Supplement to "Statistical Inference for High-Dimensional Generalized Linear Models with Binary Outcomes"

T. Tony Cai[1], Zijian Guo[2] and Rong Ma[3]
University of Pennsylvania[1]
Rutgers University[2]
Stanford University[3]

Contents

1 Proofs of Main Theorems

1.1 Proof of Theorem 1

To prove Theorem 1, we assume for now that Theorem 2 holds. By the definition of $\tilde{\beta}_j$, we have

$$\tilde{\beta}_j - \beta_j = -\left(\frac{1}{n}\sum_{i=1}^n w(X_i^\top\hat{\beta})f'(X_i^\top\hat{\beta})\hat{u}^\top X_i X_i^\top - e_j^\top\right)(\hat{\beta}-\beta) + \frac{1}{n}\sum_{i=1}^n w(X_i^\top\hat{\beta})\hat{u}^\top X_i\epsilon_i - \frac{1}{n}\sum_{i=1}^n \Delta_i w(X_i^\top\hat{\beta})\hat{u}^\top X_i, \tag{1.1}$$

where we denote $A_n = \frac{1}{n}\sum_{i=1}^n w(X_i^\top\hat{\beta})\hat{u}^\top X_i\epsilon_i$ and $B_n = -\big(\frac{1}{n}\sum_{i=1}^n w(X_i^\top\hat{\beta})f'(X_i^\top\hat{\beta})\hat{u}^\top X_i X_i^\top - e_j^\top\big)(\hat{\beta}-\beta) - \frac{1}{n}\sum_{i=1}^n \Delta_i w(X_i^\top\hat{\beta})\hat{u}^\top X_i$. In what follows, on the one hand, we show that, in equation (1.1), under the conditions of Theorem 1, with probability at least $1-p^{-c}-n^{-c}$,

$$\left\|\frac{1}{n}\hat{u}^\top\sum_{i=1}^n w(X_i^\top\hat{\beta})f'(X_i^\top\hat{\beta})X_i X_i^\top - e_j^\top\right\|_\infty \lesssim \lambda_n, \tag{1.2}$$

$$\left|\frac{1}{n}\sum_{i=1}^n \Delta_i w(X_i^\top\hat{\beta})\hat{u}^\top X_i\right| \lesssim \tau_n\frac{k\log p}{n}. \tag{1.3}$$

Combining event $\mathcal{B}_1$ in Theorem 2 and (1.2), we have with probability at least $1-p^{-c}-n^{-c}$,

$$\left|\left(\frac{1}{n}\sum_{i=1}^n w(X_i^\top\hat{\beta})f'(X_i^\top\hat{\beta})\hat{u}^\top X_i X_i^\top - e_j^\top\right)(\hat{\beta}-\beta)\right| \le \lambda_n\|\hat{\beta}-\beta\|_1 \lesssim \frac{k\log p}{n}, \tag{1.4}$$

which along with (1.3) leads to the upper bound of B_n. On the other hand, the asymptotic normality of the stochastic term A_n in (1.1) is obtained by the following proposition.

Proposition 2. *Under the assumption of Theorem 1, conditional on $\mathcal{D}_2$ and $\{X_i\}_{i=1}^n$, it holds that*

$$v_j^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w(X_i^\top\hat{\beta})\hat{u}^\top X_i\epsilon_i \to_d N(0,1). \tag{1.5}$$

The rest of the proofs are devoted to (1.2) (1.3) and Proposition 2.

Proof of (1.2). This inequality follows from the definition of $\hat{u}$. To show that such $\hat{u}$ exists with high probability, we show that, conditional on $\mathcal{D}_2$ (i) the matrix

$$\mathbf{\Gamma} = \mathbb{E}[w(X_i^\top\hat{\beta})f'(X_i^\top\hat{\beta})X_iX_i^\top] = \mathbb{E}\bigg[\frac{[f'(X_i^\top\hat{\beta})]^2}{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}X_iX_i^\top\bigg]$$

is invertible and (ii) the j-th column u^* of $\mathbf{\Gamma}^{-1}$ is feasible for the constraints

$$\bigg\|\frac{1}{n}u^\top\sum_{i=1}^{n} w(X_i^\top\hat{\beta})f'(X_i^\top\hat{\beta})X_iX_i^\top - e_j^\top\bigg\|_\infty \lesssim \lambda_n, \qquad \max_{1\le i\le n}|X_i^\top u| \le \tau_n,$$

with probability at least $1-p^{-c}-n^{-c}$. The final result follows by averaging over the randomness of $\mathcal{D}_2$.

Proof of (i). To show statement (i), it suffices to show that, for some $\delta > 0$,

$$w^\top\mathbf{\Gamma}w = \mathbb{E}\bigg[\frac{[f'(X_i^\top\hat{\beta})]^2}{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}(X_i^\top w)^2\bigg] > \delta, \tag{1.6}$$

uniformly for any unit vector $w\in\mathbb{R}^p$. To simplify notation, we denote $h(x)=\frac{[f'(x)]^2}{f(x)(1-f(x))}$. We first show that, for sufficiently large (n,p), there exists some constant $\delta>0$ such that

$$\mathbb{E}|h(X_i^\top\hat{\beta})-h(X_i^\top\beta)|(w^\top X_i)^2 < \delta. \tag{1.7}$$

To see this, note that for any $x\neq a$,

$$h(x) = h(a)+h'(a+t(x-a))(x-a), \tag{1.8}$$

for some $t\in(0,1)$. In addition, $|h'(x)|=O(1)$ for all $x\in\mathbb{R}$, since

$$h'(x) = \frac{2f'(x)f''(x)}{f(x)(1-f(x))} - \frac{[f'(x)]^2f'(x)(1-2f(x))}{f^2(x)(1-f(x))^2},$$

where by (L2) and (L3), we have $0<\frac{2f'(x)f''(x)}{f(x)(1-f(x))}=O(1)$, and by (L2), we have

$$\bigg|\frac{[f'(x)]^3(1-2f(x))}{f^2(x)(1-f(x))^2}\bigg| \le \frac{[f'(x)]^3}{f^2(x)(1-f(x))^2} \lesssim \frac{[f'(x)]^3}{(1-f(|x|))^2} \lesssim x^2f'(x) = O(1).$$

Thus by (1.8), with probability at least $1 - p^{-c}$,

$$|\mathbb{E}[h(X_i^\top \hat{\beta}) - h(X_i^\top \beta)](w^\top X_i)^2| < \sqrt{\mathbb{E}[X_i^\top(\hat{\beta} - \beta)]^2}\sqrt{\mathbb{E}(w^\top X_i)^4} \leq C\sqrt{\mathbb{E}|X_i^\top(\beta - \hat{\beta})|^2}$$
$$\lesssim \sqrt{(\beta - \hat{\beta})\mathbf{\Sigma}(\beta - \hat{\beta})} \lesssim \frac{k \log p}{n},$$

where the second last inequality follows from (A) and the last inequality follows from Theorem 2 and (A). This implies (1.7). With this, (1.6) can be proven by showing that $\mathbb{E}h(X_i^\top \beta)(w^\top X_i)^2 > 3\delta$ for all unit vector w. To see this, note that, for any constant $T > 0$, $\mathbb{E}h(X_i^\top \beta)(w^\top X_i)^2 \geq \mathbb{E}h(X_i^\top \beta)(w^\top X_i)^2 1\{|X_i^\top \beta| < T\}$. Under the constraint $|X_i^\top \beta| < T$, by concavity of the link function f over $\mathbb{R}_+$ and its symmetry (L1), we have $f'(X_i^\top \beta) \geq f'(T) > 0$. Along with the fact that $f(X_i^\top \beta)(1 - f(X_i^\top \beta)) \leq 1/4$, we have $h(X_i^\top \beta) \geq C$ for all $|X_i^\top \beta| < T$. Thus $\mathbb{E}h(X_i^\top \beta)(w^\top X_i)^2 \geq C\mathbb{E}(w^\top X_i)^2 1\{|X_i^\top \beta| < T\}$. Now since $|\mathbb{E}X_i^\top w| = O(1)$, since by (A) $\mathbb{E}(w^\top X_i)^2 \geq c_0 > 0$, it suffices to show that

$$\mathbb{E}[(w^\top X_i)^2 - (w^\top X_i)^2 1\{|X_i^\top \beta| < T\}] \leq c_0/3. \tag{1.9}$$

Note that

$$\mathbb{E}[(w^\top X_i)^2 - (w^\top X_i)^2 1\{|X_i^\top \beta| < T\}] \leq \mathbb{E}[(w^\top X_i)^2 \cdot 1\{|X_i^\top \beta| > T\}]$$
$$\leq \sqrt{\mathbb{E}[(w^\top X_i)^4} P^{1/2}(|X_i^\top \beta| > T) \leq c_1 \exp(-c_2 T^2),$$

where we can choose T sufficiently large so that $c_1 \exp(-c_2 T^2) \leq c_0/3$. Thus it holds that $\mathbb{E}h(X_i^\top \beta)(w^\top X_i)^2 \geq Cc_0/3$, which is lower bounded by 3δ if we choose $\delta < Cc_0/9$.

Proof of (ii). Let u^* be j-th column of $\mathbf{\Gamma}^{-1}$ so that $(u^*)^\top \mathbf{\Gamma} = e_j$. Note that, for any $k \in [1 : p]$, conditional on $\hat{\beta}$, the random variables $h(X_i^\top \hat{\beta})(u^*)^\top X_i X_i^\top e_k$, $i = 1, ..., n$, are independent with mean δ_{jk}, where $\delta_{jk} = 1$ if $j = k$ and otherwise $\delta_{jk} = 0$. By conditions (L1) and (L2)

$$\lim_{x\to\infty} \frac{[f'(x)]^2}{f(x)(1 - f(x))} = \lim_{x\to-\infty} \frac{[f'(x)]^2}{f(x)(1 - f(x))} = \lim_{x\to-\infty} \frac{[f'(x)]^2}{f(x)} \leq C \lim_{x\to\infty} x f'(x) = O(1), \tag{1.10}$$

it follows that $h(x)$ is bounded on $\mathbb{R}$, which along with condition (A) implies that, conditional on $\mathcal{D}_2$, $h(X_i^\top \hat{\beta})(u^*)^\top X_i X_i^\top e_k$, $i = 1, ..., n$, are independent subexponential random variables. By applying the concentration inequality for subexponential random variables (see, for example, Proposition 5.16 of Vershynin (2010)) for given $\mathcal{D}_2$, after taking account of the randomness of $\mathcal{D}_2$, we have with probability at least $1 - p^{-c}$,

$$\left\| \frac{1}{n} \sum_{i=1}^n h(X_i^\top \hat{\beta})(u^*)^\top X_i X_i^\top - e_j^\top \right\|_\infty \leq \lambda_n, \tag{1.11}$$

where $\lambda_n \asymp \sqrt{\log p/n}$. On the other hand, since by (1.6), we know that $\|\mathbf{\Gamma}^{-1}\| = O(1)$, it follows that $\|u^*\|_2 = O(1)$. Hence, by (A), with probability at least $1 - n^{-c}$, $\|Xu^*\|_\infty = \max_{1\leq i\leq n} |X_i^\top u^*| \leq \tau_n$. Therefore, we have proven statement (ii) and hence (1.2).

Proof of (1.3). By Cauchy-Schwartz inequality, we have $\big|\frac{1}{n}\sum_{i=1}^n \Delta_i w(X_i^\top\hat\beta)\hat u^\top X_i\big| \le \max_{1\le i\le n}|\hat u^\top X_i|\cdot\frac{1}{n}\sum_{i=1}^n |w(X_i^\top\hat\beta)\Delta_i|$. By definition of $\hat u$, we have $\max_{1\le i\le n}|\hat u^\top X_i|\le\tau_n$. Now by Theorem 2 and (A), we have for $w'=(\beta-\hat\beta)/\|\beta-\hat\beta\|$, with probability at least $1-p^{-c}-n^{-c}$,

$$\max_{1\le i\le n}|X_i^\top(\beta-\hat\beta)|\le\max_{1\le i\le n}|X_i^\top w'|\cdot\|\beta-\hat\beta\|_2\le\tau_n\sqrt{\frac{k\log p}{n}}\le C,$$

which holds whenever $k\lesssim\frac{n}{\tau_n^2\log p}$. By (L3), with probability at least $1-p^{-c}-n^{-c}$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n|w(X_i^\top\hat\beta)\Delta_i|\le\frac{1}{n}\sum_{i=1}^n\frac{f'(X_i^\top\hat\beta)f''(X_i^\top\hat\beta+tX_i^\top(\beta-\hat\beta)}{f(X_i^\top\hat\beta)(1-f(X_i^\top\hat\beta))}[X_i^\top(\hat\beta-\beta)]^2\\&\lesssim\frac{1}{n}\sum_{i=1}^n[X_i^\top(\hat\beta-\beta)]^2\lesssim\frac{k\log p}{n},\end{aligned}$$

where the last inequality follows from Theorem 2. This completes the proof of (1.3).

Proof of Proposition 2. We need the following lemma.

Lemma 3. *Under the conditions of Theorem 1, with probability at least $1-p^{-c}$, it holds that $\frac{0.99^2}{\lambda_{\max}(\mathbf{\Gamma})}\le v_j\le(1+o(1))\mathbf{\Gamma}_{jj}^{-1}$. The boundedness of $\lambda_{\max}(\mathbf{\Gamma})$ follows from the proof of (1.2). Moreover, with probability at least $1-p^{-c}-n^{-c}$, we have $v_j\le(1+o(1))(\mathbf{\Gamma}^*)_{jj}^{-1}$ where $\mathbf{\Gamma}^*$ is defined the same as $\mathbf{\Gamma}$ except for $\hat\beta$ replaced by β.*

For simplicity, we omit the subscript j in v_j. Define $W_i=v^{-1/2}w(X_i^\top\hat\beta)\hat u^\top X_i\epsilon_i$. Conditional on $\mathcal{D}_2$ and $X=\{X_i\}_{i=1}^n$, $\{W_i\}_{i=1}^n$ are independent random variables with $\mathbb{E}(W_i|X,\mathcal{D}_2)=0$ and $\sum_{i=1}^n\mathrm{var}(W_i|X,\mathcal{D}_2)=n$. To establish the asymptotic normality, it suffices to check the Lindeberg's condition, that is, for any constant $\bar\epsilon>0$,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\big(W_i^2 1\{|W_i|\ge\bar\epsilon\sqrt{n}\}\big)=0.\tag{1.12}$$

Note that, given Lemma 3,

$$\max_{1\le i\le n}|W_i|=\max_{1\le i\le n}\left|\frac{f'(X_i^\top\hat\beta)\hat u^\top X_i\epsilon_i}{v^{1/2}f(X_i^\top\hat\beta)(1-f(X_i^\top\hat\beta))}\right|\lesssim\max_{1\le i\le n}\left|\frac{f'(|X_i^\top\hat\beta|)}{1-f(|X_i^\top\hat\beta|)}\hat u^\top X_i\right|$$

By (L2), under the event $\mathcal{B}$ defined in Theorem 2, if we denote $w'=\frac{\beta-\hat\beta}{\|\beta-\hat\beta\|_2}$, then

$$\begin{aligned}&\frac{f'(|X_i^\top\hat\beta|)}{1-f(|X_i^\top\hat\beta|)}\le|X_i^\top\hat\beta|\le C_1|X_i^\top\beta|+C_2|X_i^\top(\beta-\hat\beta)|\le C_1|X_i^\top\beta|+C_2|X_i^\top w'|\cdot\|\beta-\hat\beta\|_2\\&\le C_1|X_i^\top\beta|+C_2|X_i^\top w'|\left(\frac{k\log p}{n}\right)\le C(|X_i^\top\beta|+|X_i^\top w'|)\end{aligned}\tag{1.13}$$

Then, $\max_{1\le i\le n}\left|\frac{f'(|X_i^\top\hat\beta|)}{1-f(|X_i^\top\hat\beta|)}\hat u^\top X_i\right| \lesssim \max_{1\le i\le n}(|X_i^\top w'|+|X_i^\top\beta|)|X_i^\top\hat u| \lesssim \tau_n^2 \lesssim \sqrt{n}$ with probability at least $1-n^{-c}$, and (1.5) follows from (1.12) and the Lindeberg's central limit theorem.

