal{N}_n(C)$ be arbitrary. Every bound below is a product of a stochastic order that does not involve $\boldsymbol{\theta}$ with the deterministic quantity $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|_1 \leq Cn^{-1/2}$, and is therefore uniform over $\mathcal{N}_n(C)$. Consequently, the supremum in the statement is of the same order. Since p is fixed, the maximum over the finitely many indices in each block preserves the order as well.

For $j \in [p]$,

$$\begin{aligned}
 & G_j(\boldsymbol{\theta}) - G_j(\boldsymbol{\theta}^\circ) \\
 &= \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} G_j(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} \right\}^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ), \text{ by the mean value theorem,} \\
 &= \left[\frac{1}{n} \sum_{i=1}^n \mathcal{G}(\bar{\eta}_i) \{1 - \mathcal{G}(\bar{\eta}_i)\} \hat{T}_j(X_{ij}) \begin{pmatrix} 1 \\ \hat{\mathbf{T}}_i \end{pmatrix} \right]^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ), \quad (\text{D.12})
 \end{aligned}$$

where $\bar{\boldsymbol{\theta}} = (\bar{\beta}_0, \bar{\boldsymbol{\beta}}^\top)^\top$ lies on the line segment between $\boldsymbol{\theta}^\circ$ and $\boldsymbol{\theta}$, and thus $\bar{\boldsymbol{\theta}} \in \mathcal{N}_n(C)$ as $\mathcal{N}_n(C)$ is a convex ℓ_1 -ball, and $\bar{\eta}_i = \bar{\beta}_0 + \hat{\mathbf{T}}_i^\top \bar{\boldsymbol{\beta}}$. Denote by $\tilde{\mathbf{H}}_{j\cdot}(\bar{\boldsymbol{\theta}})$ the vector inside the square brackets in (D.12).

Because $\mathcal{G}(\cdot)$ is infinitely differentiable with all derivatives bounded, the $\boldsymbol{\theta}$ -derivative of $\tilde{\mathbf{H}}_{j\cdot}(\boldsymbol{\theta})$ is bounded in supremum norm by a constant multiple of $n^{-1} \sum_{i=1}^n \|\hat{\mathbf{T}}_i\|_2^3 = L_a$, which is $O_p(1)$ by (D.6) and does not depend on j . Hence $\tilde{\mathbf{H}}_{j\cdot}$ is Lipschitz on $\mathcal{N}_n(C)$ with a random constant of order $O_p(1)$, so that $\|\tilde{\mathbf{H}}_{j\cdot}(\bar{\boldsymbol{\theta}}) - \tilde{\mathbf{H}}_{j\cdot}(\boldsymbol{\theta}^\circ)\|_\infty = O_p(\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\|_1) = O_p(n^{-1/2})$. Applying Hölder's inequality to (D.12) therefore gives

$$\begin{aligned}
 G_j(\boldsymbol{\theta}) - G_j(\boldsymbol{\theta}^\circ) &= \tilde{\mathbf{H}}_{j\cdot}(\boldsymbol{\theta}^\circ)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) + O_p(\|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|_1^2) \\
 &= \tilde{\mathbf{H}}_{j\cdot}(\boldsymbol{\theta}^\circ)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) + O_p(n^{-1}). \quad (\text{D.13})
 \end{aligned}$$

In what follows, we analyze the leading term in (D.13), that is,

$$\tilde{\mathbf{H}}_{j\cdot}(\boldsymbol{\theta}^\circ)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) = H_{j0}(\boldsymbol{\theta}^\circ)(\beta_0 - \beta_0^\circ) + \sum_{\ell=1}^p H_{j\ell}(\boldsymbol{\theta}^\circ)(\beta_\ell - \beta_\ell^\circ), \quad (\text{D.14})$$

where

$$\begin{aligned} H_{j0}(\boldsymbol{\theta}^\circ) &= \frac{1}{n} \sum_{i=1}^n \mathcal{G}(\hat{\eta}_i) \{1 - \mathcal{G}(\hat{\eta}_i)\} \hat{T}_j(X_{ij}), \\ H_{j\ell}(\boldsymbol{\theta}^\circ) &= \frac{1}{n} \sum_{i=1}^n \mathcal{G}(\hat{\eta}_i) \{1 - \mathcal{G}(\hat{\eta}_i)\} \hat{T}_j(X_{ij}) \hat{T}_\ell(X_{i\ell}), \text{ for } \ell \in [p], \end{aligned}$$

and $\hat{\eta}_i = \beta_0^\circ + \hat{\mathbf{T}}_i^\top \boldsymbol{\beta}^\circ$ denotes the fitted linear predictor at $\boldsymbol{\theta}^\circ$ based on the estimated transformed features (not to be confused with η_i° in Appendix C, which uses the oracle features). Recall that $\mathcal{G}(\cdot)\{1 - \mathcal{G}(\cdot)\} \in [0, 0.25]$, and that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|_\infty \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|_1 \leq Cn^{-1/2}$ for $\boldsymbol{\theta} \in \mathcal{N}_n(C)$.

In the proof for Theorem 3, we established that $\hat{T}_j(X_{ij}) = T_j^*(X_{ij}) + O_p(n^{-1/2})$, for $j \in [p]$. In particular, for $j \in \mathcal{I}_N$, because $T_j^*(X_{ij}) = 0$, $\hat{T}_j(X_{ij}) = O_p(n^{-1/2})$, and thus $H_{j0}(\boldsymbol{\theta}^\circ) = O_p(n^{-1/2})$, $H_{j\ell}(\boldsymbol{\theta}^\circ) = O_p(n^{-1})$ if $\ell \in \mathcal{I}_N$, and $H_{j\ell}(\boldsymbol{\theta}^\circ) = O_p(n^{-1/2})$ if $\ell \in \mathcal{I}$. Since $\beta_\ell^\circ = 0$ for every $\ell \in \mathcal{I}_N$, each increment $\beta_\ell - \beta_\ell^\circ$ appearing in (D.14) is $O(n^{-1/2})$, so that

$$\begin{aligned} & |\tilde{\mathbf{H}}_{j\cdot}(\boldsymbol{\theta}^\circ)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)| \\ & \leq \underbrace{O_p(n^{-1/2})O(n^{-1/2})}_{\text{intercept}} + \underbrace{O_p(n^{-1/2})O(n^{-1/2})}_{\ell \in \mathcal{I}} + \underbrace{O_p(n^{-1})O(n^{-1/2})}_{\ell \in \mathcal{I}_N} \\ & = O_p(n^{-1}), \end{aligned}$$

and thus $G_j(\boldsymbol{\theta}) - G_j(\boldsymbol{\theta}^\circ) = O_p(n^{-1})$ by (D.13).

