Elastic-net Regularized High-dimensional Negative Binomial Regression: Consistency and Weak Signal Detection

Huiming Zhang[1,2], Jinzhu Jia[1]

[1]*Peking University,* [2]*University of Macau;*

Supplementary Material

S1 Main Proofs

S1.1 Proof of Theorem 1

With the aim of deriving the targeted oracle inequalities (2.11), we first prove the lower bound for symmetric Bregman divergence $D_g^s(\boldsymbol{\beta}+\boldsymbol{\delta},\boldsymbol{\beta})$ with $g=0$.

Lemma 1. *Assume that (C.1) and (C.2) are satisfied, then we have*

$$D^s(\boldsymbol{\beta}+\boldsymbol{\delta},\boldsymbol{\beta}) \geq \boldsymbol{\delta}^T \ddot{\ell}(\boldsymbol{\beta})\boldsymbol{\delta} e^{-2L\|\boldsymbol{\delta}\|_1}.$$

Proof. We assume that $\boldsymbol{X}_i^T\boldsymbol{\delta} \neq 0$ by identifiability (C.2) for $\boldsymbol{\beta}$. Use the expression of $\dot{\ell}_n(\boldsymbol{\beta})$, we obtain

$$\begin{aligned}
\boldsymbol{\delta}^T[\dot{\ell}_n(\boldsymbol{\beta}+\boldsymbol{\delta}) - \dot{\ell}(\boldsymbol{\beta})] &= -\boldsymbol{\delta}^T \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\theta[\frac{\theta+Y_i}{\theta+e^{\boldsymbol{X}_i^T(\boldsymbol{\beta}+\boldsymbol{\delta})}} - \frac{\theta+Y_i}{\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}] \\
&= \boldsymbol{\delta}^T \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T\theta \cdot \frac{(\theta+Y_i)e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{[\theta+e^{\boldsymbol{X}_i^T(\boldsymbol{\beta}+\boldsymbol{\delta})}][\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}]} \cdot \frac{e^{\boldsymbol{X}_i^T\boldsymbol{\delta}}-1}{\boldsymbol{X}_i^T\boldsymbol{\delta}-0}\boldsymbol{\delta} \\
&\geq \boldsymbol{\delta}^T \frac{1}{n}\sum_{i=1}^n \left\{ \boldsymbol{X}_i\boldsymbol{X}_i^T \cdot \frac{\theta(\theta+Y_i)e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{[\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}]^2} \cdot \frac{\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\theta+e^{\boldsymbol{X}_i^T(\boldsymbol{\beta}+\boldsymbol{\delta})}} e^{-(|\boldsymbol{X}_i^T\boldsymbol{\delta}|\vee 0)} \right\}\boldsymbol{\delta}
\end{aligned}$$

where the last inequality is from $\frac{e^x-e^y}{x-y} \geq e^{-(|x|\vee|y|)}$. It remains to prove that

$$\frac{\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\theta+e^{\boldsymbol{X}_i^T(\boldsymbol{\beta}+\boldsymbol{\delta})}} \geq e^{-L\|\boldsymbol{\delta}\|_1}. \tag{S1.1}$$

To show the (S1.1), just note that by (C.1)

$$\begin{cases} \frac{\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\theta+e^{\boldsymbol{X}_i^T(\boldsymbol{\beta}+\boldsymbol{\delta})}} \geq e^{-\boldsymbol{X}_i^T\boldsymbol{\delta}} \geq e^{-L\|\boldsymbol{\delta}\|_1} & \text{if } \boldsymbol{X}_i^T\boldsymbol{\delta} \geq 0 \\ \frac{\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\theta+e^{\boldsymbol{X}_i^T(\boldsymbol{\beta}+\boldsymbol{\delta})}} \geq 1 & \text{if } \boldsymbol{X}_i^T\boldsymbol{\delta} \leq 0. \end{cases}$$

Last, combining inequality $\min\{e^{-|\boldsymbol{X}_i^T\boldsymbol{\delta}|}, 1\} \geq e^{-L\|\boldsymbol{\delta}\|_1}$ and (S1.1), it implies by the expression of $\ddot{\ell}(\boldsymbol{\beta})$ that

$$\boldsymbol{\delta}^T[\dot{\ell}(\boldsymbol{\beta}+\boldsymbol{\delta})-\dot{\ell}(\boldsymbol{\beta})] \geq b^T\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\boldsymbol{X}_i\boldsymbol{X}_i^T\cdot\theta(\theta+Y_i)e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{(\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}})^2}\right\}\boldsymbol{\delta}e^{-2L\|\boldsymbol{\delta}\|_1} = \boldsymbol{\delta}^T\ddot{\ell}(\boldsymbol{\beta})\boldsymbol{\delta}e^{-2L\|\boldsymbol{\delta}\|_1}.$$

□

Next, we give the proof of Theorem 1 based on Lemma 1.

Proof. Let $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \neq 0$ and $\boldsymbol{b} = \tilde{\boldsymbol{\beta}}/\|\tilde{\boldsymbol{\beta}}\|_1$, and then $\ell(\boldsymbol{\beta}^* + \boldsymbol{b}x)$ is a convex function in x due to the convexity of $\ell(\boldsymbol{\beta})$. By (2.10), we have

$$\boldsymbol{b}^T[\dot{\ell}(\boldsymbol{\beta}^*+\boldsymbol{b}x)-\dot{\ell}(\boldsymbol{\beta}^*)] \leq \frac{2\,\zeta\lambda_1}{\zeta+1}\|\boldsymbol{b}_H\|_1 - \frac{2\lambda_1}{\zeta+1}\|\boldsymbol{b}_{H^C}\|_1 \leq \frac{2\,\zeta\lambda_1}{\zeta+1}\|\boldsymbol{b}_H\|_1 \quad \text{(S1.2)}$$

holds for $x \in [0, \|\tilde{\boldsymbol{\beta}}\|_1]$ and $\boldsymbol{b} \in \mathrm{S}(\zeta, H)$.

By the Lemma 1, we get $(\boldsymbol{b}x)^T[\dot{\ell}_n(\boldsymbol{\beta}^*+\boldsymbol{b}x)-\dot{\ell}_n(\boldsymbol{\beta}^*)] \geq e^{-2Lx}(\boldsymbol{b}x)^T\ddot{\ell}_n(\boldsymbol{\beta})(\boldsymbol{b}x)$. Since $x \geq 0$, then

$$\boldsymbol{b}^T[\dot{\ell}_n(\boldsymbol{\beta}^* + \boldsymbol{b}x) - \dot{\ell}_n(\boldsymbol{\beta}^*)] \geq xe^{-2Lx}\boldsymbol{b}^T\ddot{\ell}_n(\boldsymbol{\beta})\boldsymbol{b}. \quad \text{(S1.3)}$$

Assume we know the Hessian matrix at the true coefficient $\boldsymbol{\beta}^*$, write compatibility factor as $\mathrm{C}(\zeta, H) =: \mathrm{C}(\zeta, H, \ddot{\ell}_n(\boldsymbol{\beta}^*))$. By the definition of compatibility factor and the two inequality above, we have

$$\begin{aligned} Lxe^{-2Lx}[C(\zeta,H)]^2\|\boldsymbol{b}_H\|_1^2/d_H^* &\leq Lxe^{-2Lx}\boldsymbol{b}^T\ddot{\ell}_n(\boldsymbol{\beta})\boldsymbol{b} \\ \text{(by (S1.3))} \quad &\leq L\boldsymbol{b}^T[\dot{\ell}_n(\boldsymbol{\beta}^* + \boldsymbol{b}x) - \dot{\ell}_n(\boldsymbol{\beta}^*)] \\ \text{(by (S1.2))} \quad &\leq L(\frac{2\,\zeta\lambda_1}{\zeta+1}\|\boldsymbol{b}_H\|_1 - \frac{2\lambda_1}{\zeta+1}\|\boldsymbol{b}_{H^C}\|_1) \\ &= L[\frac{2\,\zeta\lambda_1}{\zeta+1}\|\boldsymbol{b}_H\|_1 - \frac{2\lambda_1}{\zeta+1}(1 - \|\boldsymbol{b}_H\|_1)] \\ &\leq L(2\lambda_1\|\boldsymbol{b}_H\|_1 - \frac{2\lambda_1}{\zeta+1}) \leq \frac{L(\zeta+1)\|\boldsymbol{b}_H\|_1^2\lambda_1}{2}. \end{aligned}$$

where the last step is due to the elementary inequality $\frac{2\lambda_1}{\zeta+1} + \frac{(\zeta+1)\|\mathbf{b}_H\|_1^2\lambda_1}{2} \geq 2\lambda_1\|\mathbf{b}_H\|_1$.

Then we have

$$Lxe^{-2Lx} \leq \frac{L(\zeta+1)d_H^*\lambda_1}{2[C(\zeta,H)]^2} =: \tau \tag{S1.4}$$

for any $x \in [0, \|\tilde{\boldsymbol{\beta}}\|_1]$. a_τ is the small solution of the equation $\{z : ze^{-2z} = \tau\}$. Notice that the maximum of ze^{-2z} is $\frac{1}{2}e^{-1}$, we need to assume $\tau \leq \frac{1}{2}e^{-1}$.

Again, since $\ell_n(\boldsymbol{\beta})$ is a convex in $\boldsymbol{\beta}$, then $\boldsymbol{b}^T[\dot{\ell}_n(\boldsymbol{\beta}+\boldsymbol{b}x) - \ell_n(\boldsymbol{\beta})]$ is increasing in x. Thus the solution of (S1.4) w.r.t. x is a closed interval $x \in [0, \tilde{x}]$. By the fact that $x \in [0, \|\tilde{\boldsymbol{\beta}}\|_1]$ implies $x \in [0, \tilde{x}]$, thus we have $\|\tilde{\boldsymbol{\beta}}\|_1 \leq \tilde{x}$. Use (S1.4) again, it implies $L\tilde{x}e^{-2L\tilde{x}} \leq \tau$. Then, for $\forall x \in [0, \tilde{x}]$, we have

$$\|\tilde{\boldsymbol{\beta}}\|_1 \leq \tilde{x} \leq \frac{a_\tau}{L} = \frac{e^{2a_\tau}\tau}{L} = \frac{e^{2a_\tau}(\zeta+1)d_H^*\lambda_1}{2[C(\zeta,H)]^2} \tag{S1.5}$$

where the last equality is by the definition of τ.

Similarly, by the definition of weak CIF, we have

$$\begin{aligned} xe^{-2Lx} &\leq \frac{xe^{-2Lx}\boldsymbol{b}^T\ddot{\ell}_n(\boldsymbol{\beta})\boldsymbol{b}}{C_q(\zeta,H)(\|\boldsymbol{b}_H\|_1/({d_H^*}^{1/q})\|\boldsymbol{b}\|_q} \leq \frac{\boldsymbol{b}^T[\dot{\ell}_n(\boldsymbol{\beta}^*+\boldsymbol{b}x) - \dot{\ell}_n(\boldsymbol{\beta}^*)]}{C_q(\zeta,H)(\|\boldsymbol{b}_H\|_1/({d_H^*}^{1/q})\|\boldsymbol{b}\|_q} \\ &\qquad \text{(by (S1.2))} \quad \leq \frac{2\zeta {d_H^*}^{1/q}\lambda_1}{(\zeta+1)\,C_q(\zeta,H)\|\boldsymbol{b}\|_q}. \end{aligned}$$

Let $x = \|\tilde{\boldsymbol{\beta}}\|_1$, by the identity $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q = \|\tilde{\boldsymbol{\beta}}\|_1\|\boldsymbol{b}\|_q$, we have $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q \leq \frac{2e^{2a_\tau}\zeta {d_H^*}^{1/q}\lambda_1}{(\zeta+1)C_q(\zeta,H)}$ due to the same argument in (S1.5). □

S1.2 Proof of Theorem 2

To show the high probability events $\mathcal{K} \cap \mathcal{E}_c$ (or $\mathcal{K} \cap \mathcal{E}_w$), we will adopt the sub-Gaussian type concentration inequalities for the exponential family random variables with restricted parameter space.

Lemma 2 (Lemma 6.1 in Rigollet (2012))**.** *Let $\{Y_i\}_{i=1}^n$ be a sequence of random variables whose distribution belongs to the canonical exponential family with $f(y_i;\theta_i) = c(y_i)\exp(y_i\theta_i - \psi(\theta_i))$. We assume the uniformly bounded variances condition: there exist a compact set Ω and a constant C_ψ^2 such that $\sup_{\theta_i\in\Omega}\ddot{\psi}(\theta_i) \leq C_\psi^2$ for all i. Let $\boldsymbol{w} := (w_1, \cdots, w_n)^T \in \mathbb{R}^n$ be a*

non-random and define the weighted sum $S_n^w =: \sum_{i=1}^n w_i Y_i$*, we have*

$$P\{|S_n^w - \mathrm{E}S_n^w| > t\} \leq 2\exp\{-\frac{t^2}{2C_\psi^2\|\boldsymbol{w}\|_2^2}\}. \tag{S1.6}$$

Moreover, we have $\mathrm{E}|S_n^w - \mathrm{E}S_n^w|^k \leq D_{k,C}\,\|w\|_2^k$ *where* $D_{k,C} = k(2C_\psi^2)^{k/2}\Gamma(k/2)$ *and* $\Gamma(\cdot)$ *stands for the Gamma function.*

Since dispersion parameters θ is assumed to be known, this NB distribution belongs to exponential families. With assumption (C.1) and (C.3), the boundedness of $\sup_{\theta_i\in\Omega}\ddot{\psi}(\theta_i)$ holds uniformly by noticing that

$$\max_i \sup_{\theta_i\in\Omega}\ddot{\psi}(\theta_i) = \max_i \sup_{\mu_i>0}(\mu_i + \frac{\mu_i^2}{\theta}) = \sup_{|t|\leq LB}(e^t + \frac{e^{2t}}{\theta}) = e^{LB} + \frac{e^{2LB}}{\theta} := C_{LB}^2. \tag{S1.7}$$

Now, we can apply concentration inequality Lemma 2 to go on the proof. The first step is to evaluate the event $\mathcal{K} := \left\{z^* \leq \frac{\zeta-1}{\zeta+1}\lambda_1\right\}$ from the inequality in (2.12). By assuming $B\lambda_2 := B_1\lambda_1$, we have

$$\begin{aligned} P(z^* \geq \frac{\zeta-1}{\zeta+1}\lambda_1) &\leq P(\|\dot{\ell}_n(\boldsymbol{\beta}^*)\|_\infty \geq \frac{\zeta-1}{\zeta+1}\lambda_1 - 2\lambda_2 B) \\ &\leq \sum_{j=1}^p P\left(\left|\sum_{i=1}^n \frac{x_{ij}(Y_i - \mathrm{E}Y_i)\theta}{n(\theta+\mathrm{E}Y_i)}\right| \geq \frac{\zeta-1}{\zeta+1}\lambda_1 - 2\lambda_1 B_1\right). \end{aligned}$$

and define $C_{\xi,B_1} := \frac{\zeta-1}{\zeta+1} - B_1 > 0$ for some small constant B_1.

It is worth noting that Bunea (2008) and Blazere et al. (2014) also proposed assumption $\lambda_2 B = O(\lambda_1)$ for two turning parameters in Elastic-net estimates. By using Lemma 2, we have

$$P\left\{\left|\sum_{i=1}^n \frac{1}{n}\frac{x_{ij}(Y_i - \mathrm{E}Y_i)\theta}{\theta+\mathrm{E}Y_i}\right| \geq C_{\xi,B_1}\lambda_1\right\} \leq 2\exp\{-\frac{C_{\xi,B_1}^2\lambda_1^2}{2C_{LB}^2||\boldsymbol{w}^{(j)}||_2^2}\} \leq 2\exp\{-\frac{C_{\xi,B_1}^2\lambda_1^2 n}{2C_{LB}^2L^2}\},$$

where $||\boldsymbol{w}^{(j)}||_2^2 := \sum_{i=1}^n \frac{x_{ij}^2\theta^2}{n^2(\theta+\mathrm{E}Y_i)^2} \leq \frac{L^2}{n}$. Consequently,

$$P(z^* \geq \frac{\zeta-1}{\zeta+1}\lambda_1) \leq 2p\exp\{-\frac{C_{\xi,B_1}^2\lambda_1^2 n}{2C_{LB}^2L^2}\} =: \frac{2}{p^{r-1}}, \quad r>1. \tag{S1.8}$$

The expression of tuning parameter λ_1 is solved by the equality in (S1.8), we obtain $\lambda_1 = \frac{C_{LB}L}{C_{\xi,B_1}}\sqrt{\frac{2r\log p}{n}}$.

The second step is to evaluate the probability of the event of truncated random variables:

$$\mathcal{E}_c := \{C^2(\zeta, H, \ddot{\ell}_n(\boldsymbol{\beta}^*)) \geq C_t^2(\zeta, H)\}$$

and

$$\mathcal{E}_w := \{C_q(\zeta, H, \ddot{\ell}_n(\boldsymbol{\beta}^*)) \geq C_{qt}(\zeta, H)\},$$

where $C_t^2(\zeta, H)$ and $C_{qt}(\zeta, H)$ are some constant such that these two events could hold with high probability.