1.2 Proof of Theorem 2

This theorem is a direct consequence of the following proposition concerning the relationships between some key quantities associated to the GLM Lasso problem.

Proposition 3. *Let $\hat\beta = \arg\min_\beta\left\{\ell(\beta)+\lambda\|\beta\|_1\right\}$ be the Lasso estimator for some GLM with true regression coefficient β^*, where the normalized negative log-likelihood $\ell(\beta)$ is a convex function. Let*

$$F(\xi,S;\psi,\psi_0) = \inf_{b\in\mathcal{C}(\xi,S),\psi_0(b)\le 1}\frac{\langle b,\ell'(\beta^*+b)-\ell'(\beta^*)\rangle e^{-\psi_0^2(b)-\psi_0(b)}}{\|b_S\|_1\psi(b)}, \tag{1.14}$$

where

$$\mathcal{C}(\xi,S) = \{b\in\mathbb{R}^p : \|b_{S^c}\|_1\le\xi\|b_S\|_1\ne 0\}, \tag{1.15}$$

$S=\{j:\beta_j^*\ne 0\}$, ψ and ψ_0 are semi-norms and $M_2>0$ is a constant. Define the event

$$\Omega = \left\{\frac{\lambda+z^*}{(\lambda-z^*)_+}\le\xi, \frac{\lambda+z^*}{F(\xi,S;\psi,\psi_0)}\le\eta e^{-\eta^2-\eta}\right\}$$

for some $\eta\le 1/2$, $z^*=\|\ell'(\beta^*)\|_\infty\le\lambda$. Then, in the event Ω, we have

$$\psi(\hat\beta-\beta^*)\le\frac{(\lambda+z^*)e^{\eta^2+\eta}}{F(\xi,S;\psi,\psi_0)}. \tag{1.16}$$

Event $\mathcal{B}_2$. From Proposition 3, it suffices to show that, (i) for $\lambda_0\asymp\sqrt{\log p/n}$, $z^*\le\lambda_0$ with probability at least $1-p^{-c}$; (ii) for some constant $c_0>0$, $F(\xi,S;\psi,\psi_0)\ge c_0/\sqrt{k}$ with probability at least $1-p^{-c}$; and (iii) the normalized negative log-likelihood $\ell(\beta)$ is a convex function with probability at least $1-p^{-c}$. Then we can choose $\lambda=2\lambda_0$ and set ψ to be the ℓ_2 norm to obtain the results concerning $\mathcal{B}_2$.

To show (i), by definition, we have

$$\ell'(\beta^*) = -\frac{1}{n}\sum_{i=1}^n\left[\frac{y_i f'(X_i^\top\beta^*)}{f(X_i^\top\beta^*)}-\frac{(1-y_i)f'(X_i^\top\beta^*)}{1-f(X_i^\top\beta^*)}\right]X_i.$$

Then

$$\|\ell'(\beta^*)\|_\infty = \max_{1\le j\le p}\left|\frac{1}{n}\sum_{i=1}^n\left[\frac{y_i f'(X_i^\top\beta^*)}{f(X_i^\top\beta^*)}-\frac{(1-y_i)f'(X_i^\top\beta^*)}{1-f(X_i^\top\beta^*)}\right]X_{ij}\right|.$$

Note that the individual components $\big[\frac{y_i f'(X_i^\top\beta^*)}{f(X_i^\top\beta^*)} - \frac{(1-y_i)f'(X_i^\top\beta^*)}{1-f(X_i^\top\beta^*)}\big]X_{ij}$ satisfies

$$\mathbb{E}\big[\frac{y_i f'(X_i^\top\beta^*)}{f(X_i^\top\beta^*)} - \frac{(1-y_i)f'(X_i^\top\beta^*)}{1-f(X_i^\top\beta^*)}\big]X_{ij} = 0,$$

and by condition (L2),

$$\big[\frac{y_i f'(X_i^\top\beta^*)}{f(X_i^\top\beta^*)} - \frac{(1-y_i)f'(X_i^\top\beta^*)}{1-f(X_i^\top\beta^*)}\big]X_{ij} \le \max\left\{\frac{f'(X_i^\top\beta^*)}{f(X_i^\top\beta^*)}, \frac{f'(X_i^\top\beta^*)}{1-f(X_i^\top\beta^*)}\right\}|X_{ij}| \le |X_{ij}X_i^\top\beta^*|,$$

which means the individual components are independent centred sub-exponential random variables. Therefore, by concentration inequality for subexponential random variable (see, for example, Proposition 5.16 of Vershynin (2010)), we have,

$$P\left(\|\ell'(\beta^*)\|_\infty \le C\|\beta^*\|_2\sqrt{\frac{\log p}{n}}\right) \ge 1 - p^{-c}, \tag{1.17}$$

which implies (i).

To show (ii), note that by Taylor expansion

$$\langle b, \ell'(\beta^* + b) - \ell'(\beta^*)\rangle = \int_0^1 \langle b, \ell''(\beta^* + tb)b\rangle dt,$$

with

$$\ell''(\beta^* + tb) = -\frac{1}{n}\sum_{i=1}^n \left[\frac{y_i(f_t'' f_t - f_t'^2)}{f_t^2} - \frac{(1-y_i)(f_t''(1-f_t) + f_t'^2)}{(1-f_t)^2}\right]X_i X_i^\top,$$

where we denoted $f_t \equiv f(X_i^\top(\beta^* + tb))$, $f_t' \equiv f'(X_i^\top(\beta^* + tb))$ and $f_t'' \equiv f''(X_i^\top(\beta^* + tb))$ for simplicity. Thus

$$\langle b, \ell'(\beta^* + b) - \ell'(\beta^*)\rangle = \int_0^1 \frac{1}{n}\sum_{i=1}^n \left[\frac{(1-y_i)(f_t''(1-f_t) + f_t'^2)}{(1-f_t)^2} - \frac{y(f_t'' f_t - f_t'^2)}{f_t^2}\right](b^\top X_i)^2 dt,$$

and by setting $\psi(b) = \|b\|$,

$$\begin{aligned} F(\xi, S; \psi) &= \inf_{b\in\mathcal{C}(\xi,S),\psi_0(b)\le 1} \frac{\langle b, \ell'(\beta^* + b) - \ell'(\beta^*)\rangle e^{\psi_0^2(b)+\psi_0(b)}}{\|b_S\|_1\psi(b)} \\ &= \inf_{b\in\mathcal{C}(\xi,S),\psi_0(b)\le 1} \frac{1}{n}\sum_{i=1}^n \int_0^1 \left[\frac{(1-y_i)(f_t''(1-f_t) + f_t'^2)}{(1-f_t)^2} - \frac{y_i(f_t'' f_t - f_t'^2)}{f_t^2}\right]dt \frac{e^{\psi_0^2(b)+\psi_0(b)}(b^\top X_i)^2}{\|b_S\|_1\|b\|_2}. \end{aligned}$$

Denote $h_t = \frac{(1-y_i)(f_t''(1-f_t)+f_t'^2)}{(1-f_t)^2} - \frac{y_i(f_t'' f_t - f_t'^2)}{f_t^2}$. By condition (L4), we have

$$\max_{1\le i\le n}\Big|\log h_t - \log h_0\Big| \le C(|X_i^\top\beta^*|^2 + t^2|X_i^\top b|^2 + t|X_i^\top b|),$$

which implies

$$\exp\{-C(|X_i^\top\beta^*|^2+t^2|X_i^\top b|^2+t|X_i^\top b|)\}\le\frac{h_t}{h_0}\le\exp\{C(|X_i^\top\beta^*|^2+t^2|X_i^\top b|^2+t|X_i^\top b|)\}. \tag{1.18}$$

Thus, for $\psi_0(b)=\|b\|_2$,

$$\begin{aligned}
F(\xi,S;\psi)&=\inf_{b\in\mathcal{C}(\xi,S),\psi_0(b)\le1}\frac{1}{n}\sum_{i=1}^n\int_0^1 h_t dt\frac{e^{\psi_0^2(b)+\psi_0(b)}(b^\top X_i)^2}{\|b_S\|_1\|b\|_2}\\
&\ge\inf_{b\in\mathcal{C}(\xi,S),\psi_0(b)\le1}\frac{1}{n}\sum_{i=1}^n\int_0^1 e^{-C|X_i^\top\beta^*|^2}h_0\frac{e^{\psi_0^2(b)+\psi_0(b)}(b^\top X_i)^2}{e^{C(t^2|X_i^\top b|^2+t|X_i^\top b|)}\|b_S\|_1\|b\|_2}dt\\
&\ge\inf_{b\in\mathcal{C}(\xi,S),\psi_0(b)\le1}\frac{1}{n}\sum_{i=1}^n\int_0^1 I\{\sqrt{C}t|X_i^\top b|<\psi_0(b)\}dt\frac{e^{-C|X_i^\top\beta^*|^2}h_0(b^\top X_i)^2}{\|b_S\|_1\|b\|_2}\\
&\ge\inf_{b\in\mathcal{C}(\xi,S),\psi_0(b)\le1}\frac{1}{n}\sum_{i=1}^n\frac{e^{-C|X_i^\top\beta^*|^2}h_0(b^\top X_i)^2}{\|b_S\|_1\|b\|_2}\min\left\{1,\frac{\psi_0(b)}{\sqrt{C}|X_i^\top b|}\right\}
\end{aligned}$$

By the scale invariance of b, we can further reduce the right-hand-side by

$$\begin{aligned}
F(\xi,S;\psi)&\ge\inf_{b\in\mathcal{C}(\xi,S),\|b\|_2=1}\frac{1}{n}\sum_{i=1}^n\frac{e^{-C|X_i^\top\beta^*|^2}h_0}{\|b_S\|_1}\min\{(b^\top X_i)^2,|b^\top X_i|/\sqrt{C}\}\\
&\ge ck^{-1/2}\inf_{b\in\mathcal{C}(\xi,S),\|b\|_2=1}\frac{1}{n}\sum_{i=1}^n e^{-C|X_i^\top\beta^*|^2}h_0\min\{(b^\top X_i)^2,|b^\top X_i|\}\\
&\ge ck^{-1/2}\inf_{b\in\mathcal{C}(\xi,S),\|b\|_2=1}\frac{1}{n}\sum_{i=1}^n e^{-C|X_i^\top\beta^*|^2}h_0\min\{(b^\top X_i)^2,|b^\top X_i|\}1\{|X_i^\top\beta^*|<T\}.
\end{aligned}$$

Now since in the region $|X_i^\top\beta^*|<T$, by condition (L1)

$$h_0\ge c_T, \tag{1.19}$$

so that for some $c>0$

$$F(\xi,S;\psi)\ge ck^{-1/2}\inf_{b\in\mathcal{C}(\xi,S),\|b\|_2=1}\frac{1}{n}\sum_{i=1}^n\min\{(b^\top X_i)^2,|b^\top X_i|\}1\{|X_i^\top\beta^*|<T\}.$$

In addition, we consider the truncation function $\varphi_T(x)$ for some $T>1$ such that

$$\varphi_T(x)=\begin{cases}\min\{|x|,x^2\} & \text{if } |x|\le T\\ 2T-|x| & \text{if } |x|>T\end{cases}.$$

Then we also have

$$F(\xi,S;\psi)\ge c_Tk^{-1/2}\inf_{b\in\mathcal{C}(\xi,S),\|b\|_2=1}\frac{1}{n}\sum_{i=1}^n\varphi_T(X_i^\top b)\cdot1\{|X_i^\top\beta|<T\}.$$

In the following, we will show that, for some $T > 0$,

$$P\bigg(\inf_{b\in\mathcal{C}(\xi,S),\|b\|_2=1}\frac{1}{n}\sum_{i=1}^{n}\varphi_T(X_i^\top b)\cdot 1\{|X_i^\top\beta|<T\} > c\bigg) \geq 1-e^{-cn-\log p}. \tag{1.20}$$

We denote

$$f_b(x) = x^\top b\cdot 1\{|x^\top\beta^*|<T\}, \qquad g_b(x)=\varphi(f_b(x)).$$

We have

$$\varphi_T(X_i^\top b)\cdot 1\{|X_i^\top\beta^*|<T\} = g_b(X_i).$$

For each $t > 0$, we define the event

$$\mathcal{E}(t) = \bigg\{\mathbb{P}_n[g_b(x)] \leq c_1\bigg(1-c_2\sqrt{\frac{\log p}{n}}t\bigg), \quad \text{for some } b\in\mathcal{C}(\xi,S) \text{ with } \|b\|_2=1 \text{ and } \|b\|_1=t\bigg\},$$

where $\mathbb{P}_n$ is the empirical expectation over n samples. The following lemma shows that this event has very small probability to happen.

Lemma 4. *Under the conditions of Proposition 3, it holds that* $P(\mathcal{E}(t)) \leq c_1\exp(-c_2n - c_3t\log p)$.

Hence, the lower bound argument (1.20) can be shown by noticing that, for any $b\in\mathcal{C}(\xi,S)$ and $\|b\|_2=1$,

$$1\leq\|b\|_1\leq(1+\xi)\|b_S\|_1\leq(1+\xi)\sqrt{k},$$

and that $k\lesssim n/\log p$.

Lastly, (iii) can be shown with the similar argument as part (i) of the proof of (1.2), along with the fact (1.19).

Combining the above arguments, by Proposition 3, we have $\mathcal{B}_2$ holds with probability at least $1-p^{-c}$.

Event $\mathcal{B}_1$. By Lemma 7 in the proof of Proposition 3, we have, under the event Ω_0 defined therein

$$\hat\beta-\beta^*\in\mathcal{C}(\xi,S), \tag{1.21}$$

which means

$$\|\hat\beta-\beta^*\|_1\leq(1+\xi)\|(\hat\beta-\beta^*)_S\|_1\leq(1+\xi)\sqrt{k}\|(\hat\beta-\beta^*)_S\|_2\lesssim\sqrt{\frac{k\log p}{n}},$$

with probability at least $1-p^{-c}$.

Event $\mathcal{B}_3$. Note that

$$\frac{1}{n}\|X(\hat{\beta}-\beta^*)\|_2 = \frac{1}{n}(\hat{\beta}-\beta^*)^\top X^\top X(\hat{\beta}-\beta^*).$$

By Theorem 1.6 of Zhou (2009) and (1.21), we have

$$\frac{1}{n}(\hat{\beta}-\beta^*)^\top X^\top X(\hat{\beta}-\beta^*) \le C\|\hat{\beta}-\beta^*\|_2^2 \lesssim \frac{k\log p}{n},$$

with probability at least $1-p^{-c}$. □

1.3 Proof of Theorem 3

First, we show that

$$|\hat{v}_j - v_j| \to_P 0. \tag{1.22}$$

To see this, note that, with probability at least $1-p^{-c}-n^{-c}$,

$$\begin{aligned}
|v_j-\hat{v}_j| &\le \left|\hat{u}^\top\left[\frac{1}{n}\sum_{i=1}^n \frac{f'^2(X_i^\top\hat{\beta})}{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}\left(\frac{f(X_i^\top\beta)(1-f(X_i^\top\beta))}{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}-1\right)X_iX_i^\top\right]\hat{u}\right| \\
&\le C\left|\left[\frac{1}{n}\sum_{i=1}^n \frac{f'^2(X_i^\top\hat{\beta})}{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}\hat{u}^\top X_iX_i^\top\right]\hat{u}\right| \cdot \max_{1\le i\le n}\left|\frac{f(X_i^\top\beta)(1-f(X_i^\top\beta))}{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}-1\right| \\
&\le C\left|\frac{1}{n}\sum_{i=1}^n \frac{f'^2(X_i^\top\hat{\beta})}{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}(u^{*\top}X_i)^2\right| \cdot \max_{1\le i\le n}\left|\frac{f(X_i^\top\beta)(1-f(X_i^\top\beta))}{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}-1\right| \\
&= o(1),
\end{aligned}$$

where the last inequality follows from the concentration inequality of subexponential random variables and the boundedness of the spectrum of $\mathbf{\Gamma}$, as well as Lemma 8.