In contrast, for $j \in \mathcal{I}$, and thus $\hat{T}_j(X_{ij}) = O_p(1)$ now that $T_j^*(\cdot) \not\equiv 0$, we have $H_{j0}(\boldsymbol{\theta}^\circ) = O_p(1)$, $H_{j\ell}(\boldsymbol{\theta}^\circ) = O_p(n^{-1/2})$ if $\ell \in \mathcal{I}_N$, and $H_{j\ell}(\boldsymbol{\theta}^\circ) = O_p(1)$

if $\ell \in \mathcal{I}$. Putting these together in (D.14) gives the order of this leading term as $O_p(n^{-1/2})$, and thus $G_j(\boldsymbol{\theta}) - G_j(\boldsymbol{\theta}^\circ) = O_p(n^{-1/2})$ by (D.13). This completes the proof. $\square$

Theorem 4. *Under Conditions (A1)–(A3) and (A8), if the penalty parameter in (D.9) is $\lambda = \lambda_n \asymp n^{-1/2}$, then the penalized MLE $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}^\top)^\top$ satisfies*

$$\lim_{n \rightarrow \infty} P(\hat{\beta}_j = 0 \text{ for all } j \in \mathcal{I}_N) = 1 \text{ and } \lim_{n \rightarrow \infty} P(\hat{\beta}_j \neq 0 \text{ for all } j \in \mathcal{I}) = 1.$$

In particular, $P(\{j \in [p] : \hat{\beta}_j \neq 0\} = \mathcal{I}) \rightarrow 1$.

Proof. Fix $\varepsilon > 0$. By Lemma 3, there exists a finite constant C_ε such that, for all sufficiently large n ,

$$P\{\hat{\boldsymbol{\theta}} \in \mathcal{N}_n(C_\varepsilon)\} \geq 1 - \varepsilon, \quad \mathcal{N}_n(C_\varepsilon) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|_1 \leq C_\varepsilon n^{-1/2}\}.$$

Condition (A3) places $\boldsymbol{\theta}^\circ$ in the interior of $\boldsymbol{\Theta}$, and the radius of $\mathcal{N}_n(C_\varepsilon)$ tends to zero. Hence $\mathcal{N}_n(C_\varepsilon) \subset \text{int}(\boldsymbol{\Theta})$ for all large n , so that $\hat{\boldsymbol{\theta}} \in \text{int}(\boldsymbol{\Theta})$ on the same event and Corollary 1 applies to $\hat{\boldsymbol{\theta}}$.

First, we show no false positives. Because p is fixed, Theorem 3 gives $\max_{j \in \mathcal{I}_N} |G_j(\boldsymbol{\theta}^\circ)| = O_p(n^{-1})$, while Lemma 4 gives

$$\max_{j \in \mathcal{I}_N} \sup_{\boldsymbol{\theta} \in \mathcal{N}_n(C_\varepsilon)} |G_j(\boldsymbol{\theta}) - G_j(\boldsymbol{\theta}^\circ)| = O_p(n^{-1}).$$

By the triangle inequality, on the event $\{\hat{\boldsymbol{\theta}} \in \mathcal{N}_n(C_\varepsilon)\}$,

$$\max_{j \in \mathcal{I}_N} |G_j(\hat{\boldsymbol{\theta}})| \leq \max_{j \in \mathcal{I}_N} |G_j(\boldsymbol{\theta}^\circ)| + \max_{j \in \mathcal{I}_N} \sup_{\boldsymbol{\theta} \in \mathcal{N}_n(C_\varepsilon)} |G_j(\boldsymbol{\theta}) - G_j(\boldsymbol{\theta}^\circ)| = O_p(n^{-1}).$$

Since $\lambda_n \asymp n^{-1/2}$ and $n^{-1} = o(n^{-1/2})$, it follows that $P\{\max_{j \in \mathcal{I}_N} |G_j(\hat{\boldsymbol{\theta}})| < \lambda_n\} \rightarrow 1$. By Corollary 1, $\hat{\beta}_j \neq 0$ would force $|G_j(\hat{\boldsymbol{\theta}})| = \lambda_n$. Therefore

$$\liminf_{n \rightarrow \infty} P(\hat{\beta}_j = 0 \text{ for all } j \in \mathcal{I}_N) \geq 1 - \varepsilon.$$

As $\varepsilon > 0$ was arbitrary, the first claim follows.

Second, we show no false negatives. Recall from the definition of the canonical sparse representative in Appendix A that $\beta_j^\circ = 1$ for every $j \in \mathcal{I}$, a value that depends neither on n nor on how informative feature j is. Consider the event $A_n = \{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\|_\infty < \frac{1}{2}\}$. On A_n , simultaneously for every $j \in \mathcal{I}$, the reverse triangle inequality gives,

$$|\hat{\beta}_j| = |\beta_j^\circ + (\hat{\beta}_j - \beta_j^\circ)| \geq |\beta_j^\circ| - |\hat{\beta}_j - \beta_j^\circ| \geq 1 - \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\|_\infty > \frac{1}{2}.$$

Consequently,

$$A_n \subseteq \left\{ \min_{j \in \mathcal{I}} |\hat{\beta}_j| > \frac{1}{2} \right\} \subseteq \{\hat{\beta}_j \neq 0 \text{ for all } j \in \mathcal{I}\}. \quad (\text{D.15})$$

Therefore, it suffices to show that $P(A_n) \rightarrow 1$. Since the supremum norm is dominated by the ℓ_1 norm on $\mathbb{R}^{p+1}$, Lemma 3 yields $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\|_\infty \leq \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\|_1 = O_p(n^{-1/2})$. As $n^{-1/2} \rightarrow 0$, this is $o_p(1)$, so that $P(A_n^c) = P(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\|_\infty \geq 1/2) \rightarrow 0$. Together with (D.15), this establishes the second assertion.