Let $\tilde{\boldsymbol{b}}_c, \tilde{\boldsymbol{b}}_w$ be the random points such that the infimum of in the following ℓ_1-ball restricted compatibility factor and weak cone invertibility factors,

$$C^2(\zeta, H, \ddot{\ell}_n(\boldsymbol{\beta}^*)) := \inf_{\boldsymbol{b}\in\Lambda} \frac{{d_H^*}^{1/2}(\boldsymbol{b}^T\ddot{\ell}_n(\boldsymbol{\beta}^*)\boldsymbol{b})}{\|\boldsymbol{b}_H\|_1^2} =: \frac{d_H^*(\tilde{\boldsymbol{b}}_c\ddot{\ell}_n(\boldsymbol{\beta}^*)\tilde{\boldsymbol{b}}_c)}{\|(\tilde{\boldsymbol{b}}_c)_H\|_1^2} > 0, \ \ (s \in \mathbb{R}),$$

$$C_q(\zeta, H, \ddot{\ell}_n(\boldsymbol{\beta}^*)) := \inf_{\boldsymbol{b}\in\Lambda} \frac{{d_H^*}^{1/q}\boldsymbol{b}^T\ddot{\ell}_n(\boldsymbol{\beta}^*)\boldsymbol{b}}{||\boldsymbol{b}_H||_1 \cdot ||\boldsymbol{b}||_q} =: \frac{{d_H^*}^{1/q}\tilde{\boldsymbol{b}}_w^T\ddot{\ell}_n(\boldsymbol{\beta}^*)\tilde{\boldsymbol{b}}_w}{||(\tilde{\boldsymbol{b}}_w)_H||_1 \cdot ||\tilde{\boldsymbol{b}}_w||_q} > 0, (\zeta \in \mathbb{R})$$

are attained respectively, where $\Lambda := \{\boldsymbol{b} \in \mathbb{R}^p : \boldsymbol{0} \neq \boldsymbol{b} \in \mathrm{S}(\zeta, H), \|\boldsymbol{b}\|_1 = 1\}$.

Consider the event $\mathcal{E}_c$ and $\mathcal{E}_w$, let

$$S_n^c(\boldsymbol{b}, Y) := \frac{d_H^*(\boldsymbol{b}^T\ddot{\ell}_n(\boldsymbol{\beta}^*)\boldsymbol{b})}{\|\boldsymbol{b}_H\|_1^2} \text{ and } S_n^w(\boldsymbol{b}, Y) := \frac{{d_H^*}^{1/q}\boldsymbol{b}^T\ddot{\ell}_n(\boldsymbol{\beta}^*)\boldsymbol{b}}{||\boldsymbol{b}_H||_1 \cdot ||\boldsymbol{b}||_q}.$$

For all $\boldsymbol{b} \in \Lambda$, the difference of $S_n^c(\boldsymbol{b}, Y)$ and $\mathrm{E}S_n^c(\boldsymbol{b}, Y)$ is bound by

$$\begin{aligned}|S_n^c(\boldsymbol{b}, Y) - \mathrm{E}S_n^c(\boldsymbol{b}, Y)| &\leq \frac{d_H^*\|\boldsymbol{b}\|_1^2}{\|\boldsymbol{b}_H\|_1^2} \max_{j,k} |(\ddot{\ell}_n(\beta^*) - \mathrm{E}\ddot{\ell}_n(\beta^*))_{j,k}| \\ &\leq d_H^*(1+\zeta)^2 \max_{j,k} |(\ddot{\ell}_n(\boldsymbol{\beta}^*) - \mathrm{E}\ddot{\ell}_n(\boldsymbol{\beta}^*))_{j,k}|\end{aligned}$$

where the last inequality is from (2.10).

Note that the term $d_H^*(1+\zeta)^2$ is a constant, so it sufficient to bound

$$\max_{j,k} |(\ddot{\ell}_n(\beta^*) - \mathrm{E}\ddot{\ell}_n(\beta^*))_{j,k}| = \max_{j,k} \left| \frac{1}{n}\sum_{i=1}^n \frac{x_{ij}x_{ik}\theta e^{X_i^T\beta^*}}{(\theta + e^{X_i^T\beta^*})^2}(Y_i - \mathrm{E}Y_i) \right|$$

by Lemma 2. Then,

$$P\{|S_n^c(\boldsymbol{b}, Y) - \mathrm{E}S_n^c(\boldsymbol{b}, Y)| \geq t, \ \forall \boldsymbol{b} \in \Lambda\} \tag{S1.9}$$

$$\leq P\{\max_{j,k}|(\ddot{\ell}_n(\beta^*)-\mathrm{E}\ddot{\ell}_n(\beta^*))_{j,k}|\geq t/d_H^*(1+\zeta)^2\}$$
$$\leq p^2P\{|(\ddot{\ell}_n(\beta^*)-\mathrm{E}\ddot{\ell}_n(\beta^*))_{j,k}|\geq t/d_H^*(1+\zeta)^2\}$$
$$\leq 2p^2\exp\{-\frac{nt^2}{2C_{LB}^2[d_H^*(1+\zeta)L^2]^2}\} \tag{S1.10}$$

where the last inequality is by using Lemma 2 with $||\boldsymbol{w}||_2^2\leq L^4/n$.

We try to define

$$P(\mathcal{E}_c):=P\{C^2(\zeta,H)\geq C_t^2(\zeta,H)\}=P\{S_n^c(\tilde{\boldsymbol{b}}_c,Y)-\mathrm{E}S_n^c(\tilde{\boldsymbol{b}}_c,Y)\geq -t\}.$$

Since the inequality (S1.9) is free of $\boldsymbol{b}$, thus by the (S1.9) for $\Lambda\ni\tilde{\boldsymbol{b}}_c$ we have

$$P(\mathcal{E}_c)=P\{S_n^c(\tilde{\boldsymbol{b}}_c,Y)-\mathrm{E}S_n^c(\tilde{\boldsymbol{b}}_c,Y)\geq -t\}=1-P\{S_n^c(\tilde{\boldsymbol{b}}_c,Y)-\mathrm{E}S_n^c(\tilde{\boldsymbol{b}}_c,Y)\leq -t\}$$
$$>1-P\{S_n^c(\tilde{\boldsymbol{b}}_c,Y)-\mathrm{E}S_n^c(\tilde{\boldsymbol{b}}_c,Y)\leq -t\}-P\{S_n^c(\tilde{\boldsymbol{b}}_c,Y)-\mathrm{E}S_n^c(\tilde{\boldsymbol{b}}_c,Y)\geq t\}$$
$$\geq P\{|S_n^c(\boldsymbol{b},Y)-\mathrm{E}S_n^c(\boldsymbol{b},Y)|\geq t,\ \forall\boldsymbol{b}\in\Lambda\}\geq 1-2p^2\exp\{-\frac{nt^2}{2[d_H^*C_{LB}(1+\zeta)L^2]^2}\}.$$

Hence, we could find $C_t^2(\zeta,H)$. For example, the t can be chosen as $\frac{1}{2}\mathrm{E}S_n^c(\tilde{\boldsymbol{b}}_c,Y)$ or others. The probability of the intersection of two events $\mathcal{K}$ and $\mathcal{E}_c$ is at least

$$P(\mathcal{K}\cap\mathcal{E}_c)\geq P(\mathcal{K})+P(\mathcal{E}_c)-1\geq 1-\frac{2}{p^{r-1}}-2p^2\exp\{-\frac{nt^2}{2[d_H^*C_{LB}(1+\zeta)L^2]^2}\}.$$

Next, we consider similar arguments for concerning $\mathcal{E}_w$. For all $\boldsymbol{b}\in\Lambda$, the absolute difference of $S_n^w(\boldsymbol{b},Y)$ and $\mathrm{E}S_n^w(\boldsymbol{b},Y)$ is bounded by

$$|S_n^w(\boldsymbol{b},Y)-\mathrm{E}S_n^w(\boldsymbol{b},Y)|\leq\frac{{d_H^*}^{1/q}\|\boldsymbol{b}\|_1^2}{||\boldsymbol{b}_H||_1\cdot||\boldsymbol{b}||_q}\max_{j,k}|(\ddot{\ell}_n(\beta^*)-\mathrm{E}\ddot{\ell}_n(\beta^*))_{j,k}|$$
$$\leq\frac{{d_H^*}^{1/q}(1+\zeta)^2\|\boldsymbol{b}_H\|_1^2}{||\boldsymbol{b}_H||_1\cdot||\boldsymbol{b}||_q}\max_{j,k}|(\ddot{\ell}_n(\beta^*)-\mathrm{E}\ddot{\ell}_n(\beta^*))_{j,k}|$$
(By Hölder's inequality) $$\leq\frac{{d_H^*}^{1/q}(1+\zeta)^2{d_H^*}^{(1-1/q)}\|b_H\|_q}{\|b_H\|_q}\max_{j,k}|(\ddot{\ell}_n(\beta^*)-\mathrm{E}\ddot{\ell}_n(\beta^*))_{j,k}|$$
$$\leq d_H^*(1+\zeta)^2\max_{j,k}|(\ddot{\ell}_n(\boldsymbol{\beta}^*)-\mathrm{E}\ddot{\ell}_n(\boldsymbol{\beta}^*))_{j,k}|$$

where the second last inequality is from (2.10).

Let $u = \frac{1}{2}\mathrm{E}S_n^w(\tilde{\boldsymbol{b}}_w, Y)$. The same derivation show that

$$P(\mathcal{E}_w) = P\{S_n^w(\tilde{\boldsymbol{b}}_w, Y) - \mathrm{E}S_n^w(\tilde{\boldsymbol{b}}_w, Y) \geq -u\} \geq 1 - 2p^2 \exp\{-\frac{nu^2}{2[d_H^* C_{LB}(1+\zeta)L^2]^2}\}$$

and

$$P(\mathcal{K} \cap \mathcal{E}_w) \geq P(\mathcal{K}) + P(\mathcal{E}_w) - 1 \geq 1 - \frac{2}{p^{r-1}} - 2p^2 \exp\{-\frac{nu^2}{2[d_H^* C_{LB}(1+\zeta)L^2]^2}\}.$$

S1.3 Proof of Lemma 3

Proof. Judging from the convexity of the loss function and the elastic-net penalty, the chief ingredients of the proof is similar in spirit to the one used by Theorem 6.4 in Bühlmann and van de Geer (2011) for initially restricting the penalized estimator in a ball centred at its true value, and see also Lemma III.4 in Blazere et al. (2014).

Put $t = \frac{M}{M + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1}$ and $\tilde{\boldsymbol{\beta}} := t\hat{\boldsymbol{\beta}} + (1-t)\boldsymbol{\beta}^*$, so $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* := t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$. Therefore,

$$t = \frac{M}{M + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1} = \frac{M}{M + \frac{1}{t}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1}.$$

Then

$$\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq M(1-t) \leq M, \quad \text{i.e. } \tilde{\boldsymbol{\beta}} \in \mathcal{S}_M.$$

By the definition, $\hat{\boldsymbol{\beta}}$ satisfies

$$\mathbb{P}_n l(\hat{\boldsymbol{\beta}}) + \lambda_1\|\hat{\boldsymbol{\beta}}\|_1 + \lambda_2\|\hat{\boldsymbol{\beta}}\|_2^2 \leq \mathbb{P}_n l(\boldsymbol{\beta}^*) + \lambda_1\|\boldsymbol{\beta}^*\|_1 + \lambda_2\|\boldsymbol{\beta}^*\|_2^2. \tag{S1.11}$$

By convexity of the optimization function (2.1), combined with (S1.11), we get

$$\mathbb{P}_n l(\tilde{\boldsymbol{\beta}}) + \lambda_1\|\tilde{\boldsymbol{\beta}}\|_1 + \lambda_2\|\tilde{\boldsymbol{\beta}}\|_2^2 \leq \mathbb{P}_n l(\hat{\boldsymbol{\beta}}) + \lambda_1\|\hat{\boldsymbol{\beta}}\|_1 + \lambda_2\|\hat{\boldsymbol{\beta}}\|_1^2 \leq \mathbb{P}_n l(\boldsymbol{\beta}^*) + \lambda_1\|\boldsymbol{\beta}^*\|_1 + \lambda_2\|\boldsymbol{\beta}^*\|_2^2.$$

Thus

$$\mathbb{P}(l(\tilde{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^*)) + \lambda_1\|\tilde{\boldsymbol{\beta}}\|_1 + \lambda_2\|\tilde{\boldsymbol{\beta}}\|_2^2 \leq (\mathbb{P}_n - \mathbb{P})(l(\boldsymbol{\beta}^*) - l(\tilde{\boldsymbol{\beta}})) + \lambda_1\|\boldsymbol{\beta}^*\|_1 + \lambda_2\|\boldsymbol{\beta}^*\|_2^2.$$

On the event $\mathcal{A}$, using Proposition 1, we have

$$(\mathbb{P}_n - \mathbb{P})(l_1(\tilde{\boldsymbol{\beta}}) - l_1(\boldsymbol{\beta}^*)) \leq \frac{\lambda_1}{4}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1.$$

Since $\tilde{\boldsymbol{\beta}} \in \mathcal{S}_M$, by definition of $\mathcal{B}$, it yields

$$(\mathbb{P}_n - \mathbb{P})(l_2(\tilde{\boldsymbol{\beta}}) - l_2(\boldsymbol{\beta}^*)) \leq \frac{\lambda_1}{4}(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \varepsilon_n).$$

These two inequalities imply

$$\mathbb{P}(l(\tilde{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*))+\lambda_1|\tilde{\boldsymbol{\beta}}\|+\lambda_2|\tilde{\boldsymbol{\beta}}\|_2^2 \leq \frac{\lambda_1}{2}\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1+\lambda_1\frac{\varepsilon_n}{4}+\lambda_1\|\boldsymbol{\beta}^*\|_1+\lambda_2\|\boldsymbol{\beta}^*\|_2^2. \tag{S1.12}$$

Note that $\mathbb{P}(l(\tilde{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^*)) \geq 0$ from the definition of $\boldsymbol{\beta}^*$, and by using the triangular inequality, we obtain

$$\begin{aligned}\lambda_1\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 &\leq \lambda_1\|\tilde{\boldsymbol{\beta}}\|_1 + \lambda_1\|\boldsymbol{\beta}^*\|_1 \leq [\mathbb{P}(l(\tilde{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^*)) + \lambda_1\|\tilde{\boldsymbol{\beta}}\|_1] + \lambda_1\|\boldsymbol{\beta}^*\|_1 \\ [\text{by } (S1.12)] \quad &\leq \frac{\lambda_1}{2}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \frac{\lambda_1\varepsilon_n}{4} + 2\lambda_1\|\boldsymbol{\beta}^*\|_1 + (\lambda_2\|\boldsymbol{\beta}^*\|_2^2 - \lambda_2\|\tilde{\boldsymbol{\beta}}\|_2^2).\end{aligned} \tag{S1.13}$$

From the assumption $8B\lambda_2 + 4M = \lambda_1$ and (H.2), then the quadratic part in last expression is bounded from above by

$$\lambda_2(\|\boldsymbol{\beta}^*\|_2^2-\|\tilde{\boldsymbol{\beta}}\|_2^2) = \sum_{j=1}^{p}\lambda_2(\beta_j^*+\tilde{\beta}_j)(\beta_j^*-\tilde{\beta}_j) \leq (2B+M)\lambda_2\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 := \frac{\lambda_1}{4}\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1$$

where the inequality in above expression is by the fact

$$\beta_j^* + \tilde{\beta}_j = t(\hat{\beta}_j - \beta_j^*) + 2\beta^* \leq M + 2B \text{ uniformly in } j.$$

Therefore, (S1.13) implies

$$\lambda_1\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{3\lambda_1}{4}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \frac{\lambda_1\varepsilon_n}{4} + 2\lambda_1\|\boldsymbol{\beta}^*\|_1.$$

Cancelling λ_1 in the inequality above, it gives $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \varepsilon_n + 8\|\boldsymbol{\beta}^*\|_1$. We have

$$t\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \varepsilon_n + 8\|\boldsymbol{\beta}^*\|_1 =: \frac{M}{2}.$$

Plugging in the definition of t, we have $\frac{M\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1}{M+\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1} \leq \frac{M}{2}$, which derives $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq M$. □

S1.4 Proof of Proposition 1 and 2

We deduce Proposition 2 by showing the following key lemma.

Lemma 3. *Let* $\lambda_1 \geq \frac{20\theta AML}{M+\varepsilon_n}\sqrt{\frac{2\log 2p}{n}}$ $(A \geq 1)$. *Then* $P(\mathcal{B}) \geq 1 - (2p)^{-A^2}$ *under (H.1).*

Lemma 3 and Proposition 1 jointly tell us that $P(\mathcal{A}), P(\mathcal{B}) \to 1$ as $p \to 0$. If λ_1 are chosen such that

$$\lambda_1 \geq \max\left(\frac{20\theta AML}{M+\varepsilon_n}\sqrt{\frac{2\log 2p}{n}}, 4(2L\tilde{C}_{LB} + A\sqrt{2\gamma})\sqrt{\frac{2\log 2p}{n}}\right),$$

thus we obtain

$$P(\mathcal{A}\cap\mathcal{B}) \geq P(\mathcal{A}) + P(\mathcal{B}) - 1 \geq 1 - 2(2p)^{-A^2},$$

which finishes the proof of Proposition 2.

It remains to show the Lemma 3 and and Proposition 1 used in the proof of Proposition 2.

Proof of Lemma 3

The proof rests on the following lemma.

Lemma 4. *Given $M > 0$, if $A \geq 1$, under (H.1), we have*

$$P(Z_M(\boldsymbol{\beta}^*) \geq \frac{5\theta AML}{(M+\varepsilon_n)}\sqrt{\frac{2\log(2p)}{n}}) \leq (2p)^{-A^2}. \tag{S1.14}$$

where $Z_M(\boldsymbol{\beta}^*) = \sup_{\boldsymbol{\beta}\in S_M}\{\frac{|(\mathbb{P}_n-\mathbb{P})(l_2(\boldsymbol{\beta}^*)-l_2(\boldsymbol{\beta}))|}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1+\varepsilon_n}\}$.