By Theorem 1, we have, for $\delta \asymp \tau_n k\log p/n$,

$$\begin{aligned}
&P_\theta(\beta_j \in \mathrm{CI}_\alpha(\beta_j,\mathcal{D})) = P_\theta(|\tilde{\beta}_j-\beta_j|\le\tilde{\rho}_j) = P_\theta(|A_n+B_n|\le\tilde{\rho}_j) \\
&\ge P_\theta(|A_n|\le\tilde{\rho}_j-\delta_n, |B_n|\le\delta_n) \ge 1-P_\theta(|A_n|\ge\tilde{\rho}_j-\delta_n)-P(|B_n|\ge\delta_n),
\end{aligned}$$

where by (1.3) and (1.4), we have $\lim_{n,p\to\infty}\sup_{\theta\in\Theta(k)}P(|B_n|\ge\delta_n)=0$. As for $P_\theta(|A_n|\ge\tilde{\rho}_j-\delta_n)$, if $k\lesssim\frac{\sqrt{n}}{\tau_n\log p}$, by (1.22) there exists some sequence $\delta_n'\to 0$ such that $\lim_{n,p\to\infty}\sup_{\theta\in\Theta(k)}P_\theta(|\hat{v}_j-v_j|z_{\alpha/2}>\delta_n')=0$, so that

$$P_\theta(|A_n|\ge\tilde{\rho}_j-\delta_n) \le P_\theta(\sqrt{n}|A_n|\ge z_{\alpha/2}v_j-\delta_n'-\sqrt{n}\delta_n)+P_\theta(|\hat{v}_j-v_j|z_{\alpha/2}>\delta_n').$$

By (1.5) and the fact that $|\delta_n'+\sqrt{n}\delta_n|\to 0$, we have $\overline{\lim}_{n,p\to\infty}\sup_{\theta\in\Theta(k)}P_\theta(\sqrt{n}|A_n|\ge z_{\alpha/2}v_j-\delta_n'-\sqrt{n}\delta_n|\mathcal{D}_2,X)\le\alpha$, so that

$$\begin{aligned}
\overline{\lim_{n,p\to\infty}}\sup_{\theta\in\Theta(k)}P_\theta(|A_n|\ge\tilde{\rho}_j-\delta_n) &\le \overline{\lim_{n,p\to\infty}}\sup_{\theta\in\Theta(k)}\int P_\theta(\sqrt{n}|A_n|\ge z_{\alpha/2}v_j-\delta_n'-\sqrt{n}\delta_n|\mathcal{D}_2,X)dP_{\mathcal{D}_2,X} \\
&\le \overline{\lim_{n,p\to\infty}}\sup_{\theta\in\Theta(k)}\int \sup_{\theta\in\Theta(k)}P_\theta(\sqrt{n}|A_n|\ge z_{\alpha/2}v_j-\delta_n'-\sqrt{n}\delta_n|\mathcal{D}_2,X)dP_{\mathcal{D}_2,X} \\
&\le \alpha.
\end{aligned}$$

If instead $\frac{\sqrt{n}}{\tau_n \log p} \lesssim k \ll \frac{n}{\tau_n^2 \log p}$, then by (1.5) and Lemma 3, we have $\overline{\lim}_{n,p\to\infty} \sup_{\theta\in\Theta(k)} P_\theta(|A_n| \geq \tilde{\rho}_j - \delta_n) \leq \overline{\lim}_{n,p\to\infty} \sup_{\theta\in\Theta(k)} P_\theta(|A_n| \geq Ck\tau_n \log p/n - \delta_n) = 0$ since we can choose $C > 0$ such that $Ck\tau_n \log p/n \gtrsim \delta_n$. Therefore,

$$\underline{\lim}_{n,p\to\infty} \inf_{\theta\in\Theta(k)} P_\theta(\beta_j \in \mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \geq 1 - \overline{\lim}_{n,p\to\infty} \sup_{\theta\in\Theta(k)} P_\theta(|A_n| \geq \tilde{\rho}_j - \delta_n) \geq 1 - \alpha.$$

The second statement follows directly from the definition of $\tilde{\rho}_j$ and that of $\hat{v}_j$.

1.4 Proof of Theorem 4

Without loss of generality, we set $j = 1$. The proof of the lower bound relies on the construction of two parameter spaces which are different but not "testable" by any statistical procedure, which we refer as null and alternative.

The null space is taken as $\mathcal{H}_0 = \{(0, \mathbf{I})\}$, which consists of only one point. In the following proof, we first construct $\mathcal{H}_1$ and show that $\mathcal{H}_1 \subset \Theta(k)$; then we will control the distribution distance $TV(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}})$, where $f_{\pi_{\mathcal{H}_1}}$ is the distribution with parameter $(\beta, \mathbf{\Sigma})$ that has a prior $\pi_{\mathcal{H}_1}$ over $\mathcal{H}_1$ and $f_{\pi_{\mathcal{H}_0}}$ is the distribution with parameter $(\beta^*, \mathbf{\Sigma}^*) \in \mathcal{H}_0$ (here $\pi_{\mathcal{H}_0}$ is simply a point mass at $(0, \mathbf{I})$); lastly, we calculate the distance $\beta_1 - \beta_1^*$ where $(\beta, \mathbf{\Sigma}) \in \mathcal{H}_1$ and $(\beta^*, \mathbf{\Sigma}^*) \in \mathcal{H}_0$. Then we can apply the following lemma.

Lemma 5 (Cai and Guo 2017)**.** *Assume $T(\theta) = \mu_0$ for $\theta \in \mathcal{H}_0$ and $T(\theta) = \mu_1$ for $\theta \in \mathcal{H}_1$ and $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$. For any $CI_\alpha(T, Z) \in \mathcal{I}_\alpha(T, \mathcal{H})$,*

$$L(CI_\alpha(T, Z), \mathcal{H}) \geq L(CI_\alpha(T, Z), \mathcal{H}_0) \geq |\mu_1 - \mu_0|(1 - 2\alpha - TV(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}))_+. \tag{1.23}$$

In our theorem, the lower bound consists of two pieces. Accordingly, the rest of our proof is separated into two parts, one corresponding to the high-dimensional rate $k \log p/n$, and the other concerning the parametric rate $1/\sqrt{n}$.

1.4.1 The High-Dimensional Rate

We first prove the high-dimensional rate $k \log p/n$ following the steps as we stated earlier.

Step 1: Construction of $\mathcal{H}_1$. To construct $\mathcal{H}_1$, we define $\ell(M, n)$ as the set of all the n-element subsets of M. Define the covariance matrix $\mathbf{\Sigma}_I(\rho)$ such that

$$\mathbf{\Sigma}_I(\rho)^{-1} = \mathbf{\Omega}_I(\rho) = \mathbf{I}_p + \begin{bmatrix} 0 & \delta_I^\top(\rho) \\ \delta_I(\rho) & 0 \end{bmatrix}, \tag{1.24}$$

where $\delta_I(\rho) = (\delta_{I1}, ..., \delta_{I,p-1}) \in \mathbb{R}^{p-1}$ such that $\delta_{I,j-1} = \rho \cdot 1\{j \in I\}$. Based on this definition, the covariance matrix can be derived as

$$\mathbf{\Sigma}_I(\rho) = \begin{bmatrix} \frac{1}{1-\delta_I(\rho)^\top \delta_I(\rho)} & -\frac{\delta_I(\rho)^\top}{1-\delta_I(\rho)^\top \delta_I(\rho)} \\ -\frac{\delta_I(\delta)}{1-\delta_I(\rho)^\top \delta_I(\rho)} & \mathbf{I}_{p-1} + \frac{\delta_I(\rho)\delta_I(\rho)^\top}{1-\delta_I(\rho)^\top \delta_I(\rho)} \end{bmatrix}. \tag{1.25}$$

We define the alternative parameter space $\mathcal{H}_1 = \{(\beta_I, \mathbf{\Sigma}_I) : \|\beta_I\|_0 = k, \beta_{I1} = (k-1)\rho^2, \beta_{Ij} = \rho 1\{j \in I\} \mathbf{\Sigma}_I = \mathbf{\Sigma}_I(\rho), \text{ for } I \in \ell([2:p], k-1)\}$, so that $\mathcal{H}_1$ contains all the $(\beta_I, \mathbf{\Sigma}_I)$ for k-sparse β_I with constant first component being $(k-1)\rho^2$ and other $(k-1)$ nonzero components being ρ (indexed by I), and $\mathbf{\Sigma}_I(\rho)$ defined by (1.25) with the $(k-1)$ sparse vector $\delta_I(\rho)$ supported on the same set I.

Based on this definition, it can be shown that $\mathbf{\Sigma}_I(\rho)$ satisfies the condition

$$\frac{1}{M} \leq \lambda_{\min}(\mathbf{\Sigma}_I(\rho)) \leq \lambda_{\max}(\mathbf{\Sigma}_I(\rho)) \leq M. \tag{1.26}$$

Note that the second matrix on the right hand side of (1.24) is of spectral norm $\|\delta_I(\rho)\|_2 = \rho\sqrt{k-1}$. Then by Weyl's inequality $\max\{|\lambda_{\min}(\mathbf{\Omega}_I(\rho)) - 1|, |\lambda_{\max}(\mathbf{\Omega}_I(\rho)) - 1|\} \leq \|\delta_I(\rho)\|_2$. So that when $\|\delta_I(\rho)\|_2$ is chosen such that $\|\delta_I(\rho)\|_2 \leq \min\left\{1 - \frac{1}{M}, M - 1\right\}$, we will have (1.26) hold. Thus $\mathcal{H}_1 \subseteq \Theta(k)$.

Step 2: Control $TV(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}})$. To control $TV(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}})$, it suffices to control $\chi^2(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}})$, and we can applythe inequality $TV(f_1, f_0) \leq \sqrt{\chi^2(f_1, f_0)}$ (see p.90 of Tsybakov (2009)). Let π denote the uniform prior on I over $\ell([2:p], k-1)$. This prior induces a prior distribution $\pi_{\mathcal{H}_1}$ over the parameter space $\mathcal{H}_1$. By definition, the joint distribution of $Z_i = (y_i, X_i)$ is

$$\begin{aligned} p(X_i, y_i; \beta, \mathbf{\Sigma}) &= \frac{1}{\sqrt{(2\pi)^p |\mathbf{\Sigma}|}} \exp\left\{ -\frac{1}{2} X_i^\top \mathbf{\Sigma}^{-1} X_i \right\} f(X_i^\top \beta)^{y_i} (1 - f(X_i^\top \beta)))^{1-y_i} \\ &= \frac{1}{\sqrt{(2\pi)^p |\mathbf{\Sigma}|}} \exp\left\{ -\frac{1}{2} X_i^\top \mathbf{\Sigma}^{-1} X_i \right\} \left(\frac{f(X_i^\top \beta)}{1 - f(X_i^\top \beta)} \right)^{y_i} (1 - f(X_i^\top \beta)). \end{aligned}$$

For $(0, \mathbf{I}) \in \mathcal{H}_0$, the corresponding joint distribution of the data $(Z_1, ..., Z_n)$ is

$$g_0 = \prod_{i=1}^n p(X_i, y_i; 0, \mathbf{I}) = \frac{1}{(2\pi)^{np/2}} \prod_{i=1}^n \frac{1}{2} e^{-\|X_i\|_2^2/2}.$$

Similarly, the marginal distribution of the samples with parameter in $\mathcal{H}_1$ is denoted as

$$g_1 = \int_{\mathcal{H}_1} \prod_{i=1}^n p(X_i, y_i; \beta, \mathbf{\Sigma}) \pi_{\mathcal{H}_1} = \frac{1}{\binom{p-1}{k-1}^n} \sum_{(\beta, \mathbf{\Sigma}) \in \mathcal{H}_1} \prod_{i=1}^n p(X_i, y_i; \beta, \mathbf{\Sigma}).$$

Therefore we have

$$\begin{aligned} \chi^2(g_1, g_0) &= \int \frac{g_1^2}{g_0} - 1 \\ &= \frac{1}{\binom{p-1}{k-1}^{2n}} \sum_{(\beta, \mathbf{\Sigma}) \in \mathcal{H}_1} \sum_{(\beta', \mathbf{\Sigma}') \in \mathcal{H}_1} \prod_{i=1}^n \int \frac{p(X_i, y_i; \beta, \mathbf{\Sigma}) p(X_i, y_i; \beta', \mathbf{\Sigma}')}{p(X_i, y_i; 0, \mathbf{I})} - 1, \end{aligned} \tag{1.27}$$

where by Lemma 1,

$$\int \frac{p(X_i, y_i; \beta, \mathbf{\Sigma}) p(X_i, y_i; \beta', \mathbf{\Sigma}')}{p(X_i, y_i; 0, \mathbf{I})} = \frac{4 \det(\mathbf{\Omega}) \det(\mathbf{\Omega}')}{\det(\mathbf{\Omega} + \mathbf{\Omega}' - I)} \cdot \mathbb{E} f(X^\top \beta) f(X^\top \beta'),$$

where $X \sim N(0, (\mathbf{\Omega} + \mathbf{\Omega}' - I)^{-1})$, $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ and $\mathbf{\Omega}' = (\mathbf{\Sigma}')^{-1}$. By construction, we have

$$\mathbf{\Omega} + \mathbf{\Omega}' - \mathbf{I} = \begin{bmatrix} 1 & (\delta + \delta')^\top \\ \delta + \delta' & \mathbf{I} \end{bmatrix},$$

and

$$(\mathbf{\Omega} + \mathbf{\Omega}' - \mathbf{I})^{-1} = \begin{bmatrix} \frac{1}{1 - 2(\|\delta\|_2^2 + \delta^\top \delta')} & -\frac{(\delta + \delta')^\top}{1 - 2(\|\delta\|_2^2 + \delta^\top \delta')} \\ -\frac{\delta + \delta'}{1 - 2(\|\delta\|_2^2 + \delta^\top \delta')} & \mathbf{I} + \frac{(\delta + \delta')(\delta + \delta')^\top}{1 - 2(\|\delta\|_2^2 + \delta^\top \delta')} \end{bmatrix}.$$

Hence, $X^\top \beta \sim N(0, \beta^\top (\mathbf{\Omega} + \mathbf{\Omega}' - \mathbf{I})^{-1} \beta)$ and $X^\top \beta' \sim N(0, \beta'^\top (\mathbf{\Omega} + \mathbf{\Omega}' - \mathbf{I})^{-1} \beta')$, where

$$\begin{aligned} \beta^\top (\mathbf{\Omega} + \mathbf{\Omega}' - \mathbf{I})^{-1} \beta = \beta'^\top (\mathbf{\Omega} + \mathbf{\Omega}' - \mathbf{I})^{-1} \beta' &= \|\delta\|_2^2 + \frac{((k-1)\rho^2 - (\|\delta\|_2^2 + \delta^\top \delta'))^2}{1 - 2(\|\delta\|_2^2 + \delta^\top \delta')} \\ &= \|\delta\|_2^2 + \frac{(\delta^\top \delta')^2}{1 - 2(\|\delta\|_2^2 + \delta^\top \delta')}, \end{aligned}$$

and $\mathrm{Cov}(X^\top \beta, X^\top \beta') = \beta^\top (\mathbf{\Omega} + \mathbf{\Omega}' - \mathbf{I})^{-1} \beta' = \delta^\top \delta' + \frac{(\delta^\top \delta')^2}{1 - 2(\|\delta\|_2^2 + \delta^\top \delta')} \le 2\delta^\top \delta'$, as long as $\|\delta\|_2^2 = (k-1)\rho^2 = o(1)$. As a result, by Lemma 2, we have

$$\mathbb{E}_{\mathbf{\Sigma}, \mathbf{\Sigma}'} h(V; \beta, \beta') \le 1 + C\left(\delta^\top \delta' + \frac{(\delta^\top \delta')^2}{1 - 2(\|\delta\|_2^2 + \delta^\top \delta')}\right) \le 1 + 2C\delta^\top \delta' = 1 + 2Cj\rho^2$$

where $j = |\mathrm{supp}(\delta) \cap \mathrm{supp}(\delta')| = |I \cap I'|$ is the number of intersected components between δ and δ'. In addition, we have $\det(\mathbf{\Omega}) = 1 - \|\delta\|_2^2$ and $\det(\mathbf{\Omega} + \mathbf{\Omega}' - \mathbf{I}) = 1 - 2(\|\delta\|_2^2 + \delta^\top \delta')$. Then as long as $\|\delta\|_2^2 = o(1)$, we have $\det(\mathbf{\Omega}) \le 1/2$ and $\det(\mathbf{\Omega} + \mathbf{\Omega}' - \mathbf{I}) \ge 1/2$ for some sufficiently large (n, p). Hence $\chi^2(g, f) \le c\mathbb{E}(1 + CJ\rho^2)^n - 1$, where J follows a hypergeometric distribution $P(J = j) = \frac{\binom{k-1}{j}\binom{p-k}{k-1-j}}{\binom{p-1}{k-1}}, j = 0, 1, ..., k-1$. Then $\chi^2(g, f) \le c\mathbb{E} \exp(n \log(1 + C\rho^2 J)) - 1 \le c\mathbb{E} e^{Cn\rho^2 J} - 1$. As shown on page 173 of Aldous (1985), J has the same distribution as the random variable $\mathbb{E}(Z|\mathcal{B}_n)$ where Z is a binomial random variable of parameters $(k, k/p)$ and $\mathcal{B}_n$ some suitable σ-algebra. Thus by Jensen's inequality we have $\mathbb{E} e^{CnJ\rho^2} \le \left(1 - \frac{k-1}{p-1} + \frac{k-1}{p-1} e^{Cn\rho^2}\right)^{k-1}$. Hence, let $\rho^2 = \frac{1}{Cn} \log\left(1 + \frac{p-1}{c_1 (k-1)^2}\right)$ for some constant $c_1 > 0$, we have $\chi^2(g, f) \le c_2$ for some small constant $c_2 > 0$.