Two points of the argument are worth noting. First, only the consistency $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ\|_\infty = o_p(1)$ has been used, not the $n^{-1/2}$ rate. Second, the threshold $1/2$ may be replaced by any fixed $c \in (0, 1)$, since all that is needed is a level strictly between 0 and $\min_{j \in \mathcal{I}} |\beta_j^\circ| = 1$.

Let $\hat{\mathcal{A}} = \{j \in [p] : \hat{\beta}_j \neq 0\}$ denote the selected set, and put

$$B_n = \{\hat{\beta}_j = 0 \text{ for all } j \in \mathcal{I}_N\}, \quad C_n = \{\hat{\beta}_j \neq 0 \text{ for all } j \in \mathcal{I}\},$$

so that $B_n \cap C_n = \{\hat{\mathcal{A}} = \mathcal{I}\}$. The two assertions established above state that $P(B_n) \rightarrow 1$ and $P(C_n) \rightarrow 1$, whence Bonferroni's inequality gives

$$P(\hat{\mathcal{A}} = \mathcal{I}) \geq 1 - P(B_n^c) - P(C_n^c) \rightarrow 1.$$

This completes the proof. □

Appendix E: Additional Simulation Results in Section

4

This appendix provides details on the twelve data-generating mechanisms described in Section 4, and the computation times across the eight classifiers.

E.1 Design of the twelve simulation cases

Table E.1 lists the class-conditional distributions of the informative features in $\mathbf{X}_{\mathcal{I}}$ and indicates whether they are conditionally independent given the class label. Unless stated otherwise, we set $n = 800$, $p = 30$, and $|\mathcal{I}| = 20$. The class labels are drawn independently from Bernoulli(0.5), and the $p - |\mathcal{I}|$ noise features are generated from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{p-|\mathcal{I}|})$ in both classes. Throughout, $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 ; $\text{Cauchy}(a, b)$ refers to a Cauchy distribution with median parameter a and scale parameter b ; $\text{Lognormal}(a, b)$ is defined by $\log X \sim \mathcal{N}(a, b^2)$; t_ν is the Student- t distribution with ν degrees of freedom; $\Phi(\cdot)$ denotes the standard normal cumulative distribution function (CDF); and $U(a, b)$ is a uniform distribution supported on (a, b) . Lastly, $\mathbf{1}_d$ and $\mathbf{I}_d$ denote the d -dimensional all-ones vector and $d \times d$ identity matrix, respectively.

The twelve cases are organized around four departures from the standard setting. Cases 1–4 vary the shape of the class-conditional distributions while maintaining a location shift between classes. Specifically, the distributions are Gaussian in Case 1, heavy-tailed Cauchy (with no defined moments) in Case 2, and skewed log-normal in Case 3. In Case 4, the two classes follow entirely different distribution families: a standard normal distribution versus a shifted t_2 .

Table E.1: Generative models of the informative features $\mathbf{X}_{\mathcal{I}}$ in the twelve simulation cases.

Case	$X_j \mid Y = 0$	$X_j \mid Y = 1$	$X_j \perp X_{j'} \mid Y$
1	$\mathcal{N}(-0.5, 4)$	$\mathcal{N}(0.5, 4)$	Yes
2	Cauchy(0, 1)	Cauchy(0.5, 1)	Yes
3	Lognormal(0, 1)	Lognormal(0.5, 1)	Yes
4	$\mathcal{N}(0, 1)$	$t_2 + 0.5$	Yes
5	$0.9\mathcal{N}(\mu_{0j}, \sigma_j^2)$ $+ 0.1\mathcal{N}(\mu_{0j}, 100\sigma_j^2)$	$0.9\mathcal{N}(\mu_{1j}, \sigma_j^2)$ $+ 0.1\mathcal{N}(\mu_{1j}, 100\sigma_j^2)$	Yes
6	$\mathbf{X}_{\mathcal{I}} \mid Y = 0 \sim \mathcal{N}(\mathbf{0}_{ \mathcal{I} }, \mathbf{\Sigma})$	$\mathbf{X}_{\mathcal{I}} \mid Y = 1 \sim \mathcal{N}(\mathbf{0.31}_{ \mathcal{I} }, 2\mathbf{\Sigma})$	No
7	$\mathbf{X}_{\mathcal{I}} \mid Y = 0 \sim \mathcal{N}(\mathbf{0.51}_{ \mathcal{I} }, \tilde{\mathbf{\Sigma}})$	$\mathbf{X}_{\mathcal{I}} \mid Y = 1 \sim \mathcal{N}(-\mathbf{0.51}_{ \mathcal{I} }, \tilde{\mathbf{\Sigma}})$	No
8	$\mathbf{X}_{\mathcal{I}} \sim \mathcal{N}(\mathbf{0}_{ \mathcal{I} }, \mathbf{I}_{ \mathcal{I} }), \quad Y \mid \mathbf{X}_{\mathcal{I}} \sim \text{Bernoulli}\{\Phi(h(\mathbf{X}_{\mathcal{I}}))\}$		No
9	$\mathcal{N}(0, 1)$	$t_2 + 0.5$	Yes
10	$-\{Z - E(Z)\}$	$Z - E(Z)$	Yes
11	bimodal mixture normal	mirror image of f_{j0}	No
12	bimodal mixture t	mirror image of f_{j0}	No

Case 5 adds contamination. For each class, the distribution is a two-component Gaussian mixture whose components share the same location parameter (μ_{0j} or μ_{1j}) but differ in scale, the second component scaling up

σ_j^2 by a factor of 100. Since these components are concentric, the resulting marginal distribution preserves unimodality while generating heavy tails and occasional extreme observations instead of bimodality.

Cases 6 and 7 relax the conditional independence assumption. Both employ an autoregressive covariance structure, $\Sigma_{jj'} = \rho^{|j-j'|}$ for $j, j' = 1, \dots, |\mathcal{I}|$, with $\rho = 0.2$ in Case 6 and $\rho = 0.8$ in Case 7. Additionally, Case 6 introduces unequal class covariances by setting the class-1 covariance to 2Σ , so that the two classes differ in both location and scale.

Case 8 reverses the generative direction: the informative features are independently sampled from a standard normal distribution across both classes, while the class label is generated conditionally via a probit regression model with a non-linear predictor given by

$$h(\mathbf{X}_{\mathcal{I}}) = \sum_{j=1}^8 0.4 (X_j^2 - 1) + \sum_{j=9}^{14} 0.5 X_j X_{j+6} + \sum_{j=15}^{20} 0.5 \sin(3\pi(X_{j-1} + X_j)),$$

which incorporates quadratic, interaction, and oscillatory effects. This scenario therefore removes an explicitly specified class-conditional marginal model and challenges classifiers whose discrimination relies primarily on linear location shifts.