In order to apply following McDiarmid's inequality (also called bounded difference inequality, see Theorem 3.3.14 of Giné and Nickl (2015)), we replaced X_i by $X_i^{'}$ meamwhile maintaining the others fixed.

Theorem 1 (McDiarmid's inequality)**.** *Let A be a measurable set. Assume $f : A^n \to \mathbb{R}$ is a multivariate measurable function with bounded differences conditions*

$$\sup_{x_1,...,x_n,x_i^{'}\in A} |f(x_1,...,x_n) - f(x_1,...,x_{i-1},x_i^{'},x_{i+1},...,x_n)| \leq c_i.$$

Let $X_1,..,X_n$ be independent random variables with values in the set A. Then, for all $t > 0$, we have

$$P(f(X_1,...,X_n) - \mathrm{E}f(X_1,...,X_n) \geq t) \leq e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

First, we will to show that $Z_M(\boldsymbol{\beta}^*)$ is fluctuated of no more than $\frac{2\theta LM}{n(M+\varepsilon_n)}$. Let us check it. Put

$$\mathbb{P}_n = \frac{1}{n}\sum_{j=1}^n 1_{\boldsymbol{X}_j,Y_j} \quad \text{and} \quad \mathbb{P}_n^{'} = (\frac{1}{n}\sum_{j=1,j\neq i}^n 1_{\boldsymbol{X}_j,Y_j} + 1_{\boldsymbol{X}_i^{'},Y_i^{'}}),$$

it deduces

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta}\in S_M} \frac{|(\mathbb{P}_n-\mathbb{P})(l_2(\boldsymbol{\beta}^*)-l_2(\boldsymbol{\beta}))|}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1+\varepsilon_n} - \sup_{\boldsymbol{\beta}\in S_M} \frac{|(\mathbb{P}'_n-\mathbb{P})(l_2(\boldsymbol{\beta}^*)-l_2(\hat{\boldsymbol{\beta}}))|}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1+\varepsilon_n} \\
\le &\sup_{\boldsymbol{\beta}\in S_M} \frac{|l_2(\boldsymbol{\beta}^*,\boldsymbol{X}_i)-l_2(\boldsymbol{\beta},\boldsymbol{X}_i)-l_2(\boldsymbol{\beta}^*,\boldsymbol{X'}_i)+l_2(\boldsymbol{\beta},\boldsymbol{X'}_i)|}{n(\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1+\varepsilon_n)} \\
\le &\sup_{\boldsymbol{\beta}\in S_M} \frac{1}{n}\left|\frac{\theta e^{\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}}}{\theta+e^{\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}}}\right|\cdot\frac{|\boldsymbol{X}_i^T\boldsymbol{\beta}^*-\boldsymbol{X}_i^T\boldsymbol{\beta}|}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1+\varepsilon_n} + \sup_{\boldsymbol{\beta}\in S_M}\frac{1}{n}\left|\frac{\theta e^{\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}}}{\theta+e^{\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}}}\right|_1\cdot\frac{|\boldsymbol{X'}_i^T\boldsymbol{\beta}^*-\boldsymbol{X'}_i^T\boldsymbol{\beta}|}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1+\varepsilon_n} \\
\le &\sup_{\boldsymbol{\beta}\in S_M}\frac{2\theta L}{n}\frac{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1+\varepsilon_n} \le \frac{2\theta LM}{n(M+\varepsilon_n)}
\end{aligned}
$$

with $\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}$ ($\boldsymbol{X'}_i^T\tilde{\boldsymbol{\beta}}$) being an intermediate point between $\boldsymbol{X}_i^T\boldsymbol{\beta}$ ($\boldsymbol{X'}_i^T\boldsymbol{\beta}$) and $\boldsymbol{\beta}^{*T}\boldsymbol{X}_i$ ($\boldsymbol{\beta}^{*T}\boldsymbol{X'}_i$) from the Taylor's expansion of function $f(x) := \log(\theta+e^x)$, and the first inequality stems from

$$|f(x)| - \sup_x |g(x)| \le |f(x)-g(x)| \text{ (and take suprema over } x \text{ again).}$$

Apply McDiarmid's inequality to $Z_M(\boldsymbol{\beta}^*)$, thus we have

$$P(Z_M(\boldsymbol{\beta}^*) - \mathrm{E}Z_M(\boldsymbol{\beta}^*) \ge \lambda) \le \exp\{-\frac{n(M+\varepsilon_n)^2\lambda^2}{2M^2L^2\theta^2}\}.$$

Now, we put $\lambda \ge \frac{\theta AML}{(M+\varepsilon_n)}\sqrt{\frac{2\log(2p)}{n}}$ for $A>0$, therefore

$$P(Z_M(\boldsymbol{\beta}^*) - \mathrm{E}Z_M(\boldsymbol{\beta}^*) \ge \lambda) \le (2p)^{-A^2}. \tag{S1.15}$$

The next step is to estimate the sharper upper bounds of $\mathrm{E}Z_M(\boldsymbol{\beta}^*)$ by the symmetrization theorem and the contraction inequality below. It can be found in van der Vaart and Wellner (1996), Bühlmann and van de Geer (2011).

Lemma 5 (Symmetrization Theorem)**.** *Let $\varepsilon_1,...,\varepsilon_n$ be a Rademacher sequence with uniform distribution on $\{-1,1\}$, independent of $X_1,...,X_n$ and $f\in\mathcal{F}$. Then we have*

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\left[f(X_i)-\mathrm{E}\left\{f(X_i)\right\}\right]\right|\right] \le 2\mathrm{E}\left[\mathrm{E}_\epsilon\left\{\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\epsilon_i f(X_i)\right|\right\}\right].$$

where $\mathrm{E}[\cdot]$ refers to the expectation w.r.t. $X_1,...,X_n$ and $\mathrm{E}_\epsilon\{\cdot\}$ w.r.t. $\epsilon_1,...,\epsilon_n$.

Lemma 6 (Ledoux-Talagrand contraction)**.** *Let $x_1, ..., x_n$ be the non-random elements of $\mathcal{X}$ and $\varepsilon_1, ..., \varepsilon_n$ be Rademacher sequence. Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. Consider c-Lipschitz functions g_i, i.e.*

$$|g_i(s) - g_i(t)| \leq c\,|s-t|\,, \forall s, t \in R.$$

Then for any function $h : \mathcal{X} \to \mathbb{R}$, we have

$$\mathrm{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i\left[g_i\left\{f(x_i)\right\} - g_i\left\{h(x_i)\right\}\right]\right|\right] \leq 2c\mathrm{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i\left\{f(x_i) - h(x_i)\right\}\right|\right].$$

Note that $(\mathbb{P}_n - \mathbb{P})\left\{l_2(\boldsymbol{\beta}^*) - l_2(\boldsymbol{\beta})\right\} = \mathbb{P}_n\left\{l_2(\boldsymbol{\beta}^*) - l_2(\boldsymbol{\beta})\right\} - \mathrm{E}\{l_2(\boldsymbol{\beta}^*) - l_2(\boldsymbol{\beta})\}$, after using symmetrization theorem, the expected terms is canceled. To see contraction theorem, for

$$nZ_M(\boldsymbol{\beta}^*) = \sup_{\boldsymbol{\beta}\in S_M}\left\{\frac{|\sum_{k=i}^n \theta[\log(\theta + e^{\boldsymbol{X}^T\boldsymbol{\beta}^*}) - \log(\theta + e^{\boldsymbol{X}^T\boldsymbol{\beta}})] - n\mathrm{E}[l_2(\boldsymbol{\beta}^*) - l_2(\boldsymbol{\beta})]|}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 + \varepsilon_n}\right\}$$

as the suprema of the normalized empirical process (a local random Lipschitz constant), it is required to check the Lipschitz property of g_i in Lemma 6 with $\mathcal{F} = \mathbb{R}^p$. Let $f(x_i) := f_{\boldsymbol{\beta}}(x_i) := \frac{x_i^T\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1+\varepsilon_n}$, $h(x_i) = \frac{x_i^T\boldsymbol{\beta}}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1+\varepsilon_n}$, and $g_i(t) = \frac{\log[\theta+e^{t(\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1+\varepsilon_n)}]}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1+\varepsilon_n}$. Then

The function $g_i(t)$, $(|t| \leq LB)$ here is θ-Lipschitz. In fact

$$|g_i(s) - g_i(t)| = \frac{\theta e^{\tilde{t}}}{\theta + e^{\tilde{t}}}\cdot|s-t| \leq \theta|s-t|,\ t, s \in [-LB, LB]/(\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1+\varepsilon_n)$$

where $\tilde{t} \in [-LB, LB]/(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 + \varepsilon_n)$ is an intermediate point between t and s, by applying Lagrange mean value theorem for function $g_i(t)$.

Via the symmetrization theorem and the contraction inequality we have

$$\begin{aligned}
\mathrm{E}Z_M(\boldsymbol{\beta}^*) &\leq \frac{4\theta}{n}\mathrm{E}\left(\sup_{\beta\in\mathcal{S}_M}\left|\sum_{i=1}^n \frac{\epsilon_i\boldsymbol{X}_i^T(\boldsymbol{\beta}^* - \boldsymbol{\beta})}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 + \varepsilon_n}\right|\right) \\
&\leq \frac{4\theta}{n}\mathrm{E}\left(\sup_{\beta\in\mathcal{S}_M}\max_{1\leq j\leq p}\left|\sum_{i=1}^n \epsilon_i X_{ij}\right|\cdot\frac{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 + \varepsilon_n}\right) \\
\text{[due to } \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq M\text{]} \ &\leq \frac{4\theta M}{n(M+\varepsilon_n)}\mathrm{E}\left(\max_{1\leq j\leq p}\left|\sum_{i=1}^n \epsilon_i X_{ij}\right|\right)
\end{aligned}$$

$$= \frac{4\theta M}{n(M+\varepsilon_n)} \mathrm{E}\left(\mathrm{E}_\epsilon \max_{1\le j\le p} \left| \sum_{i=1}^{n} \epsilon_i X_{ij} \right| \right)$$

where E_ϵ is the conditional expectation $\mathrm{E}[\cdot|\mathbf{X}]$.

From Proposition 1, with $\mathrm{E}_\epsilon[\epsilon_i X_{ij}] = 0$ we get

$$\frac{4\theta M}{n(M+\varepsilon_n)} \mathrm{E}(\mathrm{E}_\epsilon \max_{1\le j\le p} |\sum_{i=1}^{n} \epsilon_i X_{ij}|) \le \frac{4\theta M}{n(M+\varepsilon_n)} \sqrt{2\log 2p} \cdot \sqrt{nL^2} = \frac{4\theta ML}{(M+\varepsilon_n)} \sqrt{\frac{2\log 2p}{n}}.$$

Thus, for $A \ge 1$ we have

$$\mathrm{E}Z_M(\boldsymbol{\beta}^*) \le \frac{4\theta ML}{(M+\varepsilon_n)} \sqrt{\frac{2\log 2p}{n}} \le \frac{4\theta AML}{(M+\varepsilon_n)} \sqrt{\frac{2\log 2p}{n}}. \tag{S1.16}$$

So we can conclude from (S1.15) and (S1.16) that

$$P(Z_M(\boldsymbol{\beta}^*) \ge \frac{5\theta AML}{(M+\varepsilon_n)} \sqrt{\frac{\log 2p}{n}}) \le P(Z_M(\boldsymbol{\beta}^*) \ge \lambda + \mathrm{E}Z_M(\boldsymbol{\beta}^*)) \le (2p)^{-A^2}. \tag{S1.17}$$

Finally, we complete the proof of Lemma 3 by letting $\frac{\lambda_1}{4} \ge \frac{5\theta AML}{(M+\varepsilon_n)} \sqrt{\frac{2\log 2p}{n}}$ and setting $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ in $Z_M(\boldsymbol{\beta}^*)$.

Proof of Proposition 1

Applying the Lagrange form of Taylor's expansion $\log(\theta + e^x) = \log(\theta + e^a) + \frac{e^{\tilde{a}}}{\theta + e^{\tilde{a}}}(x-a)$ for some real number $\tilde{a}$ between a and x, let $\boldsymbol{X}_i^T \tilde{\boldsymbol{\beta}}$ be a point between $\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_i^T \boldsymbol{\beta}^*$, i.e. $\tilde{\boldsymbol{\beta}} = \begin{pmatrix} t_1\hat{\beta}_1 \\ \vdots \\ t_p\hat{\beta}_p \end{pmatrix} + \begin{pmatrix} (1-t_1)\beta_1^* \\ \vdots \\ (1-t_p)\beta_p^* \end{pmatrix}$ for $\{t_j\}_{j=1}^p \subset [0,1]$. Observe that

$$\begin{aligned}
(\mathbb{P}_n - \mathbb{P})(l_1(\boldsymbol{\beta}^*) - l_1(\hat{\boldsymbol{\beta}})) &= \frac{-1}{n} \sum_{i=1}^{n} (Y_i - \mathrm{E}Y_i) \boldsymbol{X}_i^T [(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) - \log(\frac{\theta + \exp\{\mathbf{X}_i^T \boldsymbol{\beta}^*\}}{\theta + \exp\{\mathbf{X}_i^T \hat{\boldsymbol{\beta}}\}})] \\
&= \frac{-1}{n} \sum_{i=1}^{n} (Y_i - \mathrm{E}Y_i) \boldsymbol{X}_i^T [(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) - \frac{\exp\{\boldsymbol{X}_i^T \tilde{\boldsymbol{\beta}}\} \boldsymbol{X}_i^T (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})}{\theta + \exp\{\boldsymbol{X}_i^T \tilde{\boldsymbol{\beta}}\}}] \\
&= \frac{-1}{n} \sum_{i=1}^{n} (Y_i - \mathrm{E}Y_i) \frac{\theta \boldsymbol{X}_i^T (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})}{\theta + \exp\{\boldsymbol{X}_i^T \tilde{\boldsymbol{\beta}}\}}.
\end{aligned} \tag{S1.18}$$

If we have $\hat{\boldsymbol{\beta}} \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)$ for some finite M_0, thus $\tilde{\boldsymbol{\beta}} \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)$ via

$$\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \leq \sum_{j=1}^{p} t_j |\hat{\beta}_j - \beta_1^*| \leq \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \leq M_0,$$

Note that the random sum in (S1.18) is not independent, but the weights $\{\frac{\theta}{\theta+\exp\{\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}\}}\}_{i=1}^n$ are uniformly stochastic bounded with upper bound 1. We have to alternatively analysis the *suprema of the multiplier empirical processes* instead of $\mathcal{A}$, if we can derive some concentration inequality for the process

$$f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) := \sup_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i^c \theta \boldsymbol{X}_i^T (\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*)}{(\theta + \exp\{\boldsymbol{X}_i^T \boldsymbol{\beta}_2\}) \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\|_1} \right|.$$

with exponential decay rate, where $\boldsymbol{Y} = (Z_1, \cdots, Z_n)^T$ with $\{Y_i^c := Y_i - \mathrm{E}Y_i\}_{i=1}^n$.

In the proof below, we will verify that $f(\boldsymbol{Z}, \mathbf{X})$ is Lipschitz with respect to Euclidean norm via conditioning of design matrix $\mathbf{X}$. Then we apply the concentration inequalities of Lipschitz functions for strongly log-concave distribution distributions. We check the ℓ_2-Lipschitz condition for $f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*)$ w.r.t. $\boldsymbol{Y}$ by using the convexity of maximum function. Let $\boldsymbol{Z} = (Z_1, \cdots, Z_n)^T$ be a copy of $\boldsymbol{Y}$. Then

$$\begin{aligned}
& f_n(\boldsymbol{Z}, \mathbf{X}, \boldsymbol{\beta}^*) - f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) \\
\leq & \sup_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{X}_i^T (\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*)(Y_i^c - Z_i^c)\theta}{(\theta + \exp\{\boldsymbol{X}_i^T \boldsymbol{\beta}_2\}) \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\|_1} \right| \\
\leq & \sup_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)} \frac{1}{n} \sqrt{\sum_{i=1}^{n} \frac{[\boldsymbol{X}_i^T (\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*)\theta]^2}{[(\theta + \exp\{\boldsymbol{X}_i^T \boldsymbol{\beta}_2\}) \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\|_1]^2}} \sqrt{\sum_{i=1}^{n} (Y_i^c - Z_i^c)^2} \\
\leq & \frac{|||\mathbf{X}|||_\infty}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} (Y_i^c - Z_i^c)^2}.
\end{aligned}$$

where the second last inequality is obtained by Cauchy's inequality.

Thus the function $f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*)$ is $\frac{L}{\sqrt{n}}$-Lipschitz w.r.t. Euclidean norm of $\boldsymbol{Y}$. By using concentration inequalities of Lipschitz functions for γ-strongly log-concave discrete distributions [See Theorem C.8 in Appendix C. Note that the Theorem 3.16 in Wainwright (2019) is for continuous case],

it implies for $t > 0$

$$P(f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) - \mathrm{E}_{\boldsymbol{Y}} f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) \geq t|\mathbf{X}) \leq \exp\{-\frac{\gamma n t^2}{4|||\mathbf{X}|||_\infty^2}\}. \quad \text{(S1.19)}$$

provided that (H.4) holds. [It should be noted that (H.4) can be removed if we use Theorem 2 in Maurer and Pontil (2021) to show the sub-Gaussian concentration (S1.19).]