Step 3. Obtain the Lower Bound. Now by Lemma 5, we have

$$\inf_{\mathrm{CI}_\alpha(\beta_j, \mathcal{D}) \in \mathcal{I}_\alpha(\Theta(k), \beta_j)} \sup_{\theta \in \Theta(k)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \ge \inf_{\mathrm{CI}_\alpha(\beta_j, \mathcal{D}) \in \mathcal{I}_\alpha(\Theta(k), \beta_j)} \mathbb{E}_{\theta^*} L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \gtrsim \frac{k \log p}{n}, \tag{1.28}$$

where $\theta^* = (0, \mathbf{I}_p)$. Thus we have proven the lower bound $k \log p / n$.

1.4.2 The Parametric Rate

To prove the parametric rate lower bound, we define $\mathcal{H}_0 = \{(0, \sigma^2\mathbf{I})\}$, and $\mathcal{H}_1 = \{(\beta, \sigma^2\mathbf{I}) : \beta_1 = \rho/\sqrt{n}, \beta_i = 0, \forall i \neq 1\}$. By assumption, we have $M^{-1} < \sigma^2 < M$. Denote

$$g = g(Z) = \frac{1}{(2\pi\sigma^2)^{np/2}} \prod_{i=1}^{n} \exp\left\{ -\frac{1}{2\sigma^2} X_i^\top X_i \right\} \left(\frac{f(X_i^\top \beta)}{1 - f(X_i^\top \beta)} \right)^{y_i} (1 - f(X_i^\top \beta)), \quad (1.29)$$

which is the joint density with parameter $(\beta, \sigma^2\mathbf{I}) \in \mathcal{H}_1$. Also denote g_0 as before as the joint density with $(\beta, \sigma^2\mathbf{I}) \in \mathcal{H}_0$. The χ^2-divergence between distribution g and g_0 is $\chi^2 = \int \frac{g^2}{g_0} - 1 = (\mathbb{E}\Delta(X_i; \beta))^n - 1$, with $\Delta(X_i; \beta) = 2 + 4f(X_i^\top \beta)(f(X_i^\top \beta) - 1) \leq 2 + 4\Phi(C_1 X_i^\top \beta)(\Phi(C_1 X_i^\top \beta) - 1)$, where the last inequality follows from condition (L2) and the monotonicity of the function $2 + 4x(x-1)$ for $x \geq 0$.

Lemma 6. *For any $x \in \mathbb{R}$, we have $4\Phi^2(x) - 4\Phi(x) + 1 = (2\Phi(x) - 1)^2 \leq x^2$.*

From the above lemma, we have $\Delta(X_i; \beta) \leq 1 + (C_1 X_i^\top \beta)^2$ for $X_i^\top \beta \in \mathbb{R}$. Then $\mathbb{E}\Delta(X_i; \beta) \leq 1 + C_1^2\sigma^2 \sum_{j=1}^{p} \beta_j^2 = 1 + \frac{C_1^2\sigma^2\rho^2}{n}$, where the last inequality follows from the fact that $\sum_{i=1}^{p} \beta_i x_i \sim N(0, \sigma^2\|\beta\|_2^2)$ and $\beta_i = \rho/\sqrt{n} I(i = 1)$. Then we have $\chi^2 \leq \big(1 + \frac{C_1^2\sigma^2\rho^2}{n}\big)^n - 1$, Again, by Lemma 5, we obtain that, for sufficiently large n,

$$\begin{aligned} \inf_{\mathrm{CI}_\alpha(\beta_j, \mathcal{D}) \in \mathcal{I}_\alpha(\Theta(k), \beta_j)} \sup_{\theta \in \Theta(k)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) &\geq \frac{\rho}{\sqrt{n}} \left[2 - 2\alpha - \left(1 + \frac{C_1^2\sigma^2\rho^2}{n} \right)^n \right] \\ &\geq \frac{\rho}{\sqrt{n}} \left[2 - 2\alpha - e^{C_1^2\rho^2/M} \right]. \end{aligned}$$

Optimizing the last inequality, we have

$$\inf_{\mathrm{CI}_\alpha(\beta_j, \mathcal{D}) \in \mathcal{I}_\alpha(\Theta(k), \beta_j)} \sup_{\theta \in \Theta(k)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \geq \frac{C(M, C_1, \alpha)}{\sqrt{n}}, \quad (1.30)$$

where $C(M, c, \alpha) = \min_{\rho \geq 0} \rho(2 - 2\alpha - e^{C_1^2\rho^2/M})$.

1.5 Proof of Theorem 5

The proof of Theorem 5 follows from the same argument of the proof of Theorem 4. On the one hand, by the proof of the high-dimensional rate (Step 3) and the proof of the parametric rate in Theorem 4, we have

$$\inf_{\mathrm{CI}_\alpha(\beta_j, \mathcal{D}) \in \mathcal{I}_\alpha(\Theta(k), \beta_j)} \mathbb{E}_{\theta^*} L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \gtrsim \frac{k \log p}{n} + \frac{1}{\sqrt{n}}.$$

Since $\theta^* = (0, \mathbf{I}_p) \in \Theta(k_1)$, it holds that

$$L_\alpha^*(\Theta(k_1), \Theta(k), \beta_j) \geq \inf_{\mathrm{CI}_\alpha(\beta_j, \mathcal{D}) \in \mathcal{I}_\alpha(\Theta(k), \beta_j)} \mathbb{E}_{\theta^*} L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \gtrsim \frac{k \log p}{n} + \frac{1}{\sqrt{n}}. \quad (1.31)$$

□

1.6 Proof of Theorem 6

The classical resularity conditions for asymptotic normality of MLEs can be found, for example, in Theorem 4.16 of Shao (2003). In the following, we explicitly state these regularity conditions. Under these regularity conditions, the two statement of the theorem follows directly from Theorem 4.17 of Shao (2003). For simplicity, we denote $\eta = \eta_j$.

(C1) Compact parameter space: the parameter space Θ of η is finite dimensional, closed and bounded.

(C2) Differentiability: for every $(y, W) \in \{0, 1\} \times \mathbb{R}$, the function $p_\eta(y, W)$ is twice continuously differentiable in η and satisfies

$$\frac{\partial}{\partial \eta} \int \psi_\eta(y, W) dy dW = \int \frac{\partial}{\partial \eta} \psi_\eta(y, W) dy dW$$

for $\psi_\eta(y, W) = p_\eta(y, W)$ and $= \partial p_\eta(y, W)/\partial \eta$.

(C3) Positive definite Fisher information: the Fisher information matrix

$$I(\eta) = \mathbb{E}_\eta \left[\frac{\partial \log p_\eta(y, W)}{\partial \eta} \right]^2$$

is positive definite.

(C4) Integrability: for each $\eta_0 \in \Theta$, there exists $h(y, W)$ such that in a neighborhood of η_0, all the 2nd order partial derivatives of $\log p_\eta(y, W)$ is bounded by $h(y, W)$ with $\mathbb{E}_{\eta_0} |h(y, W)| < \infty$.

1.7 Proof of Theorem 7

The upper bound

$$\sup_{\theta \in \Theta_0(k)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha^{**}(\beta_j, \mathcal{D}')) \lesssim \frac{1}{\sqrt{n}}$$

follows directly from Theorem 6. The rest of the proof is devoted to the lower bound. In other words, we show that

$$\inf_{\mathrm{CI}_\alpha(\beta_j, \mathcal{D}') \in \mathcal{I}_\alpha(\Theta_0(k), \beta_j)} \sup_{\theta \in \Theta_0(k)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D}')) \gtrsim \frac{1}{\sqrt{n}}. \tag{1.32}$$

Without loss of generality, we set $j = 1$. Specifically, for some given $(k-1)$-sparse vector $\alpha \in \mathbb{R}^{p-1}$ such that $\|\alpha\|_2 \asymp 1$, we define $\mathcal{H}_0 = \{(\beta^*, \mathbf{\Sigma}_0)\}$ where $\beta^* = (0, \alpha^\top) \in \mathbb{R}^p$ and $\mathcal{H}_1 = \{(\beta^{**}, \mathbf{\Sigma}_0)\}$ where $\beta^{**} = (1/\sqrt{n}, \alpha^\top) \in \mathbb{R}^p$. Apparently, $\mathcal{H}_0 \cup \mathcal{H}_1 \subset \Theta_0(k)$. Denote

$$f = f(Z) = \frac{1}{(2\pi)^{np/2} |\mathbf{\Sigma}|^{n/2}} \prod_{i=1}^n \exp\left\{ -\frac{1}{2} X_i^\top \mathbf{\Sigma}_0^{-1} X_i \right\} \left(\frac{f(X_i^\top \beta^*)}{1 - f(X_i^\top \beta^*)} \right)^{y_i} (1 - f(X_i^\top \beta^*)), \tag{1.33}$$

which is the joint density with the parameter from $\mathcal{H}_0$, and

$$g = g(Z) = \frac{1}{(2\pi)^{np/2}|\mathbf{\Sigma}|^{n/2}} \prod_{i=1}^{n} \exp\Big\{ -\frac{1}{2} X_i^\top \mathbf{\Sigma}_0^{-1} X_i \Big\} \Big(\frac{f(X_i^\top \beta^{**})}{1 - f(X_i^\top \beta^{**})} \Big)^{y_i} (1 - f(X_i^\top \beta^{**})), \tag{1.34}$$

as the joint density with the parameter from $\mathcal{H}_1$. The χ^2-divergence between distribution g and f is $\chi^2 = \int \frac{g^2}{f} - 1 = (\mathbb{E}\Delta(X_i; \beta^*, \beta^{**}))^n - 1$, with

$$\Delta(X_i; \beta^*, \beta^{**}) = \frac{(1 - f(X_i^\top \beta^{**}))^2}{1 - f(X_i^\top \beta^*)} + \frac{f(X_i^\top \beta^{**})^2}{f(X_i^\top \beta^*)}.$$

Note that

$$\mathbb{E}\Delta(X_i; \beta^*, \beta^{**}) = \mathbb{E}\Big[\frac{(1 - f(Z/\sqrt{n} + Y))^2}{1 - f(Y)} + \frac{f^2(Z/\sqrt{n} + Y)}{f(Y)} \Big] = 1 + \mathbb{E}\Big[\frac{(f(Y) - f(Z/\sqrt{n} + Y))^2}{f(Y)(1 - f(Y))} \Big],$$

where $Z \sim N(0, (\mathbf{\Sigma}_0)_{11})$, $Y \sim N(0, \alpha^\top (\mathbf{\Sigma}_0)_{-1,-1}\alpha)$ and $\mathrm{Cov}(Z, Y) = (\mathbf{\Sigma}_0)_{1,-1}^\top \alpha$. By Taylor expansion, (L1) and (L5), we have

$$\begin{aligned} &\mathbb{E}\Big[\frac{(f(Y) - f(Z/\sqrt{n} + Y))^2}{f(Y)(1 - f(Y))} \Big] \le \frac{C}{n} \mathbb{E}\Big[\frac{Z^2 f'(Y^*)}{f(Y)(1 - f(Y))} \Big] \le \frac{C_2}{n} \mathbb{E}\Big[\frac{1}{f(|Y|)(1 - f(|Y|))} \Big] \\ &\le \frac{C_2}{n} \mathbb{E}[1 + \exp(a|Y|)] \lesssim \frac{1}{n}, \end{aligned}$$

where Y^* is between Y and $Y + Z/\sqrt{n}$. Then

$$\chi^2 \le \big(1 + \frac{C_2}{n}\big)^n - 1 \le e^{1/C_2} - 1,$$

Now by Lemma 5, we obtain (1.32).

2 Proofs of Other Technical Results

2.1 Proof of Proposition 1

The Cramér-Rao lower bound can be obtained by directly calculating the Fisher information matrix.

2.2 Proof of Proposition 3

The proof generalizes the ideas in Lemma 1 and Theorem 4 of Huang and Zhang (2012). We begin with the following key lemma.

Lemma 7. *Under the conditions of Proposition 3, for any target vector β^*, we have*

$$\langle \hat{\beta} - \beta^*, \ell'(\hat{\beta}) - \ell'(\beta^*) \rangle + (\lambda - z^*)\|b_{S^c}\|_1 \le (\lambda + z^*)\|b_S\|_1. \tag{2.1}$$

Moreover, in the event $\Omega_0 = \{(\lambda + z^*)/(\lambda - z^*) \le \xi, z^* \le \lambda\}$, $\hat{\beta} - \beta^* \ne 0$ belongs to the sign-restricted cone $\mathcal{C}_-(\xi, S) = \{b \in \mathcal{C}(\xi, S) : b_j(\ell'(\beta + b) - \ell'(\beta))_j \le 0, \forall j \in S^c\}$.