Case 9 is the high-dimensional counterpart of Case 4. The class-conditional distributions are unchanged, while $n = 300$ and $p = 700$, so that the 680 noise features far outnumber the observations and the signal must be re-

covered in a small- n -large- p regime.

Cases 10–12 have mean-matched informative features across classes: the two class-conditional distributions share a common mean, so that the classes are separated by shape as opposed to by some location shift. Specifically, informative features in Case 10 are generated from a centered TPSC distribution. Let $Z_j^* \sim \text{TPSC}(\Psi)$ with $\Psi = (\tau, w, \sigma, \delta)^\top = (0, 0.3, 1.5, 4)^\top$, and set $Z_j = Z_j^* - E(Z_j^*)$. Then, for $j \in \mathcal{I}$, X_j under the two classes are mirror images, $-Z_j$ and Z_j , sharing the same mean but exhibiting opposite skewness. Cases 11 and 12 replace the skewness with bimodality so that Hartigan’s dip test (Hartigan and Hartigan, 1985) rejects unimodality at the 0.05 significance level across all replicates. For each informative feature X_j , define a mean-zero two-component mixture distribution

$$M_j = \pi_j F(m_{1j}, s_j) + (1 - \pi_j) F(m_{2j}, s_j), \text{ with } m_{2j} = -\frac{\pi_j m_{1j}}{1 - \pi_j},$$

and the distributions of X_j under two classes are again the mirror images specified by M_j and $-M_j$. The component distribution in M_j belongs to some location-scale family, i.e., $F(m, s) = m + sZ_0$. In Case 11, $Z_0 \sim \mathcal{N}(0, 1)$, and in Case 12, $Z_0 \sim t_6$. The two cases therefore isolate the effect of adding heavy tails to a bimodal mechanism. The component parameters are drawn independently as

$$\pi_j \sim U(0.58, 0.66), \quad m_{1j} \sim U(-1.5, -1.1), \quad s_j \sim U(0.50, 0.65)$$

in Case 11, and, in Case 12,

$$\pi_j \sim U(0.58, 0.66), \quad m_{1j} \sim U(-1.6, -1.2), \quad s_j \sim U(0.48, 0.58).$$

Dependence is introduced through a Gaussian copula, with the copula function defined as the CDF of a multivariate Gaussian. Specifically, two independent latent vectors $\mathbf{G}^{(1)}, \mathbf{G}^{(2)} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ with $\mathbf{R}_{jj'} = 0.6^{|j-j'|}$ are drawn per observation. The first selects the mixture component of each feature and the second generates the within-component variation, so that the informative features are dependent as well as bimodal. Cases 11 and 12 are the only settings whose informative features are multimodal.

E.2 Computing times

Table E.2 reports the median (across 300 Monte Carlo replicates) computation time per fit-and-predict cycle for the eight classifiers from the simulation experiment in Section 4.

Appendix F: Numerical Implementation Details

This appendix describes the two implementation steps of S-RoLLR: estimating the class-specific TPSC parameters $\Psi = (\tau, w, \sigma, \delta)^\top$ and selecting the penalty parameter λ_n . Sections F.1–F.3 assess the numerical and statistical behavior of the first step, and Section F.4 gives the tuning protocol

Table E.2: Median computing times in seconds across 300 Monte Carlo replicates in twelve cases, with the lowest value in bold and the highest value italicized in each case.

Method	1	2	3	4	5	6	7	8	9	10	11	12
RoLLR	3.96	2.43	3.20	3.10	2.79	3.64	3.58	3.64	43.80	3.00	4.20	4.03
S-RoLLR	4.22	2.40	3.47	3.86	3.05	3.82	3.88	3.92	41.10	3.27	4.41	4.15
MOKE	0.37	0.31	0.33	0.36	0.36	0.36	0.37	0.35	4.59	0.33	0.35	0.34
GMM-NB	<i>28.40</i>	<i>17.50</i>	<i>22.60</i>	<i>26.00</i>	<i>22.30</i>	<i>29.00</i>	<i>30.20</i>	<i>29.30</i>	<i>291.0</i>	<i>25.40</i>	<i>26.50</i>	<i>26.60</i>
L1-RL	0.61	0.46	0.50	0.50	0.45	0.44	0.66	0.39	1.36	0.36	0.39	0.36
SVM	3.81	3.29	3.19	3.27	3.67	3.67	3.30	4.03	13.50	3.61	3.79	3.66
RF	2.08	1.93	1.77	2.04	2.00	2.11	1.80	2.29	14.30	2.05	1.69	1.80
LightGBM	3.77	2.92	2.64	3.18	3.22	3.19	3.41	3.88	3.02	2.26	2.52	2.71

for the second.

F.1 Estimating the TPSC parameters

TPSC joins two Student- t components on disjoint half-lines at a common mode and hence avoids the label-switching ambiguity of ordinary finite mixtures. Its identifiability is established in Liu et al. (2024). We maximize its log-likelihood with `optim` in R.

The mode τ is unconstrained, while $w \in [0.05, 0.95]$, $\sigma > 10^{-6}$, and $\delta \in [0.5, 100]$. Let $\tilde{x}$ be the sample median and $\text{MAD} = 1.4826 \text{median}(|x_i -$

$\tilde{x}|)$. We run L-BFGS-B (Byrd et al., 1995) from the four deterministic initial values $(\tilde{x}, 0.3, \text{MAD}, 3)$, $(\tilde{x}, 0.7, \text{MAD}, 10)$, $(\tilde{x}, 0.5, 0.5 \text{ MAD}, 2)$, and $(\tilde{x}, 0.5, 1.5 \text{ MAD}, 15)$, and retain the estimate with the highest log-likelihood as $\hat{\Psi}$.