By (H.1): $|||\mathbf{X}|||_\infty \leq L$, we get

$$P(f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) - \mathrm{E} f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) \geq t) \leq \mathrm{E} \exp\{-\frac{\gamma n t^2}{4|||\mathbf{X}|||_\infty^2}\} \leq \exp\{-\frac{\gamma n t^2}{4L^2}\}. \quad \text{(S1.20)}$$

The details of the value γ can be founded in Appendix C.

It remains to obtain the upper bound of $\mathrm{E} f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*)$ which is proved by the symmetrization theorem with difference functions.

Lemma 7 (Symmetrization theorem with difference functions)**.** *Let $\varepsilon_1, ..., \varepsilon_n$ be a Rademacher sequence with uniform distribution on $\{-1, 1\}$, independent of $Y_1, ..., Y_n$ and $g_i \in \mathcal{G}_i$. Then we have*

$$\mathrm{E}\left(\sup_{g_1,\cdots,g_n \in \mathcal{G}_1,\cdots,\mathcal{G}_n} \left|\sum_{i=1}^n \left[g_i(Y_i) - \mathrm{E}\left\{g_i(Y_i)\right\}\right]\right|\right) \leq 2\mathrm{E}\left[\mathrm{E}_\epsilon \sup_{g_1,\cdots,g_n \in \mathcal{G}_1,\cdots,\mathcal{G}_n} \left|\sum_{i=1}^n \epsilon_i g_i(Y_i)\right|\right].$$

where $\mathrm{E}[\cdot]$ refers to the expectation w.r.t. $X_1, ..., X_n$ and $\mathrm{E}_\epsilon\{\cdot\}$ w.r.t. $\epsilon_1, ..., \epsilon_n$.

Proof. Let $\{Y_i'\}_{i=1}^n$ be an independent copy of $\{Y_i\}_{i=1}^n$. The E' denote the exportation w.r.t. $(Y_i')_{i=1}^n$, then let $\mathcal{F}_n' = \sigma(Y_1', \cdots, Y_n')$. So

$$\begin{aligned}
&\mathrm{E}\left(\sup_{g_1,\cdots,g_n \in \mathcal{G}_1,\cdots,\mathcal{G}_n} \left|\sum_{i=1}^n \left[g_i(Y_i) - \mathrm{E}\left\{g_i(Y_i)\right\}\right]\right|\right) \\
&= \mathrm{E}\left(\sup_{g_1,\cdots,g_n \in \mathcal{G}_1,\cdots,\mathcal{G}_n} \left|\mathrm{E}' \sum_{i=1}^n [g_i(X_t) - g_i(Y_i')] | \mathcal{F}_n'\right|\right) \\
&\leq \mathrm{E}\left(\sup_{g_1,\cdots,g_n \in \mathcal{G}_1,\cdots,\mathcal{G}_n} \mathrm{E}' \left|\sum_{i=1}^n [g_i(Y_t) - g_i(Y_i')]\right| | \mathcal{F}_n'\right) \text{ (Jensen's inequality of absolute function)} \\
&\leq \mathrm{E}\left(\mathrm{E}' \sup_{g_1,\cdots,g_n \in \mathcal{G}_1,\cdots,\mathcal{G}_n} \left|\sum_{i=1}^n [g_i(Y_t) - g_i(Y_i')]\right| | \mathcal{F}_n'\right) \text{ (Jensen's inequality of max function)} \\
&= \mathrm{E}\left(\sup_{f_1,\cdots,f_n \in \mathcal{G}_1,\cdots,\mathcal{G}_n} \left|\sum_{i=1}^n [g_i(Y_t) - g_i(Y_i')]\right|\right),
\end{aligned}$$

$$= \mathrm{E}\left(\sup_{g_1,\cdots,g_n \in \mathcal{G}_1,\cdots,\mathcal{G}_n} \left| \sum_{i=1}^{n} \varepsilon_i \left(g_i\left(Y_i\right) - g_i\left(Y_i'\right) \right) \right| \right) \leq 2\mathrm{E}\left[\mathrm{E}_\epsilon \sup_{g_1,\cdots,g_n \in \mathcal{G}_1,\cdots,\mathcal{G}_n} \left| \sum_{i=1}^{n} \epsilon_i g_i(\, Y_i) \right| \right],$$

where the last equality is from $\varepsilon_i[g_i\left(Y_i\right) - g_i\left(Y_i'\right)] \stackrel{d}{=} g_i\left(Y_i\right) - g_i\left(Y_i'\right)$, and the referred Jensen's inequalities are conditional expectation version. □

Conditioning on $\mathbf{X}$, then Lemma 7 implies by letting $g_i(Y_i) = \frac{Y_i \theta \boldsymbol{X}_i^T (\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*)}{(\theta + \exp\{\boldsymbol{X}_i^T \boldsymbol{\beta}_2\}) \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\|_1}$,

$$\begin{aligned}
\mathrm{E} f_n[(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*)|\mathbf{X}] &\leq \frac{2}{n}\mathrm{E}\left(\sup_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)} \left| \sum_{i=1}^{n} \frac{\epsilon_i Y_i \theta \boldsymbol{X}_i^T (\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*)}{(\theta + \exp\{\boldsymbol{X}_i^T \boldsymbol{\beta}_2\}) \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\|_1} \right| |\mathbf{X} \right) \\
&\leq \frac{2}{n}\mathrm{E}\left(\sup_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)} \max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \epsilon_i Y_i X_{ij} \right| \cdot \frac{\theta}{\theta + \exp\{\boldsymbol{X}_i^T \boldsymbol{\beta}_2\}} |\mathbf{X} \right) \\
&\leq \frac{2}{n}\mathrm{E}\left(\max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \epsilon_i Y_i X_{ij} \right| |\mathbf{X} \right).
\end{aligned}$$

Now we are going to use a maximal inequality mentioned in Lemma 14.14 in Bühlmann and van de Geer (2011), and we will give a proof in end of Appendix S2.

Proposition 1 (Maximal inequality)**.** *Let $X_1, ..., X_n$ be independent random vector that takes on a value in a measurable space $\mathcal{X}$ and $f_1, ..., f_n$ real-valued functions on $\mathcal{X}$ which satisfies for all $j = 1, ..., p$ and all $i = 1, ..., n$*

$$\mathrm{E} f_j(X_i) = 0, \quad |f_j(X_i)| \leq a_{ij}.$$

Then

$$\mathrm{E}\left(\max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} f_j(X_i) \right| \right) \leq \sqrt{2\log(2p)} \max_{1 \leq j \leq p} \sqrt{\sum_{i=1}^{n} a_{ij}^2}.$$

By Proposition 1, with $\mathrm{E}[\epsilon_i Y_i X_{ij} | \mathbf{X}, \boldsymbol{Y}] = 0$ we get

$$\begin{aligned}
\mathrm{E} f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) = \mathrm{E}[\mathrm{E} f_n[(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*)|\mathbf{X}]] &= \frac{2}{n}\mathrm{E}\left(\mathrm{E}_\epsilon \left[\max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \epsilon_i Y_i X_{ij} \right| |\boldsymbol{Y}, \mathbf{X} \right] \right) \\
&\leq \frac{2}{n}\sqrt{2\log 2p}\mathrm{E}\left(\sqrt{\sum_{i=1}^{n} (Y_i X_{ij})^2} \right) \\
\text{[By Jensen's inequality and (H.1)]} \ &\leq \frac{2L}{n}\sqrt{2\log 2p} \sqrt{\mathrm{E}\left(\sum_{i=1}^{n} Y_i^2 \right)}
\end{aligned}$$

$$\leq \frac{2L}{n}\sqrt{2\log 2p}\sqrt{n\tilde{C}_{LB}^2} = 2L\tilde{C}_{LB}\sqrt{\frac{2\log 2p}{n}}$$

where the last inequality stems from

$$\mathrm{E}(Y_i^2|\boldsymbol{X}_i) = \mathrm{Var}(Y_i|\boldsymbol{X}_i)+[\mathrm{E}(Y_i|\boldsymbol{X}_i)]^2 = \mu_i+\frac{(1+\theta)\mu_i^2}{\theta} \leq \tilde{C}_{LB}^2 =: e^{LB}+\frac{(1+\theta)e^{2LB}}{\theta},$$

using (H.1) and (H.2). Thus, we get

$$\mathrm{E}f_n(\boldsymbol{Y},\mathbf{X},\boldsymbol{\beta}^*) \leq 2L\tilde{C}_{LB}\sqrt{\frac{2\log 2p}{n}}. \tag{S1.21}$$

In equation (S1.20), if we choose $t = AL\sqrt{2\gamma}\sqrt{\frac{2\log 2p}{n}}$ such that

$$P(f_n(\boldsymbol{Y},\mathbf{X},\boldsymbol{\beta}^*) \geq t + 2L\tilde{C}_{LB}\sqrt{\frac{2\log 2p}{n}}) \leq \exp\{-\frac{\gamma n t^2}{4L^2}\} = (2p)^{-A^2}. \tag{S1.22}$$

where $A > 0$ is positive constant.

Thus with $\frac{\lambda_1}{4} \geq L(2\tilde{C}_{LB} + A\sqrt{2\gamma})\sqrt{\frac{2\log 2p}{n}}$, we have by (S1.22)

$$P\left(\sup_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\in\mathcal{S}_{M_0}(\boldsymbol{\beta}^*)}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i^c\theta\boldsymbol{X}_i^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}^*)}{(\theta+\exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}_2\})\|\boldsymbol{\beta}_1-\boldsymbol{\beta}^*\|_1}\right| \leq \frac{\lambda_1}{4}\right) \geq 1-(2p)^{-A^2}.$$

In (S1.18), observe that $\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}} \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)$, then with probability at least $1-(2p)^{-A^2}$, we have $\frac{(\mathbb{P}_n-\mathbb{P})(l_1(\boldsymbol{\beta}^*)-l_1(\hat{\boldsymbol{\beta}}))}{\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1} \leq \frac{\lambda_2}{4}$ which gives

$$P\{(\mathbb{P}_n-\mathbb{P})(l_1(\boldsymbol{\beta}^*)-l_1(\hat{\boldsymbol{\beta}})) \leq \frac{\lambda_1}{4}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1\} \geq 1-(2p)^{-A^2}.$$

S1.5 Proofs of big Theorem 3.

For this subsection, the proof techniques follow the guidelines in Wegkamp (2007), Bunea (2008).

Step1: Check $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^* \in \mathrm{V}(3.5, \frac{\varepsilon_n}{2}, H)$ from Stabil Condition

Using the mere definition of elastic-net estimate $\hat{\boldsymbol{\beta}}$, we have

$$\mathbb{P}_n l(\hat{\boldsymbol{\beta}}) + \lambda_1\sum_{j=1}^{p}|\hat{\beta}_j| + \lambda_2\sum_{j=1}^{p}|\hat{\beta}_j|^2 \leq \mathbb{P}_n l(\boldsymbol{\beta}^*) + \lambda_1\sum_{j=1}^{p}|\beta_j^*| + \lambda_2\sum_{j=1}^{p}|\beta_j^*|^2. \tag{S1.23}$$

So we obtain

$$\mathbb{P}(l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*)) \leq (\mathbb{P}_n-\mathbb{P})(l(\boldsymbol{\beta}^*)-l(\hat{\boldsymbol{\beta}}))+\lambda_1\sum_{j=1}^{p}(|\beta_j^*|-|\hat{\beta}_j|)+\lambda_2\sum_{j=1}^{p}(|\beta_j^*|^2-|\hat{\beta}_j|^2). \tag{S1.24}$$

In order to bounded the empirical process, we break down the empirical process into two parts which is or is not a function of Y_i. On the event $\mathcal{A}\cap\mathcal{B}$, the Proposition 1 and Proposition 2.21 implies.

$$\begin{aligned}&(\mathbb{P}_n-\mathbb{P})(l(\boldsymbol{\beta}^*)-l(\hat{\boldsymbol{\beta}})) = (\mathbb{P}_n-\mathbb{P})(l_1(\boldsymbol{\beta}^*)-l_1(\hat{\boldsymbol{\beta}})) + (\mathbb{P}_n-\mathbb{P})(l_2(\boldsymbol{\beta}^*)-l_2(\hat{\boldsymbol{\beta}}_n))\\ &\leq \frac{\lambda_1}{4}\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*| + \frac{\lambda_1}{4}(\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|+\varepsilon_n) = \frac{\lambda_1}{2}\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*| + \frac{\lambda_1}{4}\varepsilon_n.\end{aligned} \tag{S1.25}$$

By summing $\frac{\lambda_1}{2}\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|$ and $\lambda_2\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2$ to both sides of the inequality (S1.24), and combining with the inequality (S1.25), it gives

$$\begin{aligned}&\frac{\lambda_1}{2}\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*| + (\mathbb{P}l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*))+\lambda_2\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2\\ &\leq \lambda_1\sum_{j=1}^{p}(|\hat{\beta}_j-\beta_j^*|+|\beta_j^*|-|\hat{\beta}_j|) + \frac{\lambda_1\varepsilon_n}{4} + \lambda_2(|\beta^*|_2^2-|\hat{\beta}|_2^2) + \lambda_2\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2.\end{aligned} \tag{S1.26}$$

On the one hand, $|\hat{\beta}_j-\beta_j^*|+|\beta_j^*|-|\hat{\beta}_j| = 0$ for $j\notin H$ and $|\hat{\beta}_j|-|\beta_j^*| \leq |\hat{\beta}_j-\beta_j^*|$ for $j\in H$. On the other hand, the sum of last two terms in (S1.26) is bounded by

$$\begin{aligned}\lambda_2[(|\beta^*|_2^2-|\hat{\beta}|_2^2)+\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2] &\leq 2\lambda_2\sum_{j\in H}(|\beta_j^*|^2-\beta_j^*\hat{\beta}_j) = \lambda_2\sum_{j\in H}\beta_j^*(\beta_j^*-\hat{\beta}_j)\\ &\leq 2\lambda_2 B\sum_{j\in H}|\beta_j^*-\hat{\beta}_j| \leq \frac{1}{4}\lambda_1\sum_{j\in H}|\beta_j^*-\hat{\beta}_j|.\end{aligned}$$

due to the setting $8B\lambda_2 \leq 8B\lambda_2+4M = \lambda_1$.

Therefore the inequality (S1.26) is rewritten as

$$\begin{aligned}&\frac{\lambda_1}{2}\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*| + \mathbb{P}(l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*))+\lambda_2\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2\\ &\leq 2\lambda_1\sum_{j\in H}|\hat{\beta}_j-\beta_j^*| + \frac{\lambda_1\varepsilon_n}{4} + \frac{1}{4}\lambda_1\sum_{j\in H}|\beta_j^*-\hat{\beta}_j|.\end{aligned} \tag{S1.27}$$

Using the definition of $\boldsymbol{\beta}^*$, it implies $l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^*) + \lambda_2 \sum_{j\in H} |\hat{\beta}_j - \beta_j^*|^2 > 0$. Hence

$$\lambda_1 \sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| \leq 4.5\lambda_1 \sum_{j\in H} |\hat{\beta}_j - \beta_j^*| + \frac{\lambda_1 \varepsilon_n}{2}.$$

So we have $\sum_{j\in H^c} |\hat{\beta}_j - \beta_j^*| \leq 3.5 \sum_{j\in H} |\hat{\beta}_j - \beta_j^*| + \frac{\varepsilon_n}{2}$. Thus $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \mathrm{V}(3.5, \frac{\varepsilon_n}{2}, H)$ under the event $\mathcal{A} \cap \mathcal{B}$.

Step2: Find a lower bound for $\mathbb{P}(l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^*))$

The next proposition is a crucial result which provides a lower bound for $\mathbb{P}(l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^*))$ based on the definition of the minimizer $\boldsymbol{\beta}^*$.

Proposition 2 (Quadratic lower bound for the expected discrepancy loss)**.** *Under the (H.1) and (H.3), we have*

$$\mathbb{P}(l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^*)) \geq a\mathrm{E}^*[\mathbf{X}^{*T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2$$

with $a := \min_{\{|x|\leq L(M+B), |y|\leq K\}} \{\frac{1}{2} \frac{\theta \mathrm{e}^x (\mathrm{e}^y + \theta)}{[\theta + \mathrm{e}^x]^2}\}$.