Let $h = \hat{\beta} - \beta$ and $\Delta(\beta^* + b, \beta^*) = \langle b, \ell'(\beta^* + b) - \ell'(\beta^*) \rangle$. Since $\ell(\beta)$ is a convex function,

$$t^{-1}\Delta(\beta^* + th, \beta^*) = \frac{\partial}{\partial t}\Big\{ \ell(\beta^* + th) - t\langle h, \ell'(\beta^*) \rangle \Big\}$$

is an increasing function of t. For $0 \le t \le 1$ and in the event Ω, (2.1) implies

$$t^{-1}\Delta(\beta^* + th, \beta^*) \le \Delta(h + \beta^*, \beta^*) < (\lambda + z^*)\|h_S\|_1. \tag{2.2}$$

By (1.15) and (1.14),

$$F(\xi, S; \psi, \psi_0) \le \frac{\Delta(\beta^* + th, \beta^*) e^{\psi_0^2(th) + \psi_0(th)}}{t\|h_S\|_1 \psi(th)},$$

for $\psi_0(th) \le 1$. Next, we show that $\psi_0(th) < \eta \le 1/2$ for all $0 \le t \le 1$. If $\psi_0(h) \le 1/2$, in the event Ω, by (2.2), we have $\psi_0(th) \le 1/2 < 1$ and

$$\psi_0(th) e^{-\psi_0^2(th) - \psi_0(th)} \le \frac{\Delta(\beta^* + th, \beta^*)}{t\|h_S\|_1 F(\xi, S; \psi_0, \psi_0)} < \frac{(\lambda + z^*)}{F(\xi, S; \psi_0, \psi_0)} \le \eta e^{-\eta^2 - \eta}, \tag{2.3}$$

which, by the monotonicity of xe^{-x^2-x} on $[0, 1]$, implies $\psi_0(h) < \eta \le 1/2$ when $t = 1$. To see $\psi_0(h) \le 1/2$ holds, if instead $1/2 < \psi_0(h)$, the above inequality at $\psi(t_0 h) = 1/2$ for some t_0 would give $(1/2)e^{-(1/2)^2 - 1/2} < \eta e^{-\eta^2 - \eta}$, which is absurd. Thus $\psi_0(h) \le 1/2$ and, by (2.3), $\psi_0(th) < \eta$ for all $0 \le t \le 1$. On the other hand, by (1.14)

$$\psi(th) \le \frac{\Delta(\beta^* + th, \beta^*) e^{\psi_0^2(th) + \psi_0(th)}}{t\|h_S\|_1 F(\xi, S; \psi, \psi_0)} < \frac{(\lambda + z^*) e^{\psi_0^2(th) + \psi_0(th)}}{F(\xi, S; \psi, \psi_0)} \le \frac{(\lambda + z^*) e^{\eta^2 + \eta}}{F(\xi, S; \psi, \psi_0)}$$

This completes the proof. □

2.3 Proof of Lemma 1

Thus, since $f(x) - 1/2$ is an odd function (condition (L1)), we have

$$\begin{aligned}
&\int \frac{p(X_i, y_i; \beta, \mathbf{\Sigma}) p(X_i, y_i; \beta', \mathbf{\Sigma}')}{p(X_i, y_i; 0, \mathbf{I})} \qquad (2.4)\\
&= \frac{\det(\mathbf{\Omega})\det(\mathbf{\Omega}')}{(2\pi)^{p/2}} \int 2\exp\Big\{ -\frac{1}{2} X_i^\top (\mathbf{\Omega} + \mathbf{\Omega}' - I) X_i \Big\} (1 - f(X_i^\top \beta))(1 - f(X_i^\top \beta')) dX_i \\
&\quad + \frac{\det(\mathbf{\Omega})\det(\mathbf{\Omega}')}{(2\pi)^{p/2}} \int 2\exp\Big\{ -\frac{1}{2} X_i^\top (\mathbf{\Omega} + \mathbf{\Omega}' - I) X_i \Big\} f(X_i^\top \beta) f(X_i^\top \beta') dX_i \\
&= \frac{\det(\mathbf{\Omega})\det(\mathbf{\Omega}')}{\det(\mathbf{\Omega} + \mathbf{\Omega}' - I)} \cdot \mathbb{E}_{\mathbf{\Sigma}, \mathbf{\Sigma}'} h(X; \beta, \beta'), \qquad (2.5)
\end{aligned}$$

where $X \sim N(0, (\mathbf{\Omega} + \mathbf{\Omega}' - I)^{-1})$, $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, $\mathbf{\Omega}' = (\mathbf{\Sigma}')^{-1}$ and $h(X; \beta, \beta') = 4f(X^\top \beta) f(X^\top \beta')$. □

2.4 Proof of Lemma 2

Let $g(x) = f(x) - 1/2$. By (L1), $g(x)$ is an odd function. Since

$$\mathbb{E}f(X)f(Y) = \mathbb{E}(g(X) + 1/2)(g(Y) + 1/2) = \frac{1}{4} + \mathbb{E}g(X)g(Y),$$

it suffices to show that

$$\mathbb{E}g(X)g(Y) \leq C\sigma^2\rho.$$

Note that (X, Y) has the same distribution as $(-X, -Y)$. Then

$$\begin{aligned}
\mathbb{E}g(X)g(Y)1\{X > 0, Y > 0\} &= \mathbb{E}g(X)g(Y)1\{X < 0, Y < 0\}, \\
\mathbb{E}g(X)g(Y)1\{X < 0, Y > 0\} &= \mathbb{E}g(X)g(Y)1\{X > 0, Y < 0\},
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{E}g(X)g(Y) &= 2\mathbb{E}g(X)g(Y)1\{X > 0, Y > 0\} + 2\mathbb{E}g(X)g(Y)1\{X < 0, Y > 0\} \\
&= 2\mathbb{E}g(X)g(Y)1\{Y > 0\}.
\end{aligned}$$

Let $p(x, y; \rho, \sigma)$ be the joint distribution of (X, Y). Due to the fact that, for any $x, y > 0$, as long as $\rho \in (0, 1)$, it holds that

$$p(x, y; \rho, \sigma) \geq p(-x, y; \rho, \sigma),$$

and that, for $x \geq 0$, by condition (L2),

$$g(x) \leq \Phi(cx) - \frac{1}{2},$$

then we have

$$\begin{aligned}
\mathbb{E}g(X)g(Y)1\{Y > 0\} &= \int_{-\infty}^{\infty}\int_{0}^{\infty} g(x)g(y)p(x, y; \rho, \sigma)dydx \\
&= \int_{0}^{\infty}\int_{0}^{\infty} g(x)g(y)[p(x, y; \rho, \sigma) - p(-x, y; \rho, \sigma)]dydx \\
&\leq \int_{0}^{\infty}\int_{0}^{\infty} \left[\Phi(cx) - \frac{1}{2}\right]\left[\Phi(cy) - \frac{1}{2}\right](p(x, y; \rho, \sigma) - p(-x, y; \rho, \sigma))dydx \\
&= \int_{-\infty}^{\infty}\int_{0}^{\infty} \left[\Phi(cx) - \frac{1}{2}\right]\left[\Phi(cy) - \frac{1}{2}\right]p(x, y; \rho, \sigma)dydx \\
&= \mathbb{E}\left[\Phi(cX) - \frac{1}{2}\right]\left[\Phi(cY) - \frac{1}{2}\right]1\{Y > 0\}.
\end{aligned}$$

Hence,

$$\mathbb{E}g(X)g(Y) \leq \mathbb{E}\left[\Phi(cX) - \frac{1}{2}\right]\left[\Phi(cY) - \frac{1}{2}\right] = \mathbb{E}\Phi(cX)\Phi(cY) - \frac{1}{4}$$

The rest of the proof will focus on the upper bound of $\mathbb{E}\Phi(cX)\Phi(cY)$. By normalization, it suffices to show that, for the centred Gaussian random vector (X, Y) with covariance matrix $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, it holds that

$$\mathbb{E}\Phi(\sigma X)\Phi(\sigma Y) \leq \frac{1}{4} + C\sigma^2\rho,$$

for some universal constant $C > 0$. Note that the inner product

$$\langle X, Y \rangle = \mathbb{E}XY$$

defines a Hilbert space on $L^2(\Omega, \mathcal{F}, \mu)$. Then it suffices to show that

$$\left\langle \Phi(\sigma X), \Phi(\sigma Y) \right\rangle \leq \frac{1}{4} + C\sigma^2\langle X, Y \rangle.$$

Consider the Hermite polynomials $H_n(x), x \in \mathbb{R}, n = 0, 1, ...$ which are defined as

$$H_n = \frac{(-1)^n}{n!} e^{x^2/2} \frac{d^n}{dx^n}(e^{-x^2/2}),$$

so that in particular $H_0(x) = 1, H_1(x) = x, H_2(x) = (x^2 - 1)/2$, and in general $H_n(x)$ is a polynomial of order n. The Hermite polynomials satisfy the following basic identities

$$\begin{aligned} H_n'(x) &= H_{n-1}(x) \\ (n+1)H_{n+1}(x) &= xH_n(x) - H_{n-1}(x), \\ H_n(-x) &= (-1)^n H_n(x), \end{aligned} \tag{2.6}$$

for all $n \geq 1$. For X, Y that are $N(0, 1)$ random variables that are jointly Gaussian, it can be shown (see, for example, Lemma 1.1.1 of Nualart (2006)) that

$$\langle H_n(X), H_m(Y) \rangle = \mathbb{E}(H_n(X)H_m(Y)) = \begin{cases} 0 & \text{if } m \neq n, \\ \frac{1}{n!}(\mathbb{E}XY)^n & \text{if } m = n. \end{cases} \tag{2.7}$$

Now, by the Wigner chaos decomposition (see, for example, Theorem 1.1.1 of Nualart (2006)), we would like to expand the function $\Phi(\sigma x)$ in terms of orthogonal Hermite polynomials as

$$\Phi(\sigma x) = \sum_{n=0}^{\infty} C_n H_n(x).$$

To calculate the coefficients C_n, simply note that

$$C_n = \frac{\langle \Phi(\sigma X), H_n(X) \rangle}{\langle H_n(X), H_n(X) \rangle} = \frac{(-1)^n}{(2\pi)^{1/2}} \int \Phi(\sigma x) \frac{d^n}{dx^n}(e^{-x^2/2}) dx.$$

Denote $\phi(x) = e^{-x^2/2}$, using integration by parts, for any $n \geq 1$, we have

$$C_n = (-1)^n \int \Phi(\sigma x)\phi^{(n)}(x)dx = (-1)^{n-1}\sigma \int \phi(\sigma x)\phi^{(n-1)}(x)dx$$
$$= (n-1)!\sigma \int e^{-\frac{\sigma^2 x^2}{2} - x^2/2} H_{n-1}(x)dx = \frac{\sigma^n H_{n-1}(0)}{2\sqrt{\sigma^2+1}},$$

where the last equation follows from Equation 7.374.10 on p.804 of Gradshteyn and Ryzhik (2014). Hence, $C_1 = \frac{\sigma}{2\sqrt{\sigma^2+1}}$ and $C_n = 0$ for all $n \geq 2$. In addition,

$$C_0 = \int \Phi(\sigma x)\phi(x)dx = \sqrt{\frac{1+\sigma^2}{2\pi}} - \frac{1}{\sqrt{2\pi}} + \frac{1}{2}.$$

As a result,

$$\left\langle \Phi(\sigma X), \Phi(\sigma Y) \right\rangle = \left\langle \sum_{n=0}^{\infty} C_n H_n(X), \sum_{n=0}^{\infty} C_n H_n(Y) \right\rangle = \sum_{n=0}^{\infty} C_n^2 \langle H_n(X), H_n(Y) \rangle$$
$$= C_0^2 + C_1^2 \rho \leq \frac{1}{4} + \frac{1}{4}\sigma^2 \rho,$$

where the last inequality used the fact that $0 \leq \sigma^2 \leq 1$. This completes the proof. □

2.5 Proof of Lemma 3

We need the following lemma.

Lemma 8. *Under the conditions of Theorem 1, it holds that*

$$\max_{1 \leq i \leq n} \left| \frac{f(X_i^\top \beta)(1 - f(X_i^\top \beta))}{f(X_i^\top \hat{\beta})(1 - f(X_i^\top \hat{\beta}))} - 1 \right| = o(1),$$

with probability at least $1 - p^{-c} - n^{-c}$.

By Lemma 8, with probability at least $1 - p^{-c} - n^{-c}$,

$$0.99 \leq \frac{f(X_i^\top \beta)(1 - f(X_i^\top \beta))}{f(X_i^\top \hat{\beta})(1 - f(X_i^\top \hat{\beta}))} \leq (1 + o(1)). \tag{2.8}$$

Under the above event, we have

$$v_j \leq \frac{(1+o(1))}{n} \sum_{i=1}^{n} \frac{f'^2(X_i^\top \hat{\beta})}{f(X_i^\top \hat{\beta})(1 - f(X_i^\top \hat{\beta}))} (\hat{u}^\top X_i)^2 \leq \frac{(1+o(1))}{n} \sum_{i=1}^{n} \frac{f'^2(X_i^\top \hat{\beta})}{f(X_i^\top \hat{\beta})(1 - f(X_i^\top \hat{\beta}))} (u^{*\top} X_i)^2$$

where the last inequality follows from the objective function for $\hat{u}$. By definition $u^* = \mathbf{\Gamma}_j^{-1}$, then with probability at least $1 - p^{-c}$, we have

$$v_j \leq (1 + o(1))\mathbf{\Gamma}_{jj}^{-1},$$

by conditioning on $\mathcal{D}_2$ and using the concentration inequality for independent subexponential random variables, as in the proof of (1.11). The lower bound argument focuses on $v_j \geq \frac{0.99}{n}\sum_{i=1}^n \frac{f'^2(X_i^\top\hat{\beta})}{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}(\hat{u}^\top X_i)^2$. Towards this end, we define a proof-facilitating optimization problem

$$\tilde{u} = \min_{u\in\mathbb{R}^p} u^\top\left[\frac{1}{n}\sum_{i=1}^n \frac{f'^2(X_i^\top\hat{\beta})}{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}X_iX_i^\top\right]u,$$

subject to

$$\left|e_j^\top\left[\frac{1}{n}\sum_{i=1}^n \frac{f'^2(X_i^\top\hat{\beta})}{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}X_iX_i^\top\right]u - 1\right| \leq \lambda_n. \tag{2.9}$$

Note that $\hat{u}$ satisfies the feasible set in (2.9). If we denote $\hat{\mathbf{\Gamma}} = \frac{1}{n}\sum_{i=1}^n \frac{f'^2(X_i^\top\hat{\beta})}{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}X_iX_i^\top$, then we have $\hat{u}^\top\hat{\mathbf{\Gamma}}\hat{u} \geq \tilde{u}^\top\hat{\mathbf{\Gamma}}\tilde{u} \geq \tilde{u}^\top\hat{\mathbf{\Gamma}}\tilde{u} + t((1-\lambda_n) - e_j^\top\hat{\mathbf{\Gamma}}\tilde{u})$ for any $t \geq 0$, where the last inequality follows from (2.9). Note that for a given $t \geq 0$, we have $\tilde{u}^\top\hat{\mathbf{\Gamma}}\tilde{u} + t((1-\lambda_n) - e_j^\top\hat{\mathbf{\Gamma}}\tilde{u}) \geq \min_{u\in\mathbb{R}^p} u^\top\hat{\mathbf{\Gamma}}u + t((1-\lambda_n) - e_j^\top\hat{\mathbf{\Gamma}}u)$. By solving the right-hand side minimization problem, we have the minimizer u_0 satisfies $\hat{\mathbf{\Gamma}}u_0 = t\hat{\mathbf{\Gamma}}e_j/2$ and hence $\min_{u\in\mathbb{R}^p} u^\top\hat{\mathbf{\Gamma}}u + t((1-\lambda_n) - e_j^\top\hat{\mathbf{\Gamma}}u) = -\frac{t^2}{4}e_j^\top\hat{\mathbf{\Gamma}}e_j + t(1-\lambda_n)$. Combining the above arguments, we have $\hat{u}^\top\hat{\mathbf{\Gamma}}\hat{u} \geq \max_{t\geq 0}\big[-\frac{t^2}{4}e_j^\top\hat{\mathbf{\Gamma}}e_j + t(1-\lambda_n)\big]$. The minimum of the above function is achieved at $t^* = \frac{2(1-\lambda_n)}{e_j^\top\hat{\mathbf{\Gamma}}e_j} > 0$, which means $\hat{u}^\top\hat{\mathbf{\Gamma}}\hat{u} \geq \frac{(1-\lambda_n)^2}{e_j^\top\hat{\mathbf{\Gamma}}e_j}$. The boundedness of $e_j^\top\hat{\mathbf{\Gamma}}e_j$ follows from the concentration inequality of subexponential random variables and the boundedness of the spectrum of $\mathbf{\Gamma}$. As a result, we have, for sufficiently large (n,p), with high probability $\hat{u}^\top\hat{\mathbf{\Gamma}}\hat{u} \geq \frac{0.99}{\lambda_{\max}(\mathbf{\Gamma})}$. This completes the first part of Lemma 3.