These choices raise two distinct concerns. The first is optimization error: four starting points and a constrained search domain may fail to locate the global (or a sufficiently good) maximizer. The second is statistical error, which persists even at the global optimum because finite sample sizes can limit the identifiability and estimation accuracy of Ψ . Section F.2 addresses the former, while Section F.3 focuses on the latter. Both studies use the four class-0 configurations Ψ_0 in Table F.1. Two-class metrics use independent samples of size n from Ψ_0 and $\Psi_1 = (\tau_0 + \frac{1}{2}\sigma_0, w_0, \sigma_0, \delta_0)^\top$. Single-density metrics use one sample of size n from Ψ_0 . Each generated dataset has equal sample sizes across both classes. Section F.2 uses $n \in \{20, 50, 100, 400\}$ and 500 replicates per configuration, while Section F.3 uses $n \in \{10, 20, 30, 50, 100, 200, 500, 1000\}$ and 1000 replicates. Competing procedures are evaluated on the same generated samples.

Let $f(\cdot | \Psi)$ denote the TPSC density. For a generic feature, define the estimated and oracle transformations by

$$\hat{T}(x) = \log f(x | \hat{\Psi}_1) - \log f(x | \hat{\Psi}_0), \quad T^*(x) = \log f(x | \Psi_1) - \log f(x | \Psi_0).$$

Table F.1: Four configurations of the class-0 TPSC parameter vector Ψ_0 used in Sections F.2 and F.3

Label	Description	τ_0	w_0	σ_0	δ_0
S1	near-Gaussian symmetric	0	0.50	1	30
S2	symmetric heavy tail	0	0.50	1	2
S3	skewed moderate tail	0	0.30	1	5
S4	skewed Cauchy-like tail	0	0.25	1	1

We measure transformation accuracy by

$$\text{TMSE} = \int_{\mathbb{R}} \{\hat{T}(x) - T^*(x)\}^2 \frac{1}{2} \{f(x \mid \Psi_0) + f(x \mid \Psi_1)\} dx,$$

and density accuracy by $\text{KL}(f_0 \parallel \hat{f}) = \int f(x \mid \Psi_0) \log\{f(x \mid \Psi_0)/f(x \mid \hat{\Psi}_0)\} dx$. The integrals are evaluated by deterministic quadrature with a double-exponential rule on the tails and the density modes as interior break-points.

F.2 Optimization error

We compare the four-start procedure with two more intensive alternatives. The first uses fifty Latin-hypercube starting points (McKay et al., 2000) over the same parameter domain and serves as a near-global reference. The second maps the constrained parameters to the real line and applies

the derivative-free Nelder–Mead algorithm (Nelder and Mead, 1965) from four analogous starts. Thus, the comparison varies both search effort and optimization method.

For strategy m and class c , let $\hat{\ell}_m^{(c)}$ be the attained log-likelihood and $\hat{\ell}_\star^{(c)} = \max_m \hat{\ell}_m^{(c)}$. We regard strategy m as attaining the best available solution when $\sum_c \{\hat{\ell}_\star^{(c)} - \hat{\ell}_m^{(c)}\} \leq 10^{-4}$. Table F.2 reports this percentage, median TMSE, and mean time for a pair of class-specific fits, averaged over S1–S4.

Table F.2: Percentage of replicates attaining the best log-likelihood, median TMSE, and mean time cost for a pair of class-specific fits across 500 Monte Carlo replicates, averaged over configurations S1–S4. “Four” and “Fifty” denote the four-start and fifty-start strategies, respectively.

n	At best log-likelihood (%)			Median TMSE			Time (ms per pair)		
	Four	Fifty	Simplex	Four	Fifty	Simplex	Four	Fifty	Simplex
20	89.8	99.9	86.9	2.798	4.185	2.725	5.16	66.05	13.89
50	98.9	100.0	98.8	0.262	0.267	0.263	6.17	82.73	13.67
100	99.8	100.0	100.0	0.088	0.088	0.088	8.46	114.98	15.60
400	100.0	100.0	100.0	0.018	0.018	0.018	24.61	348.19	33.82

The four-start strategy attains the best available likelihood in 89.8% of replicates at $n = 20$ and in 98.9%–100% for $n \geq 50$. At $n = 20$, the fifty-

start search attains a higher likelihood more often but has a larger median TMSE, indicating that additional optimization can amplify small-sample estimation variability. For $n \geq 50$, the three strategies have essentially identical TMSE, while the fifty-start and simplex procedures require substantially more computational cost.

We next hold the four initial values fixed and vary one boundary of the search region at a time: $\delta_{\max} \in \{30, 500\}$, $\delta_{\min} = 0.2$, $w \in [0.01, 0.99]$, or $\sigma > 10^{-3}$. Each alternative is compared with the default region on the same class-0 samples. For each sample, we compute $\Delta_{\text{KL}} = \text{KL}(f_0 \parallel \hat{f}_{\text{region}}) - \text{KL}(f_0 \parallel \hat{f}_{\text{default}})$. Table F.3 reports the 95th percentile of $|\Delta_{\text{KL}}|$, which emphasizes the more sensitive fits.

Table F.3: Sensitivity of the fitted density to the search region, measured by the 95th percentile of $|\Delta_{\text{KL}}|$ and averaged over configurations S1–S4 in Table F.1

Search region	$n=20$	$n=50$	$n=100$	$n=400$
default (reference)	0	0	0	0
$\delta_{\max} = 30$	0.646	0.018	0.007	0
$\delta_{\max} = 500$	0.962	0.011	0.004	0
$\delta_{\min} = 0.2$	0	0	0	0
$w \in [0.01, 0.99]$	4.176	0	0	0
$\sigma > 10^{-3}$	0	0	0	0

The fitted density is insensitive to these changes except in the smallest samples. Altering the lower bounds on δ and σ has no discernible effect, and all entries are below 0.02 by $n = 50$ except for the relaxed δ ceiling, which is still only 0.011. The larger deviations at $n = 20$ arise mainly in the near-Gaussian setting, where δ is weakly identified, and when the wider interval for w permits unstable extreme fits. Their disappearance for $n \geq 50$ supports the default region and suggests that its mixing-weight bound provides mild small-sample regularization.

F.3 Statistical error

Now that we have justified the initialization strategy and search region of our estimation algorithm, we turn to inspecting the estimation quality and ask whether uncertainty in $\hat{\Psi}$, especially for the weakly identified δ , affects key quantities passed into the classifier S-RoLLR. We consider three levels of evaluation: parameter recovery for the identifiable components, estimation accuracy of the fitted density, and the accuracy of a transformed feature compared to its oracle counterpart with known δ .