Proof. Let $\boldsymbol{X}^{*T}\tilde{\boldsymbol{\beta}}$ is an intermediate point between $\boldsymbol{X}^{*T}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}^*\boldsymbol{\beta}^*$ given by the first order Taylor's expansion of $l(Y, \boldsymbol{X}, \boldsymbol{\beta}) = Y\boldsymbol{X}^T\boldsymbol{\beta} - (\theta + Y)\log(\theta + e^{\boldsymbol{X}^T\boldsymbol{\beta}})$, we have by the fact that $\boldsymbol{X}^*$ is an independent copy of $\boldsymbol{X}$:

$$\begin{aligned}
&\mathbb{P}(l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^*)) = \mathrm{E}[\mathrm{E}\{l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}^*)|\boldsymbol{X}\}]|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \mathrm{E}^*\{\mathrm{E}\{l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}^*)|\boldsymbol{X}^*\}]|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \\
&= \mathrm{E}^*\mathrm{E}\left\{[Y\boldsymbol{X}^{*T}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) + (Y+\theta)[\log(\theta + e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}}) - \log(\theta + e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}^*})]|\boldsymbol{X}^*\right\}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \\
&= \mathrm{E}^*\left[\mathrm{E}(Y|\boldsymbol{X}^*)\boldsymbol{X}^{*T}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) + (\mathrm{E}(Y|\boldsymbol{X}^*) + \theta)[\log(\theta + e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}}) - \log(\theta + e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}^*})\right]|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \\
&= \mathrm{E}^*[e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}^*}\boldsymbol{X}^{*T}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) - e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}^*}\boldsymbol{X}^{*T}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) + \frac{\theta e^{\boldsymbol{X}^{*T}\tilde{\boldsymbol{\beta}}}(e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}^*} + \theta)}{2(\theta + e^{\boldsymbol{X}^{*T}\tilde{\boldsymbol{\beta}}})^2}[\boldsymbol{X}^{*T}(\boldsymbol{\beta}^* - \boldsymbol{\beta})]^2]|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \\
&= \mathrm{E}^*[\frac{\theta \mathrm{e}^{\boldsymbol{X}^{*T}\tilde{\boldsymbol{\beta}}}(e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}^*} + \theta)}{2(\theta + e^{\boldsymbol{X}^{*T}\tilde{\boldsymbol{\beta}}})^2}[\boldsymbol{X}^{*T}(\boldsymbol{\beta}^* - \boldsymbol{\beta})]^2]|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}
\end{aligned}$$

Then, by triangle inequality and the definition of restricted parameter space $\mathcal{S}_M$, we obtain by (H.1) and (H.2)

$$|\boldsymbol{X}_i^{*T}\tilde{\boldsymbol{\beta}}| \leq |\boldsymbol{X}_i^{*T}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*| + |\boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*| \leq |\boldsymbol{X}_i^{*T}\hat{\boldsymbol{\beta}} - \boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*| + |\boldsymbol{X}_i^{T}\boldsymbol{\beta}^*| \leq L(M+B). \tag{S1.28}$$

Thus we conclude

$$\mathbb{P}(l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*)) = \mathrm{E}^*[\frac{\theta \mathrm{e}^{\boldsymbol{X}^{*T}\tilde{\boldsymbol{\beta}}}(e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}^*}+\theta)}{2(\theta+e^{\boldsymbol{X}^{*T}\tilde{\boldsymbol{\beta}}})^2}[\boldsymbol{X}^{*T}(\boldsymbol{\beta}^*-\boldsymbol{\beta})]^2]|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \geq a\mathrm{E}^*[\boldsymbol{X}^{*T}(\boldsymbol{\beta}^*-\hat{\boldsymbol{\beta}})]^2$$

by letting $a := \inf_{\{|x|\leq L(M+B),|y|\leq LB\}}\{\frac{1}{2}\frac{\theta \mathrm{e}^x(\mathrm{e}^y+\theta)}{[\theta+\mathrm{e}^x]^2}\} > 0$. □

From Propositon 2 and (S1.27) we deduce that

$$\lambda_1\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|+a\mathrm{E}^*[\boldsymbol{X}^{*T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2+2\lambda_2\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2 \leq 4.5\lambda_1\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|+\frac{\lambda_1\varepsilon_n}{2}. \tag{S1.29}$$

Step3: Derivations of error bounds from Stabil Condition

Let $\boldsymbol{\Sigma} = \mathrm{E}(\boldsymbol{X}^*\boldsymbol{X}^{*T})$ be the expected $p \times p$ covariance matrix. Taking expectation w.r.t. $\boldsymbol{X}^*$ only, we have the expected prediction error:

$$\mathrm{E}^*[\boldsymbol{X}^{*T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*).$$

Since $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^* \in \mathrm{V}(3.5, \frac{\varepsilon_n}{2}, H)$ is verified under the event $\mathcal{A}\cap\mathcal{B}$. Multiplying by the constant a, we have

$$a(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) \geq ak\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2-\frac{\varepsilon_n}{2}a.$$

Then substitute the above inequality to (S1.29),

$$\lambda_1\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|+ak\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2+2\lambda_2\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2 \leq 4.5\lambda_1\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|+\frac{\varepsilon_n(\lambda_1+a)}{2}.$$

By using Cauchy-Schwarz inequality, we get

$$\lambda_1\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|+(ak+2\lambda_2)\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2 \leq 4.5\lambda_1\sqrt{d_H^*}\sqrt{\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2}+\frac{\varepsilon_n(\lambda_1+a)}{2}. \tag{S1.30}$$

Apply the elementary inequality $2xy \leq Tx^2+y^2/T$ to (S1.30) for all $t > 0$, it leads to

$$\lambda_1\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|+(a_nk+2\lambda_2)\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2 \leq 2.25^2T\lambda_1^2d_H^*+\frac{1}{T}\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2+\frac{\varepsilon_n(\lambda_1+a)}{2}. \tag{S1.31}$$

We choice $T = \frac{1}{a_n k + 2\lambda_2}$ in (S1.31), we obtain

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 := \sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| \leq \frac{2.25^2 \lambda_1 d_H^*}{ak + 2\lambda_2} + (1 + \frac{a}{\lambda_1})\varepsilon_n.$$

For the square prediction error, we deduce from (S1.29) by dropping the term $2\lambda_2 \sum_{j \in H} |\hat{\beta}_j - \beta_j^*|^2$

$$\lambda_1 \sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| + a\mathrm{E}^*[\boldsymbol{X}^{*T}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})]^2 \leq 4.5\lambda_1 (\sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| - \sum_{j \in H^c} |\hat{\beta}_j - \beta_j^*|) + \frac{\lambda_1 \varepsilon_n}{2}. \tag{S1.32}$$

Then using the upper bounds of $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$, it derives

$$\begin{aligned} a\mathrm{E}^*[\boldsymbol{X}^{*T}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})]^2 &\leq 3.5\lambda_1 (\sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*|) + \frac{\lambda_1 \varepsilon_n}{2} \\ &\leq [\frac{3.5 \cdot 2.25^2 \lambda_1^2 d_H^*}{ak + 2\lambda_2} + 3.5\lambda_1 \varepsilon_n + 3.5a\varepsilon_n] + \frac{\lambda_1 \varepsilon_n}{2}. \end{aligned}$$

Note that the term $\sum_{j \in H^c} |\hat{\beta}_j - \beta_j^*| = \sum_{j \in H^c} |\hat{\beta}_j|$ that we have discarded in the right-hand side of (S1.32), it is very small for $j \in H^c$. Thus we have

$$\mathrm{E}^*[\boldsymbol{X}^{*T}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})]^2 \leq \frac{17.71875 d_H^* \lambda_1^2}{a(ak + 2\lambda_2)} + (\frac{4\lambda_1}{a} + 3.5)\varepsilon_n.$$

Finally we conclude the proof using Proposition 2.

S1.6 Proof of Theorem 4

The next lemma for estimating grouping effect inequality which is easily proved when we detailedly analyze the KKT conditions.

Lemma 8. *Let $\hat{\boldsymbol{\beta}}$ be the elastic-net estimate of NBR defined in* (1)*. Suppose that $\lambda_2 > 0$. Then for any $k, l \in \{1, 2, \ldots, p\}$,*

$$|\hat{\beta}_k - \hat{\beta}_l| \leq \frac{1}{2n\lambda_2} \sum_{i=1}^{n} \frac{\theta \, |x_{ik} - x_{il}| \, |e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}} - Y_i|}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}}. \tag{S1.33}$$

Then, we show the asymptotical version of grouping effect inequality as $p, n \to \infty$.

When deriving the grouping effect inequality from ℓ_1-estimation error, we need to bound some random sums by WLLN (weak law of large numbers) with high probability.

Lemma 9. *Assume that (C.1) and (C.3) is true, then*

(1). *Let* $S_n = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \mathrm{E}Y_i|^2$, *we have* $\mathrm{E}S_n \le \mu$ *for some constant* μ;

(2). *The square of centered responses have finite variance with a common bound, i.e.* $\max_{1\le i\le n}\{\mathrm{Var}|Y_i - \mathrm{E}Y_i|^2\} \le \sigma^2$ *for some constant* σ^2.

The proof of Lemma 9 is straightforward which is given in Appendix B, and we present the proof of Theorem 4 in advance.

By Lemma 8, Cauchy inequality, triangle inequality and Taylor expansion, we have

$$\begin{aligned}
|\hat{\beta}_k - \hat{\beta}_l|^2 &\le (\frac{1}{2n\lambda_2}\sum_{i=1}^{n} |X_{ik} - X_{il}||e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}} - Y_i|)^2 \\
&\le \frac{1}{4\lambda_2^2}\cdot\frac{1}{n}\sum_{i=1}^{n}|X_{ik} - X_{il}|^2 \cdot \frac{1}{n}\sum_{i=1}^{n}|e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}} - Y_i|^2 \\
&= \frac{1}{4\lambda_2^2}\cdot 2(1-\rho_{kl})\frac{1}{n}\sum_{i=1}^{n}|e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}} - e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*} + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*} - Y_i|^2 \\
&\le \frac{1}{4\lambda_2^2}\cdot 2(1-\rho_{kl})\{\frac{2}{n}\sum_{i=1}^{n}|e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}} - e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}|^2 + \frac{2}{n}\sum_{i=1}^{n}|e^{\boldsymbol{X}_i{}^T\boldsymbol{\beta}^*} - Y_i|^2\} \\
&= \frac{1}{4\lambda_2^2}\cdot 2(1-\rho_{kl})\{\frac{2}{n}\sum_{i=1}^{n}e^{2\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}}|\boldsymbol{X}_i^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|^2 + \frac{2}{n}\sum_{i=1}^{n}|e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*} - Y_i|^2\}.
\end{aligned}$$

where the last inequality is due to $(a+b)^2 \le 2(a^2+b^2)$.

Under the assumption of oracle inequality (2.15), with probability $1 - \frac{2}{p^{r-1}} - \exp\{-\frac{nt^2}{2C_{LB}^2 d_c^2 L^4}\}$, we have

$$\begin{aligned}
|\hat{\beta}_k - \hat{\beta}_l|^2 &\le \frac{1}{\lambda_2^2}\cdot(1-\rho_{kl})\{Ke^{2LM}O(\lambda_1^2) + \frac{1}{n}\sum_{i=1}^{n}|\mathrm{E}Y_i - Y_i|^2\} \\
&=: (1-\rho_{kl})[Ke^{2LM}O(1) + \frac{1}{\lambda_2^2}S_n].
\end{aligned}$$

For the second part, by using Chebyshev's inequality, it implies

$$P(|S_n - \mathrm{E}S_n| \le E) \ge 1 - \frac{\sigma_n^2}{nE^2} \Rightarrow S_n \le E + \mathrm{E}S_n \le E + \mu$$

with probability at least $1-\frac{\sigma_n^2}{nE^2}$ in the event $\mathcal{C}(E) =: \{S_n \le E+\mu\}$.

Then, on the three events $\mathcal{K}\cap\mathcal{E}_c\cap\mathcal{C}(E)$, we have

$$|\hat{\beta}_k-\hat{\beta}_l|^2 \le (1-\rho_{kl})[Ke^{2LM}O(1)+\frac{1}{\lambda_2^2}(E+\mu)]$$

with probability $P(\mathcal{K}\cap\mathcal{E}_c\cap\mathcal{C}(E)) \ge 1-2p^2e^{-\frac{nt^2}{2[d_H^* C_{LB}(1+\zeta)L^2]^2}}-\frac{\sigma^2}{nE^2}$.

Moreover, if $1-\rho_{kl}=o_p(\lambda_2^2)$, we have

$$|\hat{\beta}_k-\hat{\beta}_l| \le \sqrt{o_p(1)[\lambda_2^2 e^{2LM}O(1)+(E+\mu)]}.$$

S1.7 Proof of Theorem 5

By KKT condition (see Lemma 1 and (2.3) in the main body), then we claim that $\text{sgn}\hat{\boldsymbol{\beta}}=\text{sgn}\boldsymbol{\beta}^*$ if

$$\begin{cases} \text{sign}\hat{\beta}_j=\text{sign}\beta_j^*, j\in H \\ \dot{\ell}_j(\hat{\beta})+2\lambda_2\hat{\beta}_j=-\lambda_1\text{sign}\hat{\beta}_j, \hat{\beta}_j\neq 0 \\ |\dot{\ell}_j(\hat{\beta})|\le\lambda_1, \hat{\beta}_j=0 \end{cases} \tag{S1.34}$$

Let $\boldsymbol{\beta}_H=\{\beta_j, j\in H\}$ and $\hat{\beta}_H=\{\hat{\beta}_j, j\in H\}$ be the sub-vector for $\boldsymbol{\beta}$. Since $\text{sign}\hat{\beta}_j=\text{sign}\beta_j^*, j\in H$, then $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_H,0)^T$ is the solution of the KKT conditions. So, the (S1.34) holds if

$$\begin{cases} \text{sign}\hat{\beta}_j=\text{sign}\beta_j^*, j\in H \\ |\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H)|\le\lambda_1, j\notin H \end{cases} \Leftarrow \begin{cases} |\hat{\beta}_j-\beta_j^*|<|\beta_j^*|, j\in H \\ |\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H)|\le\lambda_1, j\notin H \end{cases} \tag{S1.35}$$

where $\hat{\boldsymbol{\beta}}_H$ is the solution of $\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H)+2\lambda_2\hat{\beta}_j=-\lambda_1\text{sign}\beta_j^*, j\in H$.

Notice that the right expression in (S1.35) holds if

$$\begin{cases} \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1<\beta_*:=\min\{|\beta_j|: j\in H\} \\ |\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H)|\le\lambda_1, j\notin H \end{cases}$$

Let $\eta\in(0,1)$, the above events hold if

$$\begin{cases} E_1: \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1<\beta_*, \\ E_2: \max\limits_{j\notin H}|\dot{\ell}_j(\boldsymbol{\beta}^*)|\le\eta\lambda_1, \\ E_3: \max\limits_{j\notin H}|\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H)-\dot{\ell}_j(\boldsymbol{\beta}^*)|\le(1-\eta)\lambda_1, \end{cases}$$

which is from the triangle inequality $|\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H)|\le|\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H)-\dot{\ell}_j(\boldsymbol{\beta}^*)|+|\dot{\ell}_j(\boldsymbol{\beta}^*)|$.

Let $E = E_1 \cap E_2 \cap E_3$, we want to show that each event in $E_i, i = 1, 2, 3$ holds with high probability. And we utilize the basic sets inequality $P(E) \geq P(E_1) + P(E_2) + P(E_3) - 2$. Put $\boldsymbol{X}_{iH} = (\cdots, \tilde{x}_{ih}, \cdots)^T$ with $\tilde{x}_{ih} = x_{ih}$ if $h \in H$ and $\tilde{x}_{ih} = 0$ if $h \notin H$.

For E_1, by Theorem 2, we have

$$P(E_1) \geq P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{e^{2a_\tau}(\zeta+1)d_H^*\lambda_1}{2C_t^2(\zeta, H)}) \geq 1 - \frac{2}{p^{r-1}} - 2p^2 e^{-\frac{nt^2}{2[d_H^* C_{LB}(1+\zeta)L^2]^2}}. \tag{S1.36}$$

For E_2, thus we get

$$\begin{aligned} P(E_2) &= P(\max_{j \notin H} |\dot{\ell}_j(\boldsymbol{\beta}^*)| \geq \eta\lambda_1) \leq \sum_{j \notin H} P\left(\left|\sum_{i=1}^n \frac{x_{ij}(Y_i - \mathrm{E}Y_i)\theta}{n(\theta + \mathrm{E}Y_i)}\right| \geq \eta\lambda_1\right) \\ &\leq 2p\exp\{-\frac{\eta^2\lambda_1^2}{2C_{LB}^2||\boldsymbol{w}^{(j)}||_2^2}\} \leq 2p\exp\{-\frac{\eta^2\lambda_1^2 n}{2C_{LB}^2 L^2}\}. \end{aligned}$$

where we use $||\boldsymbol{w}^{(j)}||_2^2 \leq \frac{L^2}{n}$ in Lemma 2.

This implies that

$$P(E_2) \leq \eta\lambda_1) \geq 1 - 2p\exp\{-\frac{\eta^2\lambda_1^2 n}{2C_{LB}^2 L^2}\} = 1 - \frac{2}{p^{1 - r\eta^2/C_{\xi,B_1}^2}}. \tag{S1.37}$$

where the last equality is by observing that $\lambda_1 = \frac{C_{LB}L}{C_{\xi,B_1}}\sqrt{\frac{2r\log p}{n}}$.

For E_3, note that

$$\begin{aligned} &\max_{j \notin H} |\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H) - \dot{\ell}_j(\boldsymbol{\beta}^*)| = \max_{j \notin H} |\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H) - \dot{\ell}_j(\boldsymbol{\beta}_H^*)| \\ &= \max_{j \notin H} \frac{1}{n}\left|\sum_{i=1}^n x_{ij}\theta[\frac{\theta + Y_i}{\theta + e^{\boldsymbol{X}_{iH}^T\hat{\boldsymbol{\beta}}_H}} - \frac{\theta + Y_i}{\theta + e^{\boldsymbol{X}_{iH}^T\boldsymbol{\beta}_H^*}}]\right| \\ &= \max_{j \notin H} \frac{1}{n}\left|\sum_{i=1}^n x_{ij}\frac{\theta(\theta + Y_i)e^{\boldsymbol{X}_{iH}^T\boldsymbol{\beta}_H^*}[e^{\boldsymbol{X}_{iH}^T(\hat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_H^*)} - 1]}{(\theta + e^{\boldsymbol{X}_{iH}^T\hat{\boldsymbol{\beta}}_H})(\theta + e^{\boldsymbol{X}_{iH}^T\boldsymbol{\beta}_H^*})}\right| \\ &\leq \frac{L}{n}\sum_{i=1}^n \left|(\theta + Y_i)[e^{\boldsymbol{X}_{iH}^T(\hat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_H^*)} - 1]\right| \\ &\leq \frac{L}{n}\sum_{i=1}^n \left|(\theta + Y_i)[\boldsymbol{X}_{iH}^T(\hat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_H^*) + o_p(|\boldsymbol{X}_{iH}^T(\hat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_H^*)|)]\right| \\ &\leq \frac{C_X L^2}{n}\sum_{i=1}^n |\theta + Y_i|||\hat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_H^*||_1. \end{aligned} \tag{S1.38}$$

where the second last inequality is by (2.15) and the boundedness of $|\boldsymbol{X}_{iH}^T(\hat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_H^*)|$, the last inequality (S1.38) stems from $\|\boldsymbol{X}_i\|_\infty \le L$ and C_X is determined by

$$|\boldsymbol{X}_{iH}^T(\hat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_H^*) + o_p(|\boldsymbol{X}_{iH}^T(\hat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_H^*)|)| \le LC_X||\hat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_H^*||_1.$$

Let $A_n := \frac{1}{n}\sum_{i=1}^n |\theta + Y_i|$, similar to the proof of Lemma 9, we have

$$\mathrm{E}A_n := \frac{1}{n}\sum_{i=1}^n \mathrm{E}|\theta + Y_i| \le \frac{1}{n}\sum_{i=1}^n \sqrt{\mathrm{E}|\theta + Y_i|^2} < \infty,\ \ \sigma_n^2(A) := \frac{1}{n}\sum_{i=1}^n \mathrm{Var}|\theta + Y_i| < \infty.$$

And we can find a constant $\mu(A) > 0$ such that $\mathrm{E}A_n \le \mu(A)$.