To prove the second statement, note that,

$$|\mathbf{\Gamma}_{jj}^{-1}/(\mathbf{\Gamma}^*)_{jj}^{-1} - 1| \leq \frac{1}{(\mathbf{\Gamma}^*)_{jj}^{-1}}\|(\mathbf{\Gamma}^*)^{-1} - \mathbf{\Gamma}^{-1}\| \leq \|(\mathbf{\Gamma}^*)^{-1}\|\|\mathbf{\Gamma}^{-1}\|\|\mathbf{\Gamma}^*\|\|\mathbf{\Gamma}^* - \mathbf{\Gamma}\|.$$

It suffices to show $\|\mathbf{\Gamma}^* - \mathbf{\Gamma}\| \to o(1)$. For the case of logistic link function, by Lemma 8, we have

$$\begin{aligned}
\|\mathbf{\Gamma}^* - \mathbf{\Gamma}\| &\leq \sup_{\|w\|_2=1}\mathbb{E}\left[\left|\frac{f(X_i^\top\hat{\beta})(1-f(X_i^\top\hat{\beta}))}{f(X_i^\top\beta)(1-f(X_i^\top\beta))} - 1\right|f(X_i^\top\beta)(1-f(X_i^\top\beta))(w^\top X_i)^2\right] \\
&\leq o(1)\cdot\left\|\mathbb{E}\left[f(X_i^\top\beta)(1-f(X_i^\top\beta))X_iX_i^\top\right]\right\| + o(1)\cdot(\|\mathbf{\Gamma}^*\| + \|\mathbf{\Gamma}\|) \\
&= o(1).
\end{aligned}$$

For general link functions satisfying (L1) to (L4), one can show in the same manner that $\|\mathbf{\Gamma}^* - \mathbf{\Gamma}\| \to o(1)$. This proves $|\mathbf{\Gamma}_{jj}^{-1}/(\mathbf{\Gamma}^*)_{jj}^{-1} - 1| = o(1)$.

2.6 Proof of Lemma 4

We proceed with the following two steps.

1. We show that for any b such that $\|b\|_2 = 1$,

$$\mathbb{E}g_b(X_i^\top b) \geq c/2, \tag{2.10}$$

   for some universal constant $c > 0$.

2. For the random variable $Z(t) = \sup_{b\in\mathcal{C}(\xi,S),\|b\|_2=1,\|b\|_1=t} |\mathbb{P}_n[g_b(x)] - \mathbb{E}[g_b(x)]|$, it holds that

$$P\bigg(Z(t) \geq c/4 + C\sqrt{\frac{\log p}{n}}t\bigg) \leq c_1 \exp(-c_2 n - c_3 t^2 \log p). \tag{2.11}$$

The above two results (2.10) and (2.11) imply that $P(\mathcal{E}(t)) \leq c_1 \exp(-c_2 n - c_3 t \log p)$.

Proof of (2.10), Note that on the set $\|b\|_2 = 1$, $\mathbb{E}\min\{|X_i^\top b|, (X_i^\top b)^2\} \geq c > 0$. Then it suffices to show that

$$\mathbb{E}[\min\{|X_i^\top b|, (X_i^\top b)^2\} - g_b(X_i^\top b)] \leq c/2. \tag{2.12}$$

Since under the events $|X_i^\top \beta| < T$ and $|X_i^\top b| < T$, we have $g_b(X_i^\top b) = \min\{|X_i^\top b|, (X_i^\top b)^2\} \geq 0$, by union bound we can write

$$\begin{aligned}\mathbb{E}[\min\{|X_i^\top b|, (X_i^\top b)^2\} - g_b(X_i^\top b)] \leq\ & \mathbb{E}[\min\{|X_i^\top b|, (X_i^\top b)^2\} \cdot 1\{|X_i^\top \beta| \geq T\}] \\ & + \mathbb{E}[\min\{|X_i^\top b|, (X_i^\top b)^2\} \cdot 1\{|X_i^\top b| \geq T\}].\end{aligned}$$

Now by Cauchy-Schwarz inequality

$$\begin{aligned}\mathbb{E}[\min\{|X_i^\top b|, (X_i^\top b)^2\} \cdot 1\{|X_i^\top \beta^*| \geq T\}] &\leq \sqrt{\mathbb{E}|X_i^\top b|^2} \cdot P^{1/2}(|X_i^\top \beta| \geq T) \\ &\leq c_0 \exp\{-c_1 T^2\},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[\min\{|X_i^\top b|, (X_i^\top b)^2\} \cdot 1\{|X_i^\top b| \geq T\}] &\leq \sqrt{\mathbb{E}|X_i^\top b|^2} \cdot P^{1/2}(|X_i^\top b| \geq T) \\ &\leq c_0 \exp\{-c_2 T^2\}.\end{aligned}$$

Therefore, by choosing T sufficiently large, we can always guarantee that the right-hand side of the above inequalities is bounded by $c/2$, which implies (2.12).

Proof of (2.11). For any unit vector b, we have $\|g_b(x)\|_\infty \le T^2$. Therefore, we can apply the Azuma-Hoeffding inequality to obtain

$$P(Z(t) \ge \mathbb{E}[Z(t)] + z) \le 2\exp(-Cnz^2)$$

for some constant $C > 0$ and any $z > 0$. Set $z^*(t) = \frac{c}{4} + c't\sqrt{\log p/n}$, we have

$$P(Z(t) \ge \mathbb{E}[Z(t)] + z^*(t)) \le 2\exp(-C_1 n - C_2 t^2 \log p).$$

Comparing the above inequality with (2.11), we can observe that it suffices to show

$$\mathbb{E}[Z(t)] \lesssim t\sqrt{\frac{\log p}{n}}. \tag{2.13}$$

Let $\{\epsilon_i\}_{i=1}^n$ be a sequence of *i.i.d.* Rademacher random variables. Then a standard symmetrization argument (Ledoux and Talagrand, 2013) yields

$$\mathbb{E}[Z(t)] \le 2\mathbb{E}_{x,\epsilon}\left[\sup_{\|b\|_2=1,\|b\|_1=t}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g_b(X_i))\right|\right] = 2\mathbb{E}_{x,\epsilon}\left[\sup_{\|b\|_2=1,\|b\|_1=t}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \varphi_T(f_b(X_i))\right|\right].$$

Now since by construction the function φ_T is Lipschitz with parameter at most $2T$, and $\varphi_T(0) = 0$. Therefore, by the Ledoux-Talagrand contraction inequality (see p.112 of Ledoux and Talagrand (2013)), we have

$$\begin{aligned}
\mathbb{E}[Z(t)] &\le C\mathbb{E}_{x,\epsilon}\left[\sup_{\|b\|_2=1,\|b\|_1=t}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_b(X_i)\right|\right] \\
&\le C\mathbb{E}_{x,\epsilon}\left[\sup_{\|b\|_2=1,\|b\|_1=t}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i^\top b \cdot 1\{|X_i^\top \beta| < T\}\right|\right] \\
&\le Ct\mathbb{E}_{x,\epsilon}\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i \cdot 1\{|X_i^\top \beta| < T\}\right\|_\infty
\end{aligned}$$

where the last inequality follows from Hölder's inequality. Finally, note that the random variables $\epsilon_i X_i \cdot 1\{|X_i^\top \beta^*| < T\}$ are centred independent subgaussian random vectors. We can then apply standard bounds for subgaussian maxima (e.g. see p.97 of Ledoux and Talagrand (2013)) to get

$$\mathbb{E}_{x,\epsilon}\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i \cdot 1\{|X_i^\top \beta^*| < T\}\right\|_\infty \lesssim \sqrt{\frac{\log p}{n}}.$$

This shows (2.13) can therefore completes our proof. □

2.7 Proof of Lemma 6

For $x \ge 0$, we claim that $0 \le 2\Phi(x) - 1 \le x$. To see this, the function $l(x) = 2\Phi(x) - 1 - x$ has derivative $l'(x) = 2\phi(x) - 1 < 0$ for all $x \ge 0$ so that the maximum of $l(x)$ is attained at $l(0) = 0$. The case of $x < 0$ can be proved similarly. □

2.8 Proof of Lemma 7

We define $g = -\ell'(\hat{\beta})$. A vector $\hat{\beta}$ is a global minimizer in $\arg\min_\beta \left\{\ell(\beta) + \lambda\|\beta\|_1\right\}$ if and only if the negative gradient at $\hat{\beta}$ satisfies the Karush-Kuhn-Tucker (KKT) conditions

$$\begin{cases} g_j = \lambda \mathrm{sgn}(\hat{\beta}_j) & \text{if } \hat{\beta}_j \neq 0 \\ g_j \in \lambda[-1,1] & \text{all } j \end{cases}.$$

Since $\ell'(\hat{\beta}) - \ell'(\beta^*) = -g - \ell'(\beta^*)$, we have

$$\begin{aligned} \langle \hat{\beta} - \beta^*, \ell'(\hat{\beta}) - \ell'(\beta^*)\rangle &= \langle \hat{\beta}, -\ell'(\beta^*)\rangle - \langle \beta^*, -\ell'(\beta^*) - g\rangle - \langle \hat{\beta}, g\rangle \\ &= \langle \hat{\beta}, -\ell'(\beta^*)\rangle - \langle \beta^*, -\ell'(\beta^*) - g\rangle - \lambda\|\hat{\beta}\|_1 \end{aligned}$$

Let $h = \hat{\beta} - \beta^*$. We have $h_{S^c} = \hat{\beta}_{S^c}$ and $\beta^*_{S^c} = 0$. Thus

$$\begin{aligned} \Delta(\hat{\beta}, \beta^*) &= \langle \hat{\beta}_{S^c}, \{-\ell'(\beta^*)\}_{S^c}\rangle - \lambda\|\hat{\beta}_{S^c}\|_1 - \langle h_S, \{-\ell'(\beta^*) - g\}_S\rangle \\ &\leq \|\hat{\beta}_{S^c}\|_1(z^* - \lambda) + \langle h_S, g_S + \{\ell'(\beta^*)\}_S\rangle \\ &\leq \|\hat{\beta}_{S^c}\|_1(z^* - \lambda) + \|h_S\|_1(z^* + \lambda). \end{aligned}$$

This gives (2.1). Since by the convexity of $\ell(\beta)$, we have $\Delta(\hat{\beta}, \beta^*) > 0$, then $h \in \mathcal{C}(\xi, S)$ when $(\lambda + z^*)/(\lambda - z^*) \leq \xi$. For $j \notin S$, $h_j(\ell'(\beta + h) - \ell'(\beta))_j = \hat{\beta}_j(-\ell'(\beta^*) - g)_j \leq |\hat{\beta}_j|(\lambda - g_j) \leq 0$. □

2.9 Proof of Lemma 8

By (L1), it suffices to show that, with high probability,

$$\max_{1\leq i\leq n}\left|\frac{f(|X_i^\top\beta|)}{f(|X_i^\top\hat{\beta}|)} - 1\right| = o(1), \qquad \max_{1\leq i\leq n}\left|\frac{1 - f(|X_i^\top\beta|)}{1 - f(|X_i^\top\hat{\beta}|)} - 1\right| = o(1). \tag{2.14}$$

Note that, by Taylor expansion,

$$f(|X_i^\top\beta|) \leq f(|X_i^\top\hat{\beta}|) + f'(a^*)|X_i^\top(\hat{\beta} - \beta)|,$$

where a^* is between $|X_i^\top\beta|$ and $|X_i^\top\hat{\beta}|$. By (L2), we have, with probability at least $1 - p^{-c} - n^{-c}$,

$$\max_{1\leq i\leq n}\left|\frac{f(|X_i^\top\beta|)}{f(|X_i^\top\hat{\beta}|)} - 1\right| \lesssim \max_{1\leq i\leq n}|X_i^\top w'|\sqrt{\frac{k\log p}{n}} \leq \tau_n\sqrt{\frac{k\log p}{n}}.$$

Therefore, the first inequality of (2.14) holds for $k \ll \frac{n}{\tau_n^2\log p}$.

Similarly, since $1 - f(|X_i^\top\beta|) = f(-|X_i^\top\beta|)$, we have

$$\max_{1\leq i\leq n}\left|\frac{1 - f(|X_i^\top\beta|)}{1 - f(|X_i^\top\hat{\beta}|)} - 1\right| \leq \max_{1\leq i\leq n}\left|\frac{f(-|X_i^\top\beta|)}{f(-|X_i^\top\hat{\beta}|)} - 1\right| \leq f'(a^*)|X_i^\top(\hat{\beta} - \beta)|,$$

which leads to the second inequality in (2.14) by following the same argument. □

3 Properties of Some Link Functions

3.1 Probit Link

The Conditions (L1) to (L3) in the main paper follows from the properties of the Gaussian distribution, such as the inequalities (3.1) and (3.2) below. In the following, we show condition (L4) hold.

Let $h(x;y)$ be defined in (L4) such that $\ell_f''(\beta) = n^{-1}\sum_{i=1}^n h(X_i^\top\beta; y_i)X_iX_i^\top$. Note that for $y_i = 0$, we have $h(x;0) = \frac{f''(x)(1-f(x))+f'(x)^2}{(1-f(x))^2} = \frac{-\phi(x)x(1-\Phi(x))+\phi^2(x)}{(1-\Phi(x))^2}$; for $y_i = 1$, we have $h(x;1) = \frac{\phi^2(x)+\Phi(x)\phi(x)x}{\Phi^2(x)}$.

Lemma 9. *Let* $g_1(x) = h(x;0) = \frac{-\phi(x)x(1-\Phi(x))+\phi^2(x)}{(1-\Phi(x))^2}$ *and* $g_2(x) = h(x;1) = \frac{\phi^2(x)+\Phi(x)\phi(x)x}{\Phi^2(x)}$. *It holds that*

$$|(\log g_\ell)'(x^*)| \le C|x^*| \le C(|x| + |t|), \qquad \ell = 1, 2,$$

for all $x, t \in \mathbb{R}$, *and some* x^* *between* x *and* $x + t$.

Proof. Since

$$\frac{x}{x^2+1}\phi(x) \le 1 - \Phi(x) \le \frac{1}{x}\phi(x), \tag{3.1}$$

we have

$$\phi(x) \le \phi(x)/\Phi(x) \le \phi(x)/\Phi(-|x|) \le \frac{x^2+1}{|x|}.$$

Then

$$\begin{aligned}(\log g_2)' &= \frac{\Phi\phi - \phi^2 x - \Phi\phi^2 x^2}{\phi(\phi + \Phi x)} - \frac{2\phi}{\Phi} \\ &\le |\frac{\phi x}{\phi + \Phi x}| + |\frac{\Phi\phi x^2}{\phi + \Phi x}| + |\frac{\Phi}{\phi + \Phi x}| + |\phi/\Phi|\end{aligned}$$

Since $\frac{\phi x}{\phi+\Phi x} = \frac{x\phi/\Phi}{\phi/\Phi+x}$, and the function $\frac{xy}{x+y}$ is increasing in x and y, we have

$$\frac{\phi x}{\phi + \Phi x} \le \frac{(x^2+1)|x|}{2x^2+1} \le C|x|.$$

The function $\frac{\Phi\phi x^2}{\phi+\Phi x}$ is increasing in x for $x > 0$ and decreasing in x for $x < 0$. Then since $\phi(x) > \Phi(-|x|)|x|$ or more precisely

$$\Phi(-|x|) \le \frac{2}{|x| + \sqrt{x^2 + 8/\pi}}\phi(x), \tag{3.2}$$

we have

$$\frac{\Phi\phi x^2}{\phi + \Phi x} \le \max\{\frac{\Phi(|x|)\phi x^2}{\phi + \Phi(|x|)|x|}, \frac{\Phi(-|x|)\phi x^2}{\phi - \Phi(-|x|)|x|}\} \le \max\{\phi|x|, \frac{\phi|x|}{1 - \frac{2|x|}{|x|+\sqrt{x^2+8/\pi}}}\} \le \phi|x|^3 \le C,$$

since $\sqrt{x^2+8/\pi}-|x| \geq c/|x|$ for $|x| > \delta$. The function $\frac{\Phi}{\phi+\Phi x} = \frac{1}{\phi/\Phi+x}$ and $\frac{1}{x+y}$ is decreasing in x and y. When $x > 0$, $\frac{\Phi}{\phi+\Phi x} \leq \min\{1/x, \phi^{-1}\} \leq C$; when $x < 0$,

$$\frac{\Phi(-|x|)}{\phi - \Phi(-|x|)|x|} \leq \frac{\Phi(-|x|)}{\phi(1 - \frac{2|x|}{|x|+\sqrt{x^2+8/\pi}})} \leq C|x|.$$

Hence, we have shown that

$$(\log g_2)' \leq C|x|.$$

Similarly, we can also show that $(\log g_1)' \leq C|x|$. □

Thus, we obtain the Lipschitz condition

$$e^{-2x^2-2t^2} \leq e^{-|xt|-t^2} \leq \frac{h(x+t)}{h(x)} \leq e^{|xt|+t^2} \leq e^{2|x|^2+2t^2}.$$

3.2 CDFs of Student's t_ν-distributions

Let $f_\nu(x)$ be the cdf of Student's t_ν-distribution. Conditions (L1) and that $f_\nu(x) \leq \Phi(x)$ for any $x \geq 0$ hold by definition of the t_v-distribution.