For (τ, w, σ) , we regress the logarithm of empirical RMSE on $\log n$; the standard parametric benchmark is a slope of $-1/2$. For density estimation, we regress the logarithm of median $D_{\text{KL}}(f_0 \| \hat{f})$ on $\log n$, with benchmark

slope -1 . We treat δ separately because it can be weakly identified near Gaussianity. Table F.4 reports the fitted slopes.

Table F.4: Empirical convergence rates, from log-log regressions over the full grid of n .

Configuration	Parameter RMSE slope			Density D_{KL} slope
	τ	w	σ	
S1	-0.563	-0.602	-0.632	-1.246
S2	-0.623	-0.608	-0.548	-1.285
S3	-0.588	-0.605	-0.550	-1.245
S4	-0.633	-0.576	-0.565	-1.158
Theory	-0.500	-0.500	-0.500	-1.000

The empirical rates are close to their parametric benchmarks, with $R^2 \geq 0.98$ for parameter RMSE and $R^2 \in [0.96, 0.99]$ for density divergence. The tail parameter is recovered quickly in the heavy-tailed settings but remains weakly identified in the near-Gaussian setting. Nevertheless, the fitted density converges at a similar rate in all four configurations, showing that uncertainty in δ need not translate into density instability.

To isolate the cost of estimating δ , we compare the full MLE with an oracle that fixes δ at its true value. Let R_{KL} and R_{TMSE} be the ratios of the median divergence and median TMSE, respectively, for the full estima-

tor relative to this oracle. Under regular likelihood behavior, the median divergence ratio for estimating four rather than three parameters is

$$R^* = \frac{\text{median}(\chi_4^2)}{\text{median}(\chi_3^2)} \approx 1.4187.$$

Table F.5 reports the comparison at $n = 1000$, together with the interquartile range of $\hat{\delta}$.

Table F.5: Comparison with the δ -oracle at $n = 1000$

Configuration	R_{KL}	Benchmark	$\text{IQR}(\hat{\delta})$	R_{TMSE}
S1	1.290	1.419	77.0	1.357
S2	1.434	1.419	0.23	1.483
S3	1.479	1.419	1.06	1.368
S4	1.392	1.419	0.08	1.395

At $n = 1000$, both ratios are close to the benchmark 1.42 despite the wide dispersion of $\hat{\delta}$ in S1. Small samples can produce appreciably larger transformation error, especially in heavy-tailed settings, but this inflation declines rapidly with n : the divergence ratios are within 0.26 of R^* for $n \geq 100$, and the transformation ratios approach the same benchmark thereafter. Thus, weak identification of δ does not create persistent instability in the fitted density or transformed feature.

Together, Sections F.2 and F.3 support the four-start estimator: it is

computationally economical, insensitive to reasonable changes in the search region beyond the smallest samples, and statistically accurate as the sample size increases.

F.4 Selection of the penalty term λ_n

The second-stage penalty, denoted by λ in this section, is selected by cross-validation. The class-specific TPSC parameters are estimated once from the full training set, after which the transformed design is held fixed and cross-validation is applied only to the penalized logistic regression.

The candidate grid is generated by `glmnet`: at most 100 logarithmically spaced values from $\lambda_{\max}$, the smallest penalty yielding an empty model, to $\varepsilon\lambda_{\max}$, where $\varepsilon = 10^{-4}$ if $n > p$ and 10^{-2} otherwise. This data-dependent path adapts to the supplied design and may terminate early under the default deviance criterion.

For S-RoLLR, we specify `standardize = FALSE` so that the estimated log-likelihood ratios retain their natural evidential scale. Column-wise standardization would change the relative L_1 -penalty weights and therefore define a different penalized estimator. In contrast, because the raw features used in L1-RL have disparate units and scales, we retain the default standardization when implementing L1-RL.

We select `lambda.min`, the minimizer of cross-validated binomial deviance, and use it for both prediction and the active set $\hat{\mathcal{I}} = \{j : \hat{\beta}_j \neq 0\}$. The one-standard-error rule is not used because it deliberately adds regularization beyond the deviance-minimizing choice.

Table F.6 summarizes the number of cross-validation folds used across all experiments. Two methodological details warrant clarification. First, L1-RL is consistently tuned using the ten-fold default in `cv.glmnet`, whereas S-RoLLR employs five folds. If this discrepancy has any substantive impact, it would inherently favor L1-RL due to the larger effective training size per fold; thus, it cannot explain the empirical performance gains reported in Sections 4 and 5 of the main article. Second, in the selection-frequency study of Section 5.1, we reduce the number of folds to three for S-RoLLR: each training sample contains only approximately 71 observations, under which a five-fold partition would yield merely around 14 samples per validation fold.

In Appendix G, fold assignments and train–test splits are shared across working distributions within each replicate. Holding the first-stage transformation fixed makes cross-validated deviance mildly optimistic as an absolute risk estimate, but the same transformation is used for every candidate λ . The final test set remains strictly isolated from both stages, so the

Table F.6: Number of cross-validation folds used to select λ

Experiment	S-RoLLR	L1-RL
Section 4, twelve cases	5	10
Section 5, overall comparison of three case studies	5	10
Section 5.1, selection frequencies	3	10
Appendix G, all S-RoLLR($\mathcal{D}$) variants	5	5

reported performance comparisons are unaffected.

Appendix G: Simulation Study for S-RoLLR($\mathcal{D}$)

This appendix provides a comprehensive report of the working-distribution experiments summarized in Section 6.1 of the main article. Across all model variants, only the first-stage specification is varied. Within each replicate, the second-stage estimator, train–test partition, and cross-validation fold assignments are held identical, ensuring that any observed performance discrepancy is strictly attributable to the choice of working distribution.

G.1: Design of the experiments

The continuous-feature comparison, described in Section 6.1, includes L1-RL and six S-RoLLR variants based on Gaussian, Student- t , skew-normal, skew- t (Azzalini and Capitanio, 2003), TPSC, and finite location-scale mix-

ture working models. For discrete features we compare Poisson, negative binomial, empirical probability mass, and TPSC specifications. Table G.2 summarizes their estimation procedures.

Table G.1 gives the data-generating mechanisms. In every setting, $n = 800$, $p = 100$, $|\mathcal{I}| = 20$, and 70% of observations are used for training in each of 300 Monte Carlo replicates. Noise features have the class-0 marginal distribution in both classes. Settings C6 and D5 misspecify every candidate family; C4, C5, and D4 match low-order class moments so that discrimination depends on distributional shape.