By Chebyshev's inequality $P(|A_n - \mathrm{E}A_n| \le A) \ge 1 - \frac{\sigma_n^2(A)}{nA^2}$, we get

$$A_n \le A + \mathrm{E}A_n \le A + \mu(A)$$

with probability at least $1 - \frac{\sigma_n^2(A)}{nA^2}$. Then (S1.38) turns to

$$\max_{j\notin H} |\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H) - \dot{\ell}_j(\boldsymbol{\beta}^*)| \le C_X L^2(A + \mu(A))||\hat{\beta}_H - \beta_H^*||_1$$

Under the event $\{A_n \le A + \mu(A)\}$, with probability $1 - \frac{\sigma_n^2(A)}{nA^2}$ we obtain

$$P(\max_{j\notin H} |\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H) - \dot{\ell}_j(\boldsymbol{\beta}^*)| \le (1-\eta)\lambda_1) \ge P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le \frac{(1-\eta)\lambda_1}{C_X L^2(A + \mu(A))})$$
$$= P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le \frac{e^{2a_\tau}(\zeta+1)d_H^*}{2C_t^2(\zeta, H)} \cdot \frac{2C_t^2(\zeta, H)(1-\eta)\lambda_1}{e^{2a_\tau}(\zeta+1)d_H^* C_X L^2(A + \mu(A))}).$$

Let

$$\tilde{\lambda}_1 =: \tilde{C}\lambda_1, \text{ with } \tilde{C} := \frac{2C_t^2(\zeta, H)(1-\eta)}{e^{2a_\tau}(\zeta+1)d_H^* C_X L^2(A + \mu(A))}$$

where $\lambda_1 = \frac{C_{LB}L}{C_{\xi,B_1}}\sqrt{\frac{2r\log p}{n}}$.

Since by (2.15) with high probability in Theorem 2 and (S1.8), we conclude that

$$P(E_3) \ge P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le \frac{e^{2a_\tau}(\zeta+1)d_H^*\tilde{\lambda}_1}{2C_t^2(\zeta, H)})$$
$$\ge P(|A_n - \mathrm{E}A_n| \le A) - 2p\exp\{-\frac{C_{\xi,B_1}^2\tilde{\lambda}_1^2 n}{2C_{LB}^2 C_X L^2}\} - 2p^2 e^{-\frac{nt^2}{2[d_H^* C_{LB}(1+\zeta)L^2]^2}}$$

$$\geq 1 - \frac{\sigma_n^2(A)}{nA^2} - \frac{2}{p^{1-r\tilde{C}^2}} - 2p^2 e^{-\frac{nt^2}{2[d_H^* C_{LB}(1+\zeta)L^2]^2}}. \tag{S1.39}$$

Combining (S1.36),(S1.37) and (S1.39), we get

$$P(\text{sign}\hat{\boldsymbol{\beta}} = \text{sign}\boldsymbol{\beta}^*) \geq 1 - \frac{2}{p^{r-1}} - 4p^2 e^{-\frac{nt^2}{2[d_H^* C_{LB}(1+\zeta)L^2]^2}} - \frac{2}{p^{1-r\eta^2/C_{\xi,B_1}^2}} - \frac{\sigma_n^2(A)}{nA^2} - \frac{2}{p^{1-r\tilde{C}^2}}.$$

Without loss of generality, we assume that $r, \tilde{C}^2 r, r\eta^2/C_{\xi,B_1}^2 > 1$ since r is a tuning parameter. Let $p, n \to \infty$, it leads to sign consistency:

$$P(\text{sign}\hat{\boldsymbol{\beta}} = \text{sign}\boldsymbol{\beta}^*) \to 1.$$

S1.8 Proof of Proposition 3

Given sample size n, Bunea (2008) studied conditions under which $P(H \subset \hat{H}) \geq 1 - \delta$ for the number of parameters p and confidence $1 - \delta$ by the following lemma.

Lemma 10. *(Lemma 3.1 in Bunea (2008)) For any true parameter $\boldsymbol{\beta}^*$ and for any estimate $\hat{\boldsymbol{\beta}}$, we have $P(H \not\subset \hat{H}) \leq P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \geq \min_{j \in H} |\beta_j^*|)$.*

Based on the lemma above, we give the proof of Proposition 3.

Proof. Note that

$$P(\mathcal{A} \cap \mathcal{B}) \geq 1 - 2(2p)^{-A^2}.$$

Solving $2(2p)^{-A^2} = \delta/p$ for p, we have $p = \exp\{\frac{1}{A^2-1} \log \frac{2^{1-A^2}}{\delta}\}$ with $A > 1$. Then

$$P(H \subset \hat{H}) \geq P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \min_{j \in H} |\beta_j^*|) \geq P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq B_0) \geq 1 - \delta/p$$

which is directly followed from Lemma 10. □

S1.9 Proof of Theorem 6

The following lemma is a fancy and tractable event by virtue of KKT condition. It derives a nice bound of $P(H \not\subset \hat{H})$, yet is worthy of to be singled out here.

Lemma 11 (Proposition 3.3 in Bunea (2008))**.**

$$P(H \not\subset \hat{H}) \leq d_H^* \max_{k \in H} P(\hat{\beta}_k = 0 \text{ and } \beta_k^* \neq 0).$$

Consider the KKT condition of $\{\hat{\beta}_k = 0\}$ (Lemma 1). That is, $\{\hat{\beta}_k = 0\}$ is a solution of (1) iff $\hat{\beta}_k$ satisfies

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_{ik} \frac{\theta(e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}} - Y_i)}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} \right| \leq \lambda_1 \quad , k = 1, 2, \ldots, p.$$

Next, the proof of Theorem 6 is divided into two steps. The key fact adopted in theoretical analysis in Step1 is that, when decomposing the nth partial sum in the KKT conditions, one must split it into four partial sum. The event of each one in sums whose absolute value exceeds the tuning parameter λ_1, is asymptotically high-dimensional negligible. The decomposing method goes back to Bunea (2008) who deal with linear and logistic regression, and our decomposition for NBR is different from linear and Logistic cases.

Step1: Find $P(H \not\subset \hat{H})$.

By Lemma 11, we have

$$\begin{aligned}
&P(H \not\subset \hat{H}) \leq d_H^* \max_{k \in H} P(\hat{\beta}_k = 0 \text{ and } \beta_k^* \neq 0) \\
&= d_H^* \max_{k \in H} P(\left| \frac{1}{n} \sum_{i=1}^{n} X_{ik} \frac{\theta(e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}} - Y_i)}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} \right| \leq \lambda_1; \beta_k^* = 0) \\
&= d_H^* \max_{k \in H} P(\frac{1}{n} | \sum_{i=1}^{n} X_{ik} \theta \{ (\frac{e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} - \frac{e^{\boldsymbol{X}_i^T \boldsymbol{\beta^*}}}{\theta + e^{\boldsymbol{X}_i^T \boldsymbol{\beta^*}}}) + (\frac{Y_i}{\theta + e^{\boldsymbol{X}_i^T \boldsymbol{\beta^*}}} - \frac{Y_i}{\theta + e^{\boldsymbol{X}_i{}^T \hat{\boldsymbol{\beta}}}}) \\
&- \frac{Y_i - e^{\boldsymbol{X}_i^T \boldsymbol{\beta^*}}}{\theta + e^{\boldsymbol{X}_i^T \boldsymbol{\beta^*}}} \} | \leq \lambda_1; \beta_k^* = 0)
\end{aligned}$$

Let

$$\begin{aligned}
A_n^{(k)} &= \frac{1}{n} \sum_{i=1}^{n} X_{ik} \theta (\frac{e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} - \frac{e^{X_i^T \boldsymbol{\beta^*}}}{\theta + e^{\boldsymbol{X}_i^T \boldsymbol{\beta^*}}}), \\
C_n^{(k)} &= \frac{1}{n} \sum_{i=1}^{n} X_{ik} \theta (\frac{Y_i}{\theta + e^{\boldsymbol{X}_i^T \boldsymbol{\beta^*}}} - \frac{Y_i}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}}), \\
D_n^{(k)} &= \frac{1}{n} \sum_{i=1}^{n} X_{ik} \theta (\frac{Y_i - e^{\boldsymbol{X}_i^T \boldsymbol{\beta^*}}}{\theta + e^{\boldsymbol{X}_i^T \boldsymbol{\beta^*}}}), \quad B_n^{(k)} = \sum_{j=1}^{p} (\hat{\beta}_j - \beta_j^*) \frac{\theta}{n} \sum_{i=1}^{n} X_{ik} X_{il},
\end{aligned}$$

thus with $\{\beta_k^* = 0\}$ and assumption $\frac{\theta}{n}\sum_{i=1}^{n} X_{ik}^2 = 1$, we have

$$\begin{aligned}\left|B_n^{(k)}\right| &= |(\hat{\beta}_k - \beta_k^*)\frac{\theta}{n}\sum_{i=1}^{n} X_{ik}^2 + \sum_{j\neq k}^{p}(\hat{\beta}_j - \beta_j^*)\frac{\theta}{n}\sum_{i=1}^{n} X_{ij}X_{ik}| \\ &\geq |\hat{\beta}_k| - |\sum_{j\neq k}^{p}(\hat{\beta}_j - \beta_j^*)\frac{\theta}{n}\sum_{i=1}^{n} X_{ij}X_{ik}|.\end{aligned}$$

Let $\tilde{B}_n^{(k)} := \sum_{j\neq k}^{p}(\hat{\beta}_j - \beta_j^*)\frac{\theta}{n}\sum_{i=1}^{n} X_{ij}X_{ik}$, thus

$$|B_n^{(k)}| \geq \min_{j\in H}|\beta_j^*| - |\tilde{B}_n^{(k)}| \geq 2\lambda_1 - |\tilde{B}_n^{(k)}| \tag{S1.40}$$

Together with the above notation, we obtain

$$\begin{aligned}&P(H \not\subset \hat{H}) \leq d_H^* \max_{k\in H} P(|B_n^{(k)} + A_n^{(k)} - B_n^{(k)} + C_n^{(k)} - D_n^{(k)}| \leq \lambda_1; \beta_k^* = 0) \\ &\leq d_H^* \max_{k\in H} P(|B_n^{(k)}| - |A_n^{(k)} - B_n^{(k)}| - |C_n^{(k)}| - |D_n^{(k)}| \leq \lambda_1; \beta_k^* = 0) \\ &\leq d_H^* \max_{k\in H} P(2\lambda_1 - |\tilde{B}_n^{(k)}| - |A_n^{(k)} - B_n^{(k)}| - |C_n^{(k)}| - |D_n^{(k)}| \leq \lambda_1; \beta_k^* = 0) \\ &= d_H^* \max_{k\in H}\{P(|\tilde{B}_n^{(k)}| + |A_n^{(k)} - B_n^{(k)}| + |C_n^{(k)}| + |D_n^{(k)}| \geq \lambda_1\} \\ &\leq d_H^* \max_{k\in H}\{P(|\tilde{B}_n^{(k)}| \geq \frac{\lambda_1}{4}) + P(|A_n^{(k)} - B_n^{(k)}| \geq \frac{\lambda_1}{4}) + P(|C_n^{(k)}| \geq \frac{\lambda_1}{4}) + P(|D_n^{(k)}| \geq \frac{\lambda_1}{4})\}.\end{aligned}$$

To bound the first probability inequality, for $i = 1, 2$, we assume that $\frac{1}{4hL_i} \geq \frac{2.25^2}{ak+2\lambda_2}$, where k is defined by Identifiable Condition and constant a is given in Theorem 3. Next, we will apply the following lemma.

Lemma 12 (Lemma 2.1 in Bunea (2008))**.** *Given the constants $k > 0, \varepsilon \geq 0$ defined in Definition 2, if Identifiable Condition holds for some $0 < h < \frac{1}{1+2c+\varepsilon}$, then the Stabil Condition with measurement error is true for any $0 < k < 1 - h(1 + 2c + \varepsilon)$.*

By Lemma 12 with $\varepsilon_n = 0$, Identifiable Condition derives Stabil Condition with $k \leq 1 - 8h$ since Theorem 3 shows that $c = 3.5$. By solving a system of two inequalities: $\frac{1}{4h} \geq \frac{2.25^2}{ak+2\lambda_2}, k \leq 1 - 8h$, it implies $h \leq \frac{ak+2\lambda_2}{20.25+8a} \wedge \frac{1}{8}$. Applying Identifiable Condition and provability bound in Proposition 3, we therefore have

$$P(|\tilde{B}_n^{(k)}| \geq \frac{\lambda_1}{4}) \leq P(\sum_{j\neq k}^{p}|\hat{\beta}_j - \beta_j^*||\frac{\theta}{n}\sum_{i=1}^{n} X_{ij}X_{ik}| \geq \frac{\lambda_1}{4})$$

$$\leq P(\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|\geq\frac{\lambda_1 d_H^*}{4h})$$
$$\leq P(\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|\geq\frac{2.25^2\lambda_1 d_H^*}{ak+2\lambda_2})\leq\frac{\delta}{p}. \tag{S1.41}$$

For the second probability, $P(|A_n^{(k)}-B_n^{(k)}|\geq\frac{\lambda_1}{4})$, by Taylor's expansion, we have

$$A_n^{(k)}=\sum_{j=1}^{p}(\hat{\beta}_j-\beta_j^*)\frac{1}{n}\sum_{i=1}^{n}\frac{X_{ik}X_{ij}\cdot\theta^2 e^{a_i}}{(\theta+e^{a_i})^2},$$

where a_i be the intermediate point between $\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_i^T\boldsymbol{\beta}^*$.

So solving a system of two inequalities: $\frac{1}{4L_1h}\geq\frac{2.25^2}{ak+2\lambda_2}, k\leq 1-8h$, we get $h\leq\frac{ak+2\lambda_2}{20.25L_1+8a}\wedge\frac{1}{8}$. Then,

$$|A_n^{(k)}-B_n^{(k)}|=|\sum_{j=1}^{p}(\hat{\beta}_j-\beta_j^*)\frac{1}{n}\sum_{i=1}^{n}\theta X_{ik}X_{ij}\cdot(1-\frac{\theta e^{a_i}}{(\theta+e^{a_i})^2})|$$
$$\leq\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*||\frac{1}{n}\sum_{i=1}^{n}\theta X_{ik}X_{ij}\cdot(1-\frac{\theta e^{a_i}}{(\theta+e^{a_i})^2})|\leq\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|\frac{hL_1}{d_H^*},$$

where the last inequality is by using WCC(1).

Therefore, by the same argument like $|\tilde{B}_n^{(k)}|$, we have

$$P(|A_n^{(k)}-B_n^{(k)}|\geq\frac{\lambda_1}{4})\leq P(\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|\geq\frac{2.25^2\lambda_1 d_H^*}{ak+2\lambda_2})\leq\frac{\delta}{p}, \tag{S1.42}$$

from Corollary 3.