To obtain other properties, we need the following upper probability tail bound for the t_ν-distribution. Specifically, if $X \sim t_\nu$ for some $\nu \geq 1$, we have, for sufficiently large x,

$$P(X \geq x) = \int_x^\infty \frac{c}{(1+t^2/\nu)^{(\nu+1)/2}} dt \asymp \int_x^\infty \frac{1}{t^{\nu+1}} dt \asymp x^{-\nu}.$$

In other words, we have

$$f_\nu(-|x|) \asymp \frac{1}{x^\nu} \tag{3.3}$$

Now for any given $\nu \geq 1$, we have

$$\lim_{x\to+\infty} \frac{f'_\nu(x)}{1-f_\nu(x)} \leq C \lim_{x\to+\infty} \frac{x^{-(\nu+1)}}{x^{-\nu}} = 0, \tag{3.4}$$

and whenever $\nu \geq 1$, we have

$$x^2 f'(x) \lesssim x^{-\nu+1} = O(1). \tag{3.5}$$

Thus condition (L2) holds for cdfs of t_ν-distribution where $\nu \geq 1$.

To show (L3), we note that for any given $\nu \geq 1$,

$$f''_\nu(x) = -Cx(1+x^2/\nu)^{-\frac{\nu+3}{2}} \asymp -x \cdot |x|^{-\nu-3}.$$

Then, for any constant w,

$$|x f''_\nu(x+w)| \lesssim |x(x+w)| \cdot (1+(x+w)^2/\nu)^{-\frac{\nu+3}{2}} \lesssim x^{-\nu-1} = O(1),$$

which proves the condition (L3).

As for condition (L4), since for $y_i = 0$, we have $h(x;0) = \frac{f''(x)(1-f(x))+f'(x)^2}{(1-f(x))^2}$ whereas for $y_i = 1$, we have $h(x;1) = \frac{f'(x)^2-f''(x)f(x)}{f^2(x)}$, then if we denote $g_1(x) = h(x;0)$, it follows that,

$$(\log g_1)'(x) = \frac{g_1'(x)}{g_1(x)} \leq \frac{(1-f)^2}{f''(1-f)+f'^2} \cdot \left[\frac{|f'''|}{1-f} + \frac{f'f''}{(1-f)^2} + \frac{|f'^3|}{(1-f)^3}\right].$$

Now since

$$\begin{aligned}
&\lim_{x\to+\infty} \frac{(1-f)^2}{|f''(1-f)+f'^2|} \cdot \left[\frac{|f'''|}{1-f} + \frac{f'f''}{(1-f)^2} + \frac{|f'^3|}{(1-f)^3}\right] \\
&= \lim_{x\to+\infty} x^{-\nu+2} \cdot \left[x^{-\nu-3}x^{\nu} + x^{2\nu} \cdot x^{-\nu-1} \cdot x^{-\nu-2} + x^{3\nu} \cdot x^{-3\nu-3}\right] \\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{x\to-\infty} \frac{(1-f)^2}{|f''(1-f)+f'^2|} \cdot \left[\frac{|f'''|}{1-f} + \frac{f'f''}{(1-f)^2} + \frac{|f'^3|}{(1-f)^3}\right] \\
&= \lim_{x\to-\infty} x^{\nu+2} \cdot \left[x^{-\nu-3} + x^{-\nu-1} \cdot x^{-\nu-2} + x^{-3\nu-3}\right] \\
&= 0.
\end{aligned}$$

This proves that

$$(\log g_1)'(x) = O(1).$$

Similarly, we can also show that $(\log g_2)'(x) = O(1)$. This completes the proof of condition (L4).

As for Condition (L5), note that by (3.3), the tail of the link function $f_\nu(x)$ has order $x^{-\nu}$, which is always larger than that of the logistic link function, which is of order e^{-x}.

3.3 Generalized Logistic Function

Condition (L1) and the fact that $f(x) \leq \Phi(Cx)$ for some $C > 0$ holds trivially.

Hereafter, without loss of generality, we assume $\varphi > 0$. Note that

$$f'(x) = \frac{\gamma\varphi}{2} \tanh^{\gamma-1}(\varphi x) \cosh^{-2}(\varphi x).$$

We have, for $\gamma \geq 1$,

$$\lim_{x\to+\infty} \frac{f'(x)}{1-f(x)} = C \lim_{x\to+\infty} \frac{\cosh^{-2}(\varphi x)}{1-\tanh^{\gamma}(\varphi x)} \leq C \lim_{x\to+\infty} \frac{e^{-2\varphi x}}{1-\tanh(\varphi x)} = C.$$

In addition, since for $x \geq 0$ and $\gamma \geq 1$,

$$x^3(1-\tanh^{\gamma}(x)) \leq x^3(1-\tanh(x)) \to 0,$$

the condition (L3) holds.

As for condition (L3), it suffices to note that

$$f''(x) \lesssim \tanh^{\gamma-2}(\varphi x)\cosh^{-4}(\varphi x) + \frac{\tanh^{\gamma-1}(\varphi x)\sinh(\varphi x)}{\cosh^3(\varphi x)},$$

which goes to 0 with an exponential rate, faster than $1/x$, as $x \to +\infty$.

Lastely, for condition (L4), using the similar notations as the previous cases, since

$$\begin{aligned}
&\lim_{x\to+\infty} \frac{(1-f)^2}{|f''(1-f)+f'^2|}\cdot\left[\frac{|f'''|}{1-f} + \frac{f'f''}{(1-f)^2} + \frac{|f'^3|}{(1-f)^3}\right] \\
&= \lim_{x\to+\infty} \frac{e^{-4\varphi x}}{e^{-6\varphi x}+e^{-4\varphi x}}\cdot\left[e^{-2\varphi x} + e^{-2\varphi x}e^{-4\varphi x}e^{4\varphi x} + 1\right] \\
&= 1
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{x\to-\infty} \frac{(1-f)^2}{|f''(1-f)+f'^2|}\cdot\left[\frac{|f'''|}{1-f} + \frac{f'f''}{(1-f)^2} + \frac{|f'^3|}{(1-f)^3}\right] \\
&= \lim_{x\to-\infty} e^{4\varphi x}\cdot\left[e^{-6\varphi x} + e^{-6\varphi x} + e^{-6\varphi x}\right] \\
&= 0
\end{aligned}$$

as long as $\nu \geq 1$. This proves that

$$(\log g_1)'(x) = O(1).$$

Similarly, we can also show that $(\log g_2)'(x) = O(1)$. This completes the proof of condition (L4).

Finally, the condition (L5) follows trivially for $\gamma \geq 1$.

4 Supplements to Section 5 of the Main Paper

In Section 5 of our main paper, we carried out simulations that compare different methods. The design covariates were generated from a multivariate Gaussian distribution, whose covariance matrix is a blockwise diagonal matrix of 10 identical unit diagonal Toeplitz matrices as follows

$$\boldsymbol{\Sigma}_M = \begin{bmatrix}
1 & \frac{6(p-2)}{10(p-1)} & \frac{6(p-3)}{10(p-1)} & \cdots & \frac{6}{10(p-1)} & 0 \\
\frac{6(p-2)}{10(p-1)} & 1 & \frac{6(p-2)}{10(p-1)} & \cdots & \frac{12}{10(p-1)} & \frac{6}{10(p-1)} \\
\vdots & & \ddots & & & \\
0 & \frac{6}{10(p-1)} & \frac{12}{10(p-1)} & \cdots & \frac{6(p-2)}{10(p-1)} & 1
\end{bmatrix}.$$

Table S4.1: Empirical performances of CIs for β_{100} under $\mathbf{\Sigma} = 0.02 \cdot \mathbf{I}_p$, $\psi = 1$, $\alpha = 0.05$ and $n = 400$

p	Coverage (%)					Length				
	LSW	wlp	lproj	rproj	rose	LSW	wlp	lproj	rproj	rose
	$k = 20$									
400	94.9	95.5	97.7	95.5	78.5	2.80	2.77	2.81	1.60	1.95
700	97.1	97.1	99.3	95.8	82.9	2.61	2.77	2.79	2.24	1.94
1000	97.6	94.3	99.0	96.8	85.8	2.79	2.77	2.81	2.52	1.95
1300	95.6	94.0	99.0	96.2	79.1	2.69	2.77	2.79	2.65	1.93
	$k = 25$									
400	95.3	95.0	97.2	96.2	80.8	2.81	2.78	2.82	1.60	1.95
700	95.0	94.1	98.1	96.2	79.8	2.62	2.77	2.81	2.24	1.94
1000	96.6	95.3	97.8	97.2	84.2	2.65	2.77	2.79	2.51	1.92
1300	95.9	94.6	99.3	95.9	78.6	2.85	2.78	2.80	2.65	1.94
	$k = 30$									
400	96.3	96.0	98.1	96.0	80.6	2.80	2.78	2.82	1.60	1.95
700	95.7	96.0	98.7	96.6	81.5	2.62	2.77	2.81	2.24	1.94
1000	95.6	95.0	98.7	97.2	84.2	2.79	2.77	2.80	2.53	1.94
1300	97.6	94.0	99.6	96.9	79.6	2.70	2.78	2.79	2.65	1.94
	$k = 35$									
400	97.0	95.0	97.3	96.2	82.3	2.81	2.77	2.82	1.60	1.95
700	97.0	96.2	96.8	94.7	78.6	2.63	2.77	2.80	2.25	1.93
1000	96.6	94.3	99.1	96.2	81.2	2.78	2.78	2.87	2.52	1.94
1300	95.9	95.3	97.9	94.5	81.3	2.69	2.77	2.81	2.66	1.95

4.1 Additional Simulations for CIs under Logistic Regression

In Tables S4.1 and S4.2, we show some additional simulations results under the settings considered in Section 5.1 of the main paper. The CIs for the nonzero coefficients are presented in the main paper, whereas in Tables S4.1 and S4.2 the coverage probabilities and average lengths of the 95% CIs for the zero coefficient β_{100} were reported. The values associated to the CIs were calculated based on 500 round of simulations.

4.2 Efficiency Loss of Sample Splitting

We compare the efficiency of the proposed methods with and without sample splitting. Specifically, we consider the setting described in Section 5.1 of the main paper with $\mathbf{\Sigma} = 0.02\mathbf{I}_p$, $n = 400$, $\psi = 1$, and let p vary from 1400 to 1700. For the sample splitting

Table S4.2: Empirical performances of CIs for β_{100} under $\mathbf{\Sigma} = \mathbf{\Sigma}_M$, $\psi = 0.5$, $\alpha = 0.05$ and $n = 400$

p	Coverage (%)					Length				
	LSW	wlp	lproj	rproj	rose	LSW	wlp	lproj	rproj	rose
	$k = 20$									
400	95.9	98.9	90.1	97.4	68.7	1.10	1.55	0.54	1.01	0.32
700	98.9	100.0	94.4	98.2	69.5	1.21	1.34	0.54	0.90	0.30
1000	98.2	100.0	97.4	97.4	64.9	1.14	1.37	0.52	0.76	0.30
1300	99.2	100.0	99.7	99.4	66.4	1.01	1.26	0.54	0.76	0.35
	$k = 25$									
400	97.6	99.7	89.9	97.8	54.1	1.14	1.62	0.55	1.01	0.33
700	98.9	99.2	96.2	98.1	44.5	1.24	1.33	0.54	0.91	0.31
1000	98.6	99.4	98.1	98.2	42.9	1.15	1.40	0.52	0.77	0.31
1300	99.2	100.0	98.6	98.6	50.1	1.02	1.33	0.54	0.76	0.35
	$k = 30$									
400	97.9	99.7	91.6	97.9	71.4	1.14	1.56	0.56	1.31	0.33
700	98.9	99.7	97.6	98.7	69.3	1.22	1.28	0.54	0.91	0.31
1000	98.5	99.2	97.5	97.5	63.4	1.25	1.35	0.52	0.76	0.31
1300	98.7	100.0	99.4	98.9	64.6	1.09	1.27	0.54	0.76	0.35
	$k = 35$									
400	96.2	100.0	92.3	98.4	65.0	1.14	1.56	0.55	1.07	0.32
700	98.9	100.0	96.0	98.9	46.2	1.23	1.64	0.54	0.91	0.31
1000	98.2	99.6	96.2	98.4	52.1	1.19	1.41	0.53	0.77	0.31
1300	99.3	100.0	98.9	99.2	58.7	1.02	1.31	0.53	0.77	0.35

procedure, we randomly select $\xi = 25\%, 50\%, 75\%$ or 90% of the samples, corresponding to $\mathcal{D}_1$, for the bias correction step, and use the rest of the samples, corresponding to $\mathcal{D}_2$ to obtain the initial Lasso estimator $\hat{\beta}$. We denote the proposed methods with data splitting proportion $\xi = 25\%, 50\%, 75\%$ and 90% as "lsw1," "lsw2," "lsw3," and "lsw4," respectively. Table S4.3 summarizes the empirical coverage probability and the averaged length of the CIs in different settings. Since all the methods achieved the correct coverage probability 95%, we may use the length of the CIs as an estimate of the efficiency of the methods. We observe that, the proposed method "LSW" without sample splitting is more efficient than its data-split counterparts across all the settings. The efficiency of "lowe" relative to its data-split counterparts is roughly $1/\sqrt{\xi}$, reflecting the information loss due to data splitting. Intuitively, the reason that a larger ξ in our simulation leads to a higher efficiency is that, the asymptotic normality is essentially established based on such samples for bias correction. Therefore, when data splitting is applied, as long as the samples used for the initial Lasso estimator are not too scarce (theoretically a non-diminishing proportion of the total samples), the more samples used for bias correction, the more efficient the proposed method is. In addition, in this specific setting, the method "rproj" is asymptotically equally efficient as "LSW," but it fails to cover the true coefficient in other settings (Table 2 of the main paper). In general, our numerical results suggest that, for practical purpose, one should apply the proposed method directly without data splitting for better performance.

4.3 Numerical Comparison with the Knockoff Method

This section includes some numerical results about the proposed multiple testing procedure and the knockoff method of Candès et al. (2018). We have remarked in the main paper that, although the knockoff approach does not require a pre-specified link function, it requires the distribution of the covariates to be known. In comparison, our approach does not require knowledge about the design distribution and can be applied to a large class of GLMs with binary outcomes. In the following, we consider the simulation settings in Section 5.1 of the main paper with $n = 400$, $p \in \{600, 800, 1000\}$, $k \in \{5, 8, 10\}$, and the signal strength $\psi = 1.5$. The random design matrix was generated from a centred multivariate Gaussian distribution with covariance $\mathbf{\Sigma} = \mathbf{I}_p$. In Table S4.4, the empirical powers and FDRs (with desired FDR level $\leq 10\%$) were calculated based on 500 rounds of simulations. The knockoff method was implemented using the `stat.lasso_coefdiff_bin` function and the `knockoff.filter` function in the R package `knockoff` with the default parameters. We observe that under the current setting, our testing procedure based on the debiased estimators is more powerful than the kockoff method for the sparser coefficient vectors; this is likely because our method takes advantage of the underlying sparse structure.