The mixed settings contain equal-sized continuous and discrete blocks, each with ten informative features. Continuous features follow C3. A proportion $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$ of the discrete features follows D4, and the remainder follows Poisson(4) in Class 0 and Poisson($4e^{0.45}$) in Class 1. Thus only the composition of the discrete block changes across E1–E5.

To compare settings of different difficulty, let $\overline{\text{AUC}}_{c,a}$ be the mean AUC of variant a in setting c and define $\text{Gap}_{c,a} = \max_{a'} \overline{\text{AUC}}_{c,a'} - \overline{\text{AUC}}_{c,a}$. The maximum over c is the variant’s worst-case AUC shortfall. Table 4 of the main article also reports median runtime, while the discrete and mixed result tables below report the largest false positive rate (FPR).

Table G.1: Generative models of $\mathbf{X}_{\mathcal{T}}$ in the extension experiments

Setting	$X_j Y = 0$	$X_j Y = 1$	Correctly specified
<i>Continuous features</i>			
C1	$\mathcal{N}(0, 1)$	$\mathcal{N}(0.55, 1)$	L1-RL
C2 [†]	$\text{SN}(0, 1.5, 5)$	$\text{SN}(0.4, 1.5, 5)$	skew-normal
C3 [†]	$\text{ST}(0, 1.5, 3, 3)$	$\text{ST}(0.75, 1.5, 3, 3)$	skew- t
C4 [‡]	$-\{Z - E(Z)\}$	$Z - E(Z)$	TPSC
C5 [*]	M	$-M$	mixture normal
C6 [•]	$0.9\mathcal{N}(0, 1) \oplus 0.1\mathcal{N}(0, 100)$	$0.55 + \{0.9\mathcal{N}(0, 1) \oplus 0.1\mathcal{N}(0, 100)\}$	none
<i>Discrete features</i>			
D1	$\text{Poisson}(5)$	$\text{Poisson}(5e^{0.3})$	Poisson
D2 [◊]	$\text{NB}(0.5, 6)$	$\text{NB}(0.5, 6e^{0.55})$	negative binomial
D3	$\min\{4, \text{Poisson}(1.2)\}$	$\min\{4, \text{Poisson}(2.2)\}$	empirical pmf
D4 [‡]	$\text{Unif}\{0, \dots, 5\}$	$\text{Multinomial}(1, \mathbf{q})$	empirical pmf
D5 [◊]	$C\{\text{Poisson}(4)\}$	$C\{\text{Poisson}(4e^{0.4})\}$	none

[†] $\text{SN}(\xi, \omega, \alpha)$ and $\text{ST}(\xi, \omega, \alpha, \nu)$ denote the skew-normal and the skew- t distributions with location ξ , scale ω , slant α , and ν degrees of freedom.

[‡] $Z \sim \text{TPSC}(\Psi)$ with $\Psi = (0, 0.3, 1.5, 4)^\top$ in (2.1) of the main article, as in Case 10 of Section 4 of the main article.

^{*} $M = 0.62\mathcal{N}(-1.3, 0.55^2) + 0.38\mathcal{N}(2.12, 0.55^2)$, so that $E(M) = 0$ and the two classes are mirror images.

[•] $\oplus$ denotes a two-component mixture in which exactly 10% of the observations of each class are drawn from the scale-inflated component.

[◊] $\text{NB}(a, b)$ denotes the negative binomial distribution with size a and mean b .

[‡] $\mathbf{q} = (0.220, 0.007, 0.273, 0.273, 0.007, 0.220)^\top$, chosen so that the class mean and the class variance coincide with those of $\text{Unif}\{0, \dots, 5\}$.

[◊] $C\{\cdot\}$ denotes contamination under which 5% of the counts are multiplied by ten.

Table G.2: Estimation procedure for each working distribution in Appendix G

Working distribution	Estimation method
Gaussian $N(\mu, \sigma^2)$	class-specific sample mean $\hat{\mu} = \bar{x}$ and sample standard deviation $\hat{\sigma}$
Student- t	maximum likelihood by <code>optim</code>
Skew-normal	maximum likelihood by the <code>sn</code> package
Skew- t	maximum likelihood by the <code>sn</code> package
Mixture	expectation-maximization, with the number of components and the degrees of freedom selected by BIC
Poisson(μ)	$\hat{\mu} = \bar{x}$, the closed-form maximum likelihood estimator
Negative binomial	method of moments, $\hat{\mu} = \bar{x}$ and $\hat{r} = \bar{x}^2/(s^2 - \bar{x})$; defaults to Poisson when $s^2 \leq \bar{x}$
Empirical pmf	Laplace-smoothed cell probabilities on the support pooled across the two training classes, $\hat{p}_k = (n_k + 0.5)/(n + 0.5K)$ for $k = 1, \dots, K$; an unseen test value is assigned zero log-likelihood ratio
Bernoulli	$\hat{p} = (n_1 + 1)/(n + 2)$, the Laplace-smoothed proportion of ones; used by the adaptive rule only for two-valued features, which do not arise in the settings of Table G.1

G.2: Working distribution for continuous features

Table 4 of the main article shows that TPSC, skew- t , and the mixture remain close to the best classifier across all six continuous settings, whereas models accommodating only skewness or only heavy tails can fail when the omitted feature is essential. Because predictive accuracy alone does not clearly separate the three competitive variants, we compare the stability of their estimated transformations.

For a representative informative feature, let $\widehat{T}^{(r)}(x) = \log \widehat{f}_1^{(r)}(x) - \log \widehat{f}_0^{(r)}(x)$ in replicate r . On a common grid $x_1, \dots, x_G$ of $G = 25$ points, define $S_{c,a}(x_g) = \text{SD}_r\{\widehat{T}^{(r)}(x_g)\}$ and $\bar{S}_{c,a} = G^{-1} \sum_g S_{c,a}(x_g)$ for setting c and variant a . Table G.3 reports the maximum and mean of $\bar{S}_{c,a}$ across settings. TPSC has the lowest sampling variability by both summaries.