To bound the third probability, notice that

$$C_n^{(k)}=\frac{1}{n}\sum_{i=1}^{n}X_{ik}\theta(\frac{Y_i}{\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}}-\frac{Y_i}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}})=\sum_{j=1}^{p}(\hat{\beta}_j-\beta_j^*)\frac{1}{n}\sum_{i=1}^{n}\frac{X_{ik}X_{ij}\theta Y_i\cdot e^{b_i}}{(\theta+e^{b_i})^2}.$$

Under the event WCC(2), similar derivation by solving the system of two inequalities: $\frac{1}{4L_2h}\geq\frac{2.25^2}{ak+2\lambda_2}, k\leq 1-8h$, we have $h\leq\frac{ak+2\lambda_2}{20.25L_2+8a}\wedge\frac{1}{8}$. Then

$$P(|C_n^{(k)}|\geq\frac{\lambda_1}{4})\leq P(\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|\geq\frac{2.25^2\lambda_1 d_H^*}{ak+2\lambda_2})\leq\frac{\delta}{p}. \tag{S1.43}$$

It remains to obtain the upper bound for the fourth term. This can adopt Lemma 2 by letting $w_i^{(j)} := \frac{\theta X_{ij}}{\theta+\mathrm{e}^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}}$, so $||\boldsymbol{w}^{(j)}||_2^2 := \sum_{i=1}^{n} \frac{X_{ij}^2\theta^2}{n^2(\theta+\mathrm{e}^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*})^2} \leq \frac{L^2}{n}$ and then conditioning on $\mathbf{X}$. With (S1.7), we get

$$P(|D_n^{(k)}| \geq \frac{\lambda_1}{4}) = P(|\frac{1}{n}\sum_{i=1}^{n} \frac{X_{ik}\theta}{\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}}(Y_i - \mathrm{E}Y_i)| \geq \frac{\lambda_1}{4}\} \leq 2\exp\{-\frac{n\lambda_1^2}{32C_{LB}^2L^2}\}. \tag{S1.44}$$

In summary, the four probabilities (S1.41),(S1.42),(S1.43) and (S1.44) imply

$$P(H \not\subset \hat{H}) \leq \frac{3d_H^*}{p}\delta + 2d_H^*\exp\{-\frac{n\lambda_1^2}{32C_{LB}^2L^2}\}.$$

Step2: Find $P(\hat{H} \not\subset H)$.
From the KKT conditions, we define the set

$$\mathcal{K} := \bigcap_{k\notin H}\{\left|\frac{1}{n}\sum_{i=1}^{n} X_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}} - Y_i)}{\theta + e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}\right| \leq \lambda_1\}.$$

Thus, we have $\hat{\beta}_k = 0$ if $k \notin H$. And thus $\forall k \notin H \Rightarrow k \notin \hat{H}$ which gives $\forall k \in \hat{H} \Rightarrow k \in H$. We conclude that event $\mathcal{K}$ implies $\hat{H} \subset H$. Subsequently,

$$\begin{aligned}
P(\hat{H} \not\subset H) \leq P(K^c) &\leq \sum_{k\notin H} P(\left|\frac{1}{n}\sum_{i=1}^{n} X_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}} - Y_i)}{\theta + e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}\right| \geq \lambda_1) \\
&= \sum_{k\notin H} P(|A_n^{(k)} + C_n^{(k)} - D_n^{(k)}| \geq \lambda_1) \\
&\leq \sum_{k\notin H}\{P(|A_n^{(k)}| \geq \frac{\lambda_1}{3}) + P(|C_n^{(k)}| \geq \frac{\lambda_1}{3}) + P(|D_n^{(k)}| \leq \frac{\lambda_1}{3})\} \\
&\leq \sum_{k\notin H}\{P(|A_n^{(k)}| \geq \frac{\lambda_1}{4}) + P(|C_n^{(k)}| \geq \frac{\lambda_1}{4}) + P(|D_n^{(k)}| \leq \frac{\lambda_1}{4})\} \\
&\leq \sum_{k\notin H} P(|A_n^{(k)}| \geq \frac{\lambda_1}{4}) + (p - d_H^*)[\frac{\delta}{p} + 2e^{-n\lambda_1^2/32C_{LB}^2L^2}],
\end{aligned}$$

where the last inequality is similarly obtained from (S1.43) and (S1.44).

It remains to bound the first term as the summation of $P(|A_n^{(k)}| \geq \frac{\lambda_1}{4})$.

By WCC (1) we have

$$|A_n^{(k)}| = |\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j^*)\frac{\theta}{n}\sum_{i=1}^{n} X_{ik}X_{ij}\cdot\frac{\theta e^{a_i}}{(\theta+e^{a_i})^2})|$$
$$\leq \sum_{j=1}^{p}|\hat{\beta}_j - \beta_j^*||\frac{\theta}{n}\sum_{i=1}^{n} X_{ik}X_{ij}\cdot\frac{\theta e^{a_i}}{(\theta+e^{a_i})^2})| \leq \sum_{j=1}^{p}|\hat{\beta}_j - \beta_j^*|\frac{hL_1}{d_H^*}.$$

So by the bounds in Proposition 3 we have

$$P(|A_n^{(k)}| \geq \frac{\lambda_1}{4}) \leq P(\sum_{j=1}^{p}|\hat{\beta}_j - \beta_j^*|\frac{hL_1}{d_H^*} \geq \frac{\lambda_1}{4}) \leq P(\sum_{j=1}^{p}|\hat{\beta}_j - \beta_j^*| \geq \frac{1}{4}\cdot\frac{d_H^*\lambda_1}{hL_1})$$
$$\leq P(\sum_{j=1}^{p}|\hat{\beta}_j - \beta_j^*| \geq \frac{2.25^2\lambda_1 d_H^*}{ak+2\lambda_2}) \leq \frac{\delta}{p}.$$

We conclude that

$$P(\hat{H} \not\subset H) \leq (p-d_H^*)\frac{\delta}{p} + (p-d_H^*)[\frac{\delta}{p} + 2e^{-n\lambda_1^2/18C_{LB}^2L^2}] \leq (p-d_H^*)[\frac{2\delta}{p} + 2e^{-n\lambda_1^2/32C_{LB}^2L^2}].$$

Judging from the above two steps and relation, we obtain

$$P(H=\hat{H},\ \text{WCC(2)}) \geq 1-P(H\not\subset\hat{H})-P(\hat{H}\not\subset H) \geq 1-(2+d_H^*/p)\delta-2pe^{-n\lambda_1^2/32C_{LB}^2L^2}.$$

Since WCC(2) holds with probability $1-\varepsilon_{n,p}$. By (3.27), it gives

$$P(H=\hat{H}) \geq 1-2(1+d_H^*/p)\delta - 2pe^{-n\lambda_1^2/32C_{LB}^2L^2} - \varepsilon_{n,p}.$$

S2 Assisted lemmas

We fix $Y_i = y_i$ in the proof of Lemma 1,8. Rewrite $\hat{\boldsymbol{\beta}}(\lambda_1,\lambda_2), \hat{\beta}_k(\lambda_1,\lambda_2), \hat{\beta}_l(\lambda_1,\lambda_2)$ as $\hat{\boldsymbol{\beta}}, \hat{\beta}_k$ and $\hat{\beta}_l$ respectively.

S2.1 Proof of Lemma 1

For $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$, define the following multivariate function:

$$F(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}[(\theta+y_i)\log(\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}) - y_i\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}] + \lambda_1\sum_{i=1}^{p}|\hat{\beta}_i| + \lambda_2\sum_{i=1}^{p}|\hat{\beta}_i|^2. \tag{S2.45}$$

And let $\mathbf{e}_k = (\underbrace{0, \cdots, 0, 1}_{k}, 0, \cdots, 0)$. Next, we simply write $\hat{\beta}_k(\lambda_1, \lambda_2)$ as $\hat{\beta}_k$.

Case 1. If $\hat{\beta}_k \neq 0$, for sufficiently small $\varepsilon \in (-|\hat{\beta}_k|, |\hat{\beta}_k|)$, we have

$$F(\hat{\boldsymbol{\beta}} + \varepsilon \mathbf{e}_k) - F(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} [(\theta + y_i) \log \frac{\theta + e^{\boldsymbol{X}^T(\hat{\boldsymbol{\beta}} + \varepsilon \mathbf{e}_k)}}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} - y_i x_{ik} \varepsilon] + \lambda_1(|\hat{\beta}_k + \varepsilon| - |\hat{\beta}_k|) + \lambda_2(2\hat{\beta}\varepsilon + \varepsilon^2).$$

Notice that the ranges of ε, we obtain $|\hat{\beta}_k + \varepsilon| - |\hat{\beta}_k| = \text{sign}(\hat{\beta}_k)\varepsilon$. The Taylor's expansion implies that

$$\begin{aligned}
\log \frac{\theta + e^{\boldsymbol{X}_i^T(\hat{\boldsymbol{\beta}} + \varepsilon \mathbf{e}_k)}}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} &= \log(1 + \frac{1}{\theta} e^{\boldsymbol{X}_i^T(\hat{\boldsymbol{\beta}} + \varepsilon \mathbf{e}_k)}) - \log(1 + \frac{1}{\theta} e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}) \\
&= \frac{1}{1 + \frac{1}{\theta} e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} \cdot \frac{1}{\theta} e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}} (e^{x_{ik}\varepsilon} - 1) + o[\frac{1}{\theta} e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}} (e^{x_{ik}\varepsilon} - 1)] \\
&= \frac{1}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} \cdot e^{\boldsymbol{X}_i{}^T \hat{\boldsymbol{\beta}}} (x_{ik}\varepsilon + o(\varepsilon)) + o[\frac{1}{\theta} e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}} (x_{ik}\varepsilon + o(\varepsilon))] \\
&= \frac{e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}} x_{ik} \varepsilon}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} + o(\varepsilon).
\end{aligned}$$

Since the aim is to minimize the object function, we must have

$$\begin{aligned}
0 < F(\hat{\boldsymbol{\beta}} + \varepsilon \mathbf{e}_k) - F(\hat{\boldsymbol{\beta}}) &= \sum_{i=1}^{n} x_{ik} [\frac{(\theta + y_i) e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} - y_i] \varepsilon + \lambda_1 \text{sign}(\hat{\boldsymbol{\beta}}_k) \varepsilon + \lambda_2 (2\hat{\beta}_k \varepsilon + \varepsilon^2) \\
&= [\sum_{i=1}^{n} x_{ik} \frac{\theta(e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}} - y_i)}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} + \lambda_1 \text{sign}(\hat{\beta}_k) + 2\lambda_2 \hat{\beta}_k] \varepsilon + \lambda_2 \varepsilon^2 + o(\varepsilon)
\end{aligned}$$

Note that $\lambda_2 \neq 0$, for any sufficiently small $\varepsilon \in (-|\hat{\beta}_k|, |\hat{\beta}_k|)$, in order to make sure that the above inequality is valid, iff

$$\sum_{i=1}^{n} [x_{ik} \frac{\theta(e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}} - y_i)}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}}] + \lambda_1 \text{sign}(\hat{\beta}_k) + 2\lambda_2 \hat{\beta}_k = 0, (k = 1, 2, \cdots, p).$$

Thus we get $\left| \sum_{i=1}^{n} x_{ik} \frac{\theta(e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}} - y_i)}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} \right| = \lambda_1 + 2\lambda_2 |\hat{\beta}_i| > \lambda_1$.

Case 2. If $\hat{\beta}_i = 0$, for sufficiently small $\varepsilon \in \mathbb{R}$, by (S2.45) we have

$$F(\hat{\boldsymbol{\beta}} + \varepsilon \mathbf{e}_k) - F(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} [(\theta + y_i) \log \frac{\theta + e^{\boldsymbol{X}_i^T(\hat{\boldsymbol{\beta}} + \varepsilon \mathbf{e}_k)}}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}} - y_i x_{ik} \varepsilon] + \lambda_1(|\varepsilon|) + \lambda_2 \varepsilon^2.$$

According to the Taylor expansions of $F(\hat{\boldsymbol{\beta}}+\varepsilon\mathbf{e}_k)-F(\hat{\boldsymbol{\beta}})$ in Case 1, and observing $\left|\sum_{i=1}^{n}[x_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}]\right|\neq 0$. We must have

$$\begin{aligned}0<F(\hat{\boldsymbol{\beta}}+\varepsilon\mathbf{e}_k)-F(\hat{\boldsymbol{\beta}})&=\sum_{i=1}^{n}[x_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}]\varepsilon+\lambda_1(|\varepsilon|)+\lambda_2\varepsilon^2+o(\varepsilon)\\&=\{\sum_{i=1}^{n}[x_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}]+\lambda_1\text{sign}\varepsilon\}\varepsilon+\lambda_2\varepsilon^2+o(\varepsilon).\end{aligned}$$

Note that $\lambda_2\neq 0$, in order to make sure that the above inequality is valid for any sufficiently small $\varepsilon\in\mathbb{R}$, iff

$$\sum_{i=1}^{n}[x_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}]+\lambda_1\text{sign}\varepsilon=0,(k=1,2,\cdots,p).\tag{S2.46}$$

In other words, $\sum_{i=1}^{n}x_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}>-\lambda_1$ for $\varepsilon\geq 0$ and $\sum_{i=1}^{n}x_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}<\lambda_1$ for $\varepsilon\leq 0$. Thus we get $\left|\sum_{i=1}^{n}x_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}\right|\leq\lambda_1$.

S2.2 Proof of Lemma 8

The KKT conditions is crucial for us to derive the upper bound of grouping effect inequality associated with the difference between the coefficient paths of predictors X_i and X_j.

Case 1. When $\hat{\beta}_k\hat{\beta}_l>0$. According to Lemma 1, we have

$$\sum_{i=1}^{n}x_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}=\text{sign}(\hat{\beta}_k)(\lambda_1+2\lambda_2|\hat{\beta}_k|),\sum_{i=1}^{n}x_{il}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)}{\theta+e^{\boldsymbol{X}_i{}^T\hat{\boldsymbol{\beta}}}}=\text{sign}(\hat{\beta}_l)(\lambda_1+2\lambda_2|\hat{\beta}_l|)$$

Taking the subtraction of two equations above, we obtain

$$2\lambda_2\left|\hat{\beta}_k(\lambda_1,\lambda_2)-\hat{\beta}_l(\lambda_1,\lambda_2)\right|=\left|\sum_{i=1}^{n}(x_{ik}-x_{il})\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}\right|\leq\sum_{i=1}^{n}\frac{\theta\,|(x_{ik}-x_{il})|\cdot|e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)|}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}.$$

and therefore inequality (S1.33) is proved.

Case 2. When $\hat{\beta}_k\hat{\beta}_l<0$, i.e. $\text{sign}(\hat{\beta}_k)=-\text{sign}(\hat{\beta}_l)$. According to Lemma 1, we have

$$\left|\sum_{i=1}^{n}(x_{ik}-x_{il})\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}\right|=\left|2[\text{sign}(\hat{\beta}_k)\lambda_1+\lambda_2(\hat{\beta}_k-\hat{\beta}_l)]\right|$$

$$= \left|2\text{sign}(\hat{\beta}_k)[\lambda_1 + \lambda_2|\hat{\beta}_k - \hat{\beta}_l|]\right| \geq \left|2\lambda_2\text{sign}(\hat{\beta}_k)|\hat{\beta}_k - \hat{\beta}_l|\right|.$$

and therefore inequality (S1.33) is also proved.

Case 3. When $\hat{\beta}_k \neq 0, \hat{\beta}_l = 0$. By the Case 1 in Lemma 1 and (S2.46), by subtracting these two expressions we have

$$\sum_{i=1}^{n}(x_{ik} - x_{il})\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}} - y_i)}{\theta + e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}} = \lambda_1[\text{sign}\varepsilon + \text{sign}(\hat{\beta}_k)] + 2\lambda_2\text{sign}(\hat{\beta}_k)|\hat{\beta}_k|).$$

If $\text{sign}(\varepsilon + \text{sign}(\hat{\beta}_k) = 0$, it is apparently that (S1.33) is true. If $\text{sign}\varepsilon + \text{sign}(\hat{\beta}_k) = -2$ (or 2), it derives that

$$\sum_{i=1}^{n}(x_{ik} - x_{il})\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}} - y_i)}{\theta + e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}} = -2\lambda_1 - 2\lambda_2|\hat{\beta}_k|, \text{ (or } 2\lambda_1{+}2\lambda_2|\hat{\beta}_k|).$$

Then

$$\left|\sum_{i=1}^{n}(x_{ik} - x_{il})\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}} - y_i)}{\theta + e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}\right| = \left|2\lambda_1 + 2\lambda_2|\hat{\beta}_k|\right| \geq 2\lambda_2|\hat{\beta}_k| = 2\lambda_2|\hat{\beta}_k - \hat{\beta}_l|.$$

Thus (S1.33) is proved. If $\hat{\beta}_l \neq 0, \hat{\beta}_k = 0$, the proof is by the same method.

Case 4. When $\hat{\beta}_k = \hat{\beta}_l = 0$, (S1.33) is obviously.