Table S4.3: Empirical performances with $\mathbf{\Sigma} = 0.02 \cdot \mathbf{I}_p$, $\psi = 1$, $\alpha = 0.05$ and $n = 400$

p	Coverage (%)						Length					
	LSW	lsw1	lsw2	lsw3	lsw4	rproj	LSW	lsw1	lsw2	lsw3	lsw4	rproj
						$k = 20$						
1400	96.3	93.5	94.0	94.7	93.5	95.7	2.71	4.45	3.73	3.17	2.89	2.68
1500	95.2	95.2	92.2	93.0	95.5	95.0	2.72	4.40	3.73	3.17	2.87	2.70
1600	94.7	95.7	93.7	94.0	93.5	93.2	2.72	4.37	3.70	3.17	2.89	2.72
1700	95.7	94.5	95.2	92.7	93.3	95.7	2.70	4.36	3.65	3.16	2.91	2.72
						$k = 25$						
1400	94.5	94.7	96.0	93.0	94.0	94.7	2.70	4.46	3.75	3.15	2.88	2.68
1500	95.0	96.0	94.2	94.0	94.3	95.5	2.69	4.43	3.70	3.14	2.90	2.69
1600	95.7	95.2	96.2	92.0	93.3	94.0	2.71	4.39	3.69	3.15	2.89	2.71
1700	94.5	94.2	93.7	91.5	94.0	94.2	2.73	4.39	3.64	3.18	2.90	2.73
						$k = 30$						
1400	95.7	94.7	95.7	94.2	92.5	95.2	2.69	4.45	3.77	3.14	2.89	2.67
1500	96.0	96.0	92.7	94.0	94.0	96.2	2.71	4.45	3.73	3.17	2.88	2.70
1600	96.5	95.5	93.2	94.2	93.5	95.5	2.70	4.39	3.69	3.17	2.89	2.71
1700	96.0	93.5	93.2	94.0	94.3	94.7	2.71	4.34	3.66	3.17	2.90	2.73

Table S4.4: Multiple testing results with $\mathbf{\Sigma} = \mathbf{I}_p$, $\psi = 1.5$, $n = 400$ and FDR $< 10\%$.

p	FDR (%)		Power	
	LSW	knockoff	LSW	knockoff
		$k = 5$		
600	1.26	1.60	0.99	0.03
800	1.02	2.62	0.92	0.05
1000	1.33	3.21	0.99	0.06
		$k = 8$		
600	0.56	5.69	1.00	0.21
800	0.11	7.38	0.98	0.31
1000	0.22	8.86	0.99	0.30
		$k = 10$		
600	0.10	8.02	1.00	1.00
800	0.10	10.08	0.98	0.99
1000	0.12	6.59	0.99	0.98

5 Analysis of Real Data using Alternative Methods

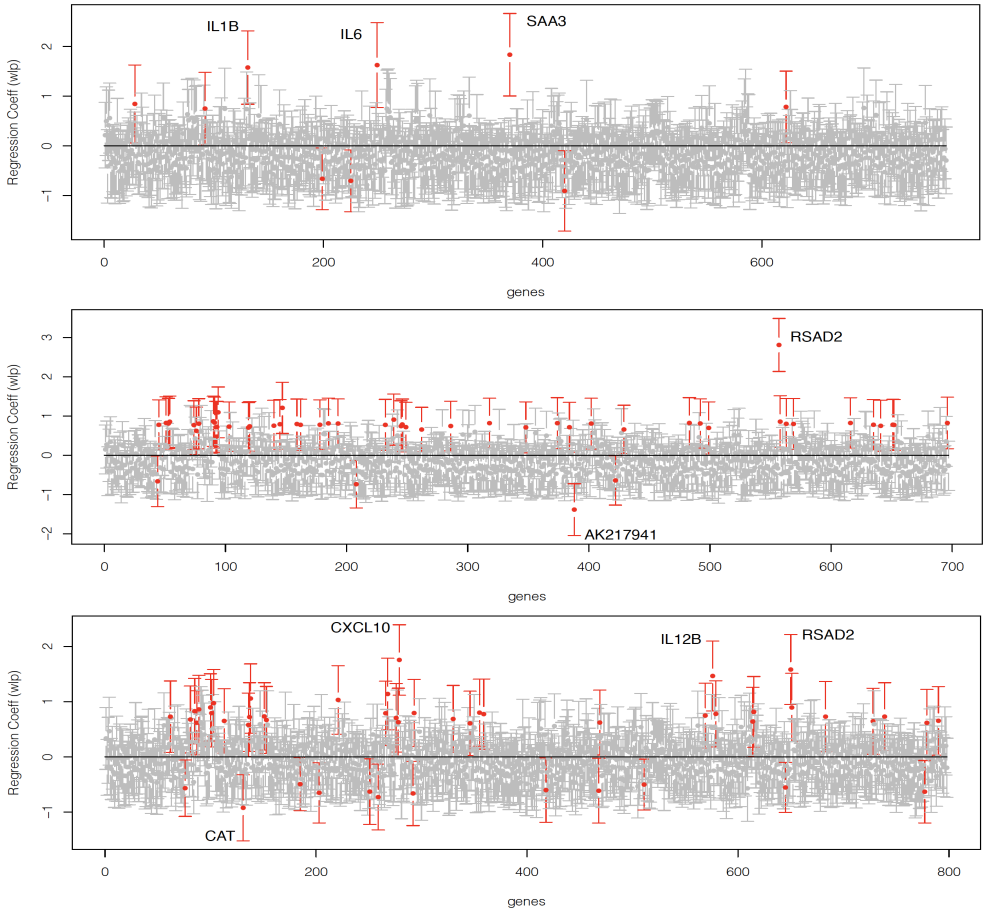

Figure S5.1: An illustration of the CIs produced by "wlp" corresponding to the stimulations by LPS (top), PIC (middle) and LPS (bottom), respectively. The CIs that do not cover zero are marked in red.

As a comparison, we applied other methods, considered in Section 5 of the main paper, to the real data set. The corresponding CIs are illustrated in Figures S5.1 to S5.3. In general, we observed that (i) "wlp" tends to produce shorter CIs, and therefore leads to a lot more CIs that do not cover zero, including the ones discovered by the proposed method, (ii) "lproj" identifies similar sets of genes whose CI does not cover zero, yet the CIs are much longer than the proposed method, and (iii) "rproj" also produces very long CIs, all of which cover zero. These results along with our cross reference with the existing literature concerning the genes identified by the proposed method suggest that the proposed CIs are effectively informative in analyzing such data set. However, more experimental and numerical evidences are needed as to determine which method produced the most precise

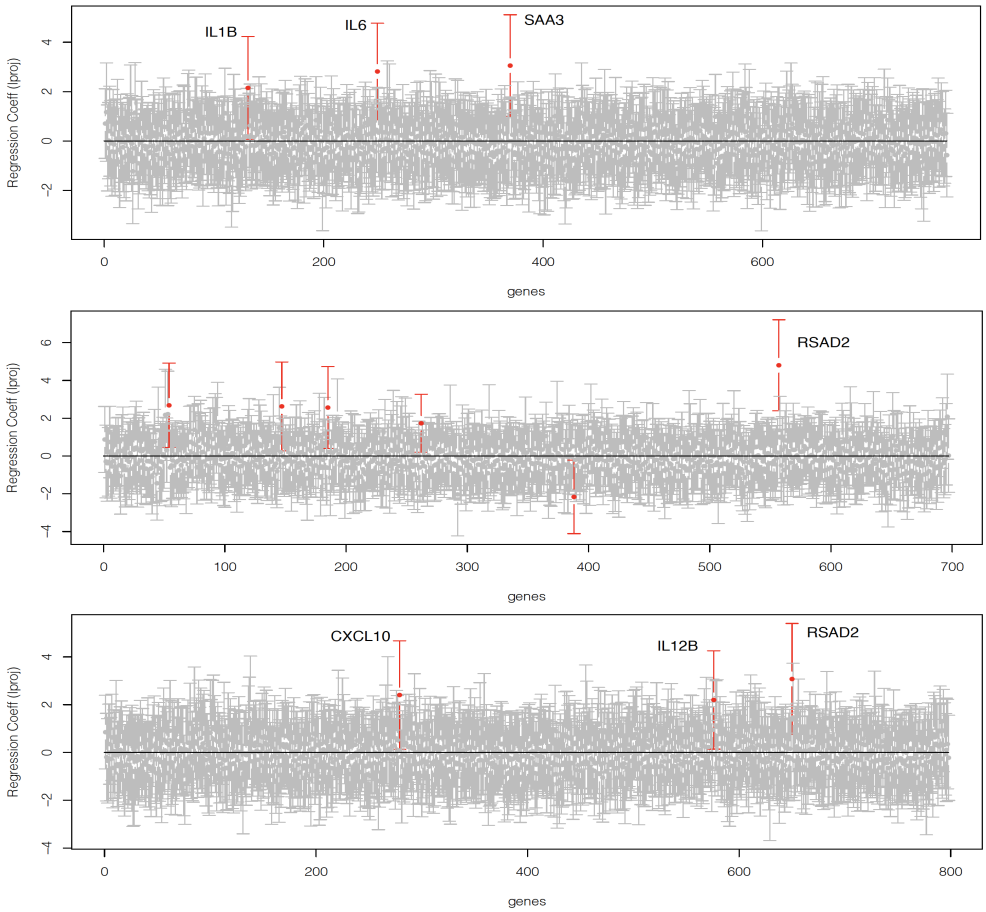

Figure S5.2: An illustration of the CIs produced by "lproj" corresponding to the stimulations by LPS (top), PIC (middle) and LPS (bottom), respectively. The CIs that do not cover zero are marked in red.

CIs.

6 Comparison with Ning and Cheng (2020)

Recently, a novel statistical framework for constructing sparse confidence sets was proposed by Ning and Cheng (2020). Comparing with this new approach to high-dimensional inference (hereafter referred as "sparse confidence sets"), our proposal has the following distinct features:

- In terms of the purpose, our proposed confidence intervals focus on a single regression coefficient, whereas the sparse confidence sets concern the whole regression vector. In particular, our proposal does not aim at controlling the overall false positive rate, but

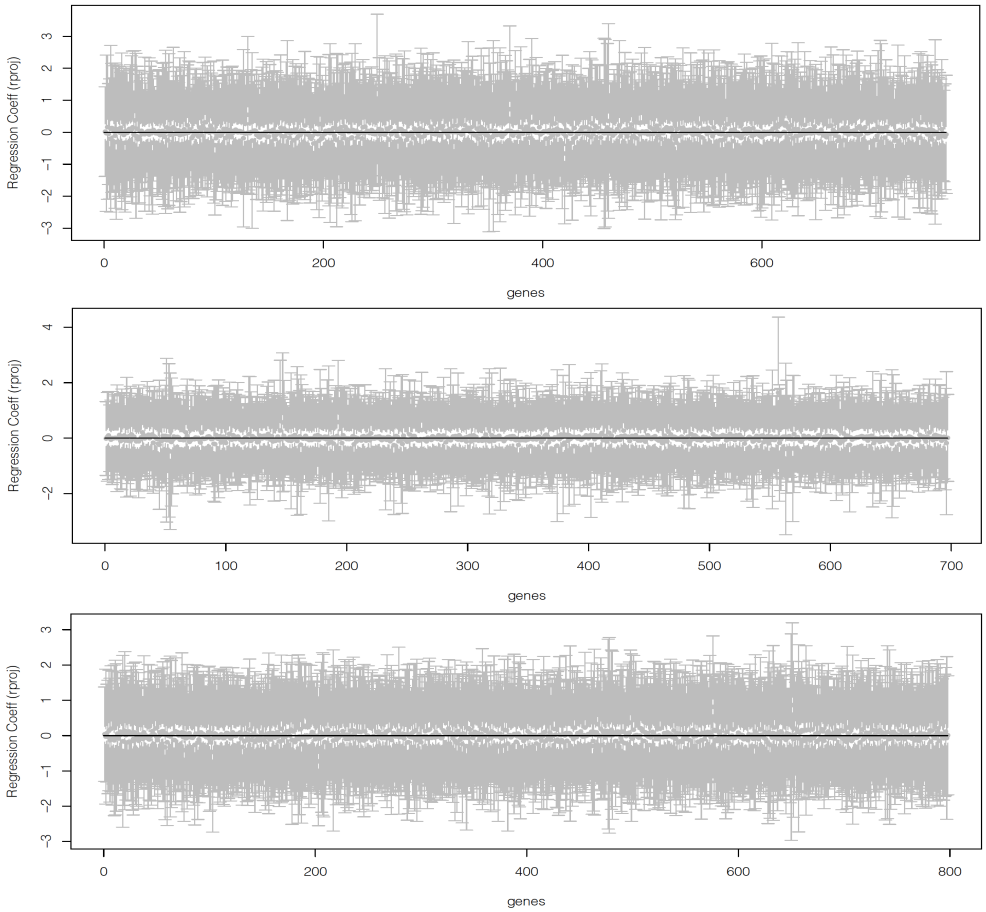

Figure S5.3: An illustration of the CIs produced by "rproj" corresponding to the stimulations by LPS (top), PIC (middle) and LPS (bottom), respectively.

could be modified properly (such as Bonferroni correction or thresholding) to achieve such goals. However, a detailed account of such an interesting issues is beyond the scope of the current paper.

- The validity of the sparse confidence sets relies on a certain "beta-min" condition, that is, the minimal absolute value of the nonzero coefficients has to be sufficiently large. See also Figure 1 of Ning and Cheng (2020) for numerical evidences. In contrast, our proposal does not require any minimum signal strength condition.

- Ning and Cheng (2020) focus on the multivariate Gaussian model with independent components and known variance, whereas the current paper concerns the more challenging high-dimensional logistic regression model with possibly correlated covariates. Therefore, direct numerical comparison between the two proposals is difficult, as the extension of their sparse confidence sets to the logistic regression model is itself highly

nontrivial. We leave this interesting problem for future investigation.

7 Derivation of the Influence Function

Define $Z_i = (X_{i,\cdot}^\intercal, y_i)$. For our outcome model, we have the joint density

$$p(Z_i, \beta) = p_\beta(y_i|X_{i,\cdot})P(X_{i,\cdot}) = [f(X_i^\intercal\beta)]^{y_i}[1 - f(X_i^\intercal\beta)]^{1-y_i}p_X(X_{i,\cdot}),$$

where $p_X(X_{i,\cdot})$ denotes the density of X_i. Then have the score function as

$$\begin{aligned} S_\beta(Z_i, \beta) = \frac{\partial \log p(Z_i, \beta)}{\partial \beta} &= f'(X_i^\intercal\beta)\left[\frac{y_i}{f(X_i^\intercal\beta)} - \frac{1 - y_i}{1 - f(X_i^\intercal\beta)}\right] X_i \\ &= \frac{f'(X_i^\intercal\beta)}{f(X_i^\intercal\beta)[1 - f(X_i^\intercal\beta)]} X_i \left[y_i - f(X_i^\intercal\beta)\right] \end{aligned} \tag{7.1}$$

Then we have

$$I(\beta) = \mathbb{E}S_\beta(Z_i, \beta)S_\beta^\intercal(Z_i, \beta) = \mathbb{E}\frac{[f'(X_i^\intercal\beta)]^2}{f(X_i^\intercal\beta)[1 - f(X_i^\intercal\beta)]} X_i X_i^\intercal$$

$$\begin{aligned} \psi^{\text{eff}}(Z_i) =& e_j^\intercal [I(\beta)]^{-1} S_\beta(Z_i, \beta) \\ =& e_j^\intercal \left[\mathbb{E}\frac{[f'(X_i^\intercal\beta)]^2}{f(X_i^\intercal\beta)[1 - f(X_i^\intercal\beta)]} X_i X_i^\intercal\right]^{-1} \frac{f'(X_i^\intercal\beta)}{f(X_i^\intercal\beta)[1 - f(X_i^\intercal\beta)]} X_i \left[y_i - f(X_i^\intercal\beta)\right] \end{aligned} \tag{7.2}$$

e_j is the j-th Euclidean basis. By the definition $w(z) = \frac{f'(z)}{f(z)(1-f(z))}$, we establish (2.13) in Remark 2.

References

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIIIÑ1983*, pp. 1–198. Springer.

Cai, T. T. and Z. Guo (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics 45*(2), 615–646.

Candès, E., Y. Fan, L. Janson, and J. Lv (2018). Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(3), 551–577.

Gradshteyn, I. S. and I. M. Ryzhik (2014). *Table of Integrals, Series, and Products*. Academic press.

Huang, J. and C.-H. Zhang (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *Journal of Machine Learning Research 13*(Jun), 1839–1864.

Ledoux, M. and M. Talagrand (2013). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media.

Ning, Y. and G. Cheng (2020). Sparse confidence sets for normal mean models. *arXiv preprint arXiv:2008.07107*.

Nualart, D. (2006). *The Malliavin Calculus and Related Topics*. Springer Science & Business Media.

Shao, J. (2003). *Mathematical Statistics*. Springer Texts in Statistics.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Zhou, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*.