Table G.3: Sampling variability of the estimated transformed features for the three competitive continuous-feature variants, over 300 Monte Carlo replicates

Working distribution	$\max_{c \in \mathcal{C}} \bar{S}_{c,a}$	$ \mathcal{C} ^{-1} \sum_{c \in \mathcal{C}} \bar{S}_{c,a}$
TPSC	0.373	0.282
Skew- t	0.730	0.388
Mixture	0.590	0.366

Re-evaluating C3 at $n \in \{80, 150, 300, 800\}$ on an independent test set

of 3,000 observations confirms that this stability matters primarily in small samples. The paired AUC advantage of TPSC over skew- t decreases from 0.061 at $n = 80$ to 0.015, 0.001, and 5.6×10^{-5} as n increases. Together with the runtimes in Table 4 of the main article, these results support TPSC as a stable and computationally efficient default. Correctly specified simpler or more specialized families may retain a modest advantage when reliable prior information is available.

G.3: Discrete and mixed features

No single working family is uniformly reliable for discrete predictors. We therefore use a feature-specific adaptive rule: binary features are assigned Bernoulli models, features with at most ten observed levels use empirical probability mass functions, non-negative integer features with larger supports use Poisson models, and all remaining features use TPSC.

Table G.4 illustrates the trade-off motivating this rule. The empirical mass function captures the non-monotone class difference in D4 and the contamination in D5, but its high variance produces an FPR of 0.813 in D2. Poisson fitting is more stable, with a largest FPR of 0.123, but cannot represent non-monotone differences. As explained in Section 6.1, its log-likelihood ratio is affine in the count, so its predictor space is the

same as that of L1-RL up to the feature-wise weighting induced by the L_1 penalty. The adaptive rule combines flexible modeling for small supports with parametric stabilization for larger count supports.

Table G.5 compares the adaptive rule with L1-RL, uniform TPSC fitting, and a fully discretized variant using eight training-sample quantile bins. The adaptive rule is never worse in AUC, has the highest F_1 score in every setting, and limits the largest FPR to 0.045. Its AUC advantage over uniform TPSC grows from 0.0005 at $\rho = 0$ to 0.0634 at $\rho = 1$, showing that feature-specific modeling becomes more valuable as non-monotone discrete structure increases. Universal binning remains competitive in AUC but loses continuous information and is consistently inferior in feature selection.

Table G.4: Average classification AUC and average feature selection F_1 score across 300 Monte Carlo replicates for L1-RL and four variants of S-RoLLR in the five discrete settings

	L1-RL	Poisson	negative binomial	empirical pmf	TPSC
<i>AUC for classification (with the highest in each setting highlighted in boldface)</i>					
D1	0.984	0.985	0.985	0.965	0.983
D2	0.864	0.864	0.841	0.592	0.700
D3	0.989	0.990	0.990	0.987	0.987
D4	0.507	0.524	0.525	0.998	0.519
D5	0.779	0.779	0.836	0.986	0.980
<i>F_1 score for feature selection (with the highest in each setting highlighted in boldface)</i>					
D1	0.690	0.966	0.948	0.565	0.856
D2	0.724	0.837	0.685	0.371	0.547
D3	0.697	0.969	0.967	0.861	0.982
D4	0.025	0.133	0.182	0.974	0.101
D5	0.698	0.772	0.612	0.667	0.774
<i>Summary across the five settings (with the lowest value highlighted in italics)</i>					
Maximum AUC shortfall	0.492	0.475	0.473	<i>0.273</i>	0.480
Largest FPR	0.234	<i>0.123</i>	0.311	0.813	0.373

Table G.5: Average classification AUC and average feature selection F_1 score across 300 Monte Carlo replicates for L1-RL and three variants of S-RoLLR in five mixed-data settings, where ρ is the proportion of discrete features generated by the non-monotone mechanism of setting D4.

	L1-RL	TPSC throughout	binned	adaptive
<i>AUC for classification (with the highest in each setting highlighted in boldface)</i>				
E1 ($\rho = 0$)	0.988	0.995	0.993	0.995
E2 ($\rho = 0.25$)	0.980	0.992	0.991	0.995
E3 ($\rho = 0.5$)	0.958	0.983	0.989	0.995
E4 ($\rho = 0.75$)	0.893	0.961	0.984	0.995
E5 ($\rho = 1$)	0.810	0.931	0.981	0.995
<i>F_1 score for feature selection (with the highest in each setting highlighted in boldface)</i>				
E1 ($\rho = 0$)	0.678	0.933	0.858	0.952
E2 ($\rho = 0.25$)	0.650	0.849	0.851	0.933
E3 ($\rho = 0.5$)	0.609	0.722	0.807	0.924
E4 ($\rho = 0.75$)	0.554	0.607	0.766	0.919
E5 ($\rho = 1$)	0.513	0.527	0.751	0.915
<i>Summary across the five settings (with the lowest value highlighted in italics)</i>				
Maximum AUC shortfall	0.185	0.063	0.013	<i>0.000</i>
Largest FPR	0.206	0.241	0.175	<i>0.045</i>

These experiments support the practical guidance in Section 6: TPSC is a competitive default for continuous features, while feature-specific working distributions are preferable for discrete and mixed predictors. In both cases the second-stage estimator is unchanged.

References

- Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B* 65(2), 367–389.
- Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing* 16(5), 1190–1208.
- Cover, T. and J. Thomas (2006). *Elements of Information Theory*. Wiley.
- Devroye, L., L. Györfi, and G. Lugosi (2013). *A Probabilistic Theory of Pattern Recognition*. Stochastic Modelling and Applied Probability. Springer New York.
- Hartigan, J. A. and P. M. Hartigan (1985). The dip test of unimodality. *The annals of Statistics*, 70–84.
- Horn, R. A. and C. R. Johnson (2013). *Matrix Analysis*. Cambridge University Press.
- Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics* 40(2), 633–643.
- Liu, Q., X. Huang, and R. Bai (2024). Bayesian modal regression based on mixture distributions.

REFERENCES

- Computational Statistics & Data Analysis* 199, 108012.
- McKay, M. D., R. J. Beckman, and W. J. Conover (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 42(1), 55–61.
- Nelder, J. A. and R. Mead (1965). A simplex method for function minimization. *The computer journal* 7(4), 308–313.
- Park, M. Y. and T. Hastie (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 69(4), 659–677.
- Resnick, S. I. (2007). *Heavy-tail phenomena: Probabilistic and Statistical Modeling*. Springer.
- van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 1–25.