S2.3 Proof of Lemma 9

The variance and kurtosis of Y_i are

$$\text{Var}Y_i = \frac{\theta p_i}{(1-p_i)^2}, \quad \text{Kurt}(Y_i) := \frac{\text{E}|Y_i - \text{E}Y_i|^4}{(\text{E}|Y_i - \text{E}Y_i|^2)^2} = 3 + \frac{6}{\theta} + \frac{(1-p_i)^2}{\theta p_i},$$

see p216 of Johnson et al. (2005). By (C.1) and (C.3), we get

$$0 < \frac{e^{-LB}}{\theta + e^{-LB}} \leq p_i = \frac{e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}}{\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}} \leq \frac{e^{LB}}{\theta + e^{LB}} < 1.$$

Let $Q_i := \frac{p_i}{(1-p_i)^2} \in [\frac{e^{-LB}(\theta+e^{-LB})}{\theta^2}, \frac{e^{LB}(\theta+e^{LB})}{\theta^2}]$, then

$$\text{E}S_n = \frac{1}{n}\sum_{i=1}^{n}\text{E}|Y_i - \text{E}Y_i|^2 = \frac{1}{n}\sum_{i=1}^{n}\theta Q_i \leq \frac{e^{LB}(\theta + e^{LB})}{\theta} := \mu.$$

For (2), we obtain

$$\begin{aligned}\mathrm{Var}|Y_i-\mathrm{E}Y_i|^2 &= \mathrm{E}|Y_i-\mathrm{E}Y_i|^4-(\mathrm{E}|Y_i-\mathrm{E}Y_i|^2)^2=(\mathrm{Var}Y_i)^2[Kurt(Y_i)-1]\\ &=\frac{\theta^2p_i^2}{(1-p_i)^4}\left(2+\frac{6}{\theta}+\frac{(1-p_i)^2}{\theta p_i}\right)=(2\theta^2+6\theta)Q_i^2+\theta Q_i.\end{aligned}$$

So, it implies

$$\mathrm{Var}|Y_i-\mathrm{E}Y_i|^2\le(2+\frac{6}{\theta})e^{2LB}(\theta+e^{LB})^2+\frac{e^{LB}(\theta+e^{LB})}{\theta}:=\sigma^2.$$

S2.4 Proof of Proposition 1

Proof. Let $V_j=\sum_{i=1}^{n}f_j(X_i)$, then by Jensen's inequality and Hoeffding's lemma, we have

$$\begin{aligned}\mathrm{E}\max_{1\le j\le p}|V_j| &= \frac{1}{\lambda}\mathrm{E}\log \mathrm{e}^{\lambda\max\limits_{1\le j\le p}|V_j|}\le\frac{1}{\lambda}\log\mathrm{E}\mathrm{e}^{\lambda\max\limits_{1\le j\le p}|V_j|}\\ &\le\frac{1}{\lambda}\log\sum_{i=1}^{n}\mathrm{E}\mathrm{e}^{\lambda|V_j|}\le\frac{1}{\lambda}\log[\sum_{j=1}^{p}2e^{\frac{1}{2}\lambda^2\sum\limits_{i=1}^{n}a_{ij}^2}]\\ &\le\frac{1}{\lambda}\log\left[2pe^{\frac{1}{2}\lambda^2\max\limits_{1\le j\le p}\sum\limits_{i=1}^{n}a_{ij}^2}\right]=\frac{1}{\lambda}\log(2p)+\frac{1}{2}\lambda\max_{1\le j\le p}\sum_{i=1}^{n}a_{ij}^2.\end{aligned}$$

Then $\mathrm{E}\max\limits_{1\le j\le p}|V_j|\le\inf\limits_{\lambda>0}\{\frac{1}{\lambda}\log(2p)+\frac{1}{2}\lambda\max\limits_{1\le j\le p}\sum\limits_{i=1}^{n}a_{ij}^2\}=\sqrt{2\log(2p)}\cdot\max\limits_{1\le j\le p}\sum\limits_{i=1}^{n}a_{ij}^2$. □

S3 The the proof of (S1.19) and the value γ

S3.1 The the proof of (S1.19)

In this section, we illustrate the use of concentration inequalities in application to empirical processes. Here we use the convex geometry method to derive various tail bounds on the suprema of empirical processes, i.e. for random variables that are generated by taking suprema of sample averages over function classes. The following discrete version of Prékopa–Leindler inequality is extracted from Theorem 1.2 in Halikias et al. (2019), it is essential the discrete variants of Brunn-Minkowski type inequalities in convex geometry, see Halikias et al. (2019). In fact, the discrete variants of

Prékopa–Leindler inequality is of paramount importance to derive concentration inequalities for strongly log-concave counting measures, similar to continuous Prékopa–Leindler inequality presented in Theorem 3.15 of Wainwright (2019).

Lemma 13 (discrete Prékopa–Leindler inequality)**.** *Let* $\lambda \in [0,1]$ *and suppose* $f, g, h, k : \mathbb{Z}^n \to [0, \infty)$ *satisfy*

$$f(\boldsymbol{x})g(\boldsymbol{y}) \leq h(\lfloor \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \rfloor) k(\lceil (1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y} \rceil) \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n \quad \text{(S3.47)}$$

where $\lfloor \boldsymbol{x} \rfloor = (\lfloor x_1 \rfloor, \ldots \lfloor x_n \rfloor)$ *and* $\lceil \boldsymbol{x} \rceil = (\lceil x_1 \rceil, \ldots, \lceil x_n \rceil)$*. Then*

$$(\sum_{\boldsymbol{x} \in Z^n} f(\boldsymbol{x}))(\sum_{\boldsymbol{x} \in Z^n} g(\boldsymbol{x})) \leq (\sum_{\boldsymbol{x} \in \mathbb{Z}^n} h(\boldsymbol{x}))(\sum_{\boldsymbol{x} \in Z^n} k(\boldsymbol{x})),$$

where $\lfloor r \rfloor = \max\{m \in \mathbb{Z}; m \leq r\}$ *is the lower integer part of* $r \in \mathbb{R}$ *and* $\lceil r \rceil = -\lfloor -r \rfloor$ *the upper integer part.*

From a geometric point of view, the Prékopa-Leindler inequality is useful tool to establish some advanced concentration inequalities of Lipschitz functions for strongly log-concave distributions. Motivated by Moriguchi et al. (2020), we define a distribution P_γ with a density $p(\boldsymbol{x})$ (w.r.t. the counting measure) is said to be strongly discrete log-concave if the log function $\psi(\boldsymbol{x}) =: -\log p(\boldsymbol{x}) : \mathbb{Z}^n \to \mathbb{R}$ is *strongly midpoint log-convex* for some $\gamma > 0$

$$\psi(\boldsymbol{x}) + \psi(\boldsymbol{y}) - \psi(\lceil \frac{1}{2}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y} \rceil) - \psi(\lfloor \frac{1}{2}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y} \rfloor) \geq \frac{\gamma}{4}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n. \quad \text{(S3.48)}$$

Let $\lambda \in [0,1]$. The (S3.48) is a slightly extension *strongly convex with modulus of convexity* γ for continuous functions on $\mathbb{R}^n$

$$\lambda\psi(\boldsymbol{x}) + (1-\lambda)\psi(\boldsymbol{y}) - \psi(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \geq \frac{\gamma}{2}\lambda(1-\lambda)\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n,$$

see Chapter 2 of Mahoney et al. (2018).

Strongly discrete log-convex property requires the restricted behavior of continuous functions on lattice space. If $\gamma = 0$, (S3.48) will leads to the definition of is *discrete midpoint convexity* for $\psi(\boldsymbol{x})$ mentioned by Moriguchi et al. (2020)

$$\psi(\boldsymbol{x}) + \psi(\boldsymbol{y}) \geq \psi(\lceil \frac{1}{2}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y} \rceil) + \psi(\lfloor \frac{1}{2}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y} \rfloor) \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n.$$

Howsoever, directly restrict some continuous function to a lattice space does not necessarily yield a discretely convex function, the counter-example in Yüceer (2002).

For P_γ being one-dimensional, we say that the probability mass function $p(x)$ are log-concave if the sequence $\{p(x)\}_{x\in\mathbb{Z}}$ is a log-concave sequence which means that for any $m, n \in \mathbb{Z}$ and $\lambda \in (0,1)$ such that $\lambda n+(1-\lambda)m \in \mathbb{Z}$, we have

$$p(\lambda n + (1-\lambda)m) \geq p(n)^\lambda p(m)^{1-\lambda}.$$

Equivalently, $p(n)^2 \geq p(n-1)p(n+1)$ for every $x \in \mathbb{Z}$ (or x in a subset of $\mathbb{Z}$), see Klartag and Lehec (2019).

Theorem 2 (Concentration for strongly log-concave discrete distributions)**.** *Let P_γ be any strongly log-concave discrete distribution index by $\gamma > 0$ on $\mathbb{Z}^n$. Then for any function $f : \mathbb{R}^n \to \mathbb{R}$ that is L-Lipschitz with respect to Euclidean norm, we have for $\boldsymbol{X} \sim P_\gamma$*

$$P_\gamma\{|f(\boldsymbol{X}) - \mathrm{E}f(\boldsymbol{X})| \geq t\} \leq 2e^{-\frac{\gamma t^2}{4L^2}},\ t > 0. \tag{S3.49}$$

The Theorem 2 allows for some dependence due to a function of vector $\boldsymbol{X}$ will be a dependence summation.

Proof. Let h be an arbitrary zero-mean function with Lipschitz constant L with respect to the Euclidean norm. It suffices to show that $\mathrm{E}e^{h(x)} \leq e^{\frac{L^2}{\tau}}$. Indeed, if this inequality holds, then, given an arbitrary function f with Lipschitz constant K and $\lambda \in \mathbb{R}$, we can apply this inequality to the zero-mean function $h(\boldsymbol{X}) := \lambda(f(\boldsymbol{X}) - \mathrm{E}f(\boldsymbol{X}))$, which has Lipschitz constant $L = \lambda K$. The zero-mean function h is L-Lipschitz and for given $\lambda \in (0,1)$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n$, define the proximity operator of h

$$l(\boldsymbol{y}) := \inf_{\boldsymbol{x}\in\mathbb{Z}^n} \left\{ h(\boldsymbol{x}) + \frac{\gamma}{4}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right\}$$

as the functional minimizer of the rescaled h with Euclidean norm.

Next, with this functional minimizer, the proof is based on adopting the discrete Prekopa-Leindler inequality Lemma 13 with $\lambda = 1/2$ and $h(\boldsymbol{t}) = k(\boldsymbol{t}) =: p(\boldsymbol{t}) = e^{-\psi(\boldsymbol{t})}$ and the pair of functions given by $f(\boldsymbol{x}) := e^{-h(\boldsymbol{x})-\psi(\boldsymbol{x})}$ and $g(\boldsymbol{y}) := e^{l(\boldsymbol{y})-\psi(\boldsymbol{y})}$.

It is sufficient to check the (S3.50) in Lemma 13 is satisfied with $\lambda = 1/2$, i.e.

$$e^{\frac{1}{2}[l(\boldsymbol{y})-h(\boldsymbol{x})-\psi(\boldsymbol{y})-\psi(\boldsymbol{x})]} \leq e^{-\frac{1}{2}\psi(\lceil\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rceil)} \cdot e^{-\frac{1}{2}\psi(\lfloor\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rfloor)} \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n \tag{S3.50}$$

Indeed, by discrete strong convexity of the function ψ and the proximity operator of h

$$\frac{1}{2}[\psi(\boldsymbol{x}) + \psi(\boldsymbol{y}) - \psi(\lceil\frac{1}{2}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y}\rceil) - \psi(\lfloor\frac{1}{2}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y}\rfloor) \geq \frac{\gamma}{8}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2,$$

we have

$$\begin{aligned}
&-\frac{1}{2}\psi(\lceil\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rceil)-\frac{1}{2}\psi(\lfloor\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rfloor)\\
&\geq\frac{1}{2}\left\{l(\boldsymbol{y})-h(\boldsymbol{x})-\frac{\gamma}{4}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right\}-\frac{1}{2}\psi(\lceil\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rceil)-\frac{1}{2}\psi(\lfloor\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rfloor)\\
&\geq\frac{1}{2}\{l(\boldsymbol{y})-h(\boldsymbol{x})\}-\frac{1}{2}\psi(\boldsymbol{y})-\frac{1}{2}\psi(\boldsymbol{x}).
\end{aligned}$$

which verifies (S3.50).

Note that $\sum_{\boldsymbol{x}\in\mathbb{Z}^n}h(\boldsymbol{x})=\sum_{\boldsymbol{x}\in\mathbb{Z}^n}k(\boldsymbol{x})=1$, the Lemma 13 implies that

$$\mathrm{E}e^{l(\boldsymbol{Y})}\mathrm{E}e^{-h(\boldsymbol{X})}=\sum_{\boldsymbol{x}\in\mathbb{Z}^n}e^{-h(\boldsymbol{x})-\psi(\boldsymbol{x})}\sum_{\boldsymbol{y}\in\mathbb{Z}^n}e^{l(\boldsymbol{y})-\psi(\boldsymbol{y})}\leq 1$$

Rearranging and Jensen's inequality yield

$$\mathrm{E}e^{l(\boldsymbol{Y})}\leq(\mathrm{E}e^{-h(\boldsymbol{X})})^{-1}\leq(e^{\mathrm{E}[-h(\boldsymbol{X})]})^{-1}=1$$

where the last equality due to $\mathrm{E}[-h(\boldsymbol{X})]=\mathrm{E}[\lambda(f(\boldsymbol{X})-\mathrm{E}f(\boldsymbol{X}))]=0$.

So we have by definition of the proximity operator

$$\begin{aligned}
1\geq\mathrm{E}e^{l(\boldsymbol{y})}=\mathrm{E}e^{\inf_{\boldsymbol{x}\in\mathbb{Z}^n}\left\{h(\boldsymbol{x})+\frac{\gamma}{4}\|\boldsymbol{x}-\boldsymbol{Y}\|_2^2\right\}}&=\mathrm{E}e^{\inf_{\boldsymbol{x}\in\mathbb{Z}^n}\left\{h(\boldsymbol{Y})+[h(\boldsymbol{x})-h(\boldsymbol{Y})]+\frac{\gamma}{4}\|\boldsymbol{x}-\boldsymbol{Y}\|_2^2\right\}}\\
&\geq\mathrm{E}e^{h(\boldsymbol{Y})+\inf_{\boldsymbol{x}\in\mathbb{R}^n}\left\{-L\|\boldsymbol{x}-\boldsymbol{Y}\|_2+\frac{\gamma}{4}\|\boldsymbol{x}-\boldsymbol{Y}\|_2^2\right\}}\\
&=\mathrm{E}e^{h(\boldsymbol{Y})-L^2/\gamma}.
\end{aligned}$$

where the second last inequality is from the fact that h is L-Lipschitz: $|h(\boldsymbol{x})-h(\boldsymbol{Y})|\leq L\|\boldsymbol{x}-\boldsymbol{Y}\|_2$.

It yields that

$$\mathrm{E}e^{\lambda(f(\boldsymbol{X})-\mathrm{E}f(\boldsymbol{X})]}\leq e^{\frac{1}{2}\cdot\lambda^2\cdot\frac{2L^2}{\gamma}}\quad\text{for all }\lambda\in\mathbb{R},$$

This implies that $f(X)-\mathrm{E}f(\boldsymbol{X})$ has a sub-Gaussian tail bound as claimed in (S3.49). □

S3.2 The value γ

For $Y_i\sim\mathrm{NBD}(\mu_i,\theta)$ with known $\theta>1$. The log-density for $\boldsymbol{y}=(y_1,\cdots,y_n)^T$ is

$$\log p(\boldsymbol{y})=:\sum_{i=1}^n\log p_i(y_i)=:\sum_{i=1}^n\psi(y_i)$$

$$= \sum_{i=1}^{n} \{\log\Gamma(\theta+y_i) + y_i \log\mu_i + \theta\log\theta - \log\Gamma(\theta) - \log y_i! - (\theta+y_i)\log(\theta+\mu_i)\}.$$

Then

$$\psi'(y_i) := \left.\frac{\partial \log p(y)}{\partial y}\right|_{y_i} = \log\frac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)} - y_i\log(\theta+\mu_i).$$

Let us find the γ. Taylor's expansion implies

$$\begin{aligned}\psi(y) &= \psi(\lceil \tfrac{1}{2}x + \tfrac{1}{2}y\rceil) + \tfrac{1}{2}\psi'(\lceil \tfrac{1}{2}x + \tfrac{1}{2}y\rceil)(y-x) + \tfrac{1}{8}(y-x)^2\psi''(a_1)\\ \psi(x) &= \psi(\lfloor \tfrac{1}{2}x + \tfrac{1}{2}y\rfloor) + \tfrac{1}{2}\psi'(\lfloor \tfrac{1}{2}x + \tfrac{1}{2}y\rfloor)(x-y) + \tfrac{1}{8}(y-x)^2\psi''(a_2)\end{aligned}$$

where $a_1 = t_1 y + (1-t_1)(x+y)/2, a_2 = t_2 y + (1-t_1)(x+y)/2$ with $t_1, t_2 \in [0,1]$.

So we have

$$\begin{aligned}\tfrac{1}{2}\psi(x) + \tfrac{1}{2}\psi(y) &= \tfrac{1}{2}\psi(\lfloor \tfrac{1}{2}x + \tfrac{1}{2}y\rfloor) + \psi(\lceil \tfrac{1}{2}x + \tfrac{1}{2}y\rceil)\\ &\quad + \tfrac{x-y}{4}\left[\psi'(\lfloor \tfrac{1}{2}x + \tfrac{1}{2}y\rfloor) - \psi'(\lceil \tfrac{1}{2}x + \tfrac{1}{2}y\rceil)\right] + \tfrac{\psi''(a_1)+\psi''(a_2)}{16}(y-x)^2\end{aligned}$$

Define

$$\Delta(x,y) := \frac{x-y}{4}\left[\psi'(\left\lfloor\frac{1}{2}x+\frac{1}{2}y\right\rfloor) - \psi'(\left\lceil\frac{1}{2}x+\frac{1}{2}y\right\rceil)\right] + \frac{\psi''(a_1)+\psi''(a_2)}{16}(y-x)^2$$

We have

$$\Delta(x,y) \geq |x-y|^2\left\{\frac{\psi''(a_1)+\psi''(a_2)}{16} - \sup_{x\neq y; x,y\in\mathbb{Z}^n}\frac{|[\psi'(\lfloor (x+y)/2\rfloor) - \psi'(\lceil (x+y)/2\rceil)]|}{4|x-y|}\right\}$$

Let

$$\begin{aligned}C_\psi &:= \sup_{x\neq y; x,y\in\mathbb{Z}^n}\tfrac{|[\psi'(\lfloor (x+y)/2\rfloor) - \psi'(\lceil (x+y)/2\rceil)]|}{4|x-y|}\\ &= \sup_{x\neq y; x,y\in\mathbb{Z}^n}\left|\left[\log\tfrac{\Gamma(\theta+\lfloor (x+y)/2\rfloor)\Gamma(\lceil (x+y)/2\rceil+1)}{\Gamma(\theta+\lceil (x+y)/2\rceil)\Gamma(\lfloor (x+y)/2\rfloor+1)} - \tfrac{(\lfloor (x+y)/2\rfloor - \lceil (x+y)/2\rceil)}{\log^{-1}(\theta+\mu_i)}\right]\right|/4|x-y|\end{aligned}$$

We can see that $C_\psi \approx \frac{|[\log(\theta+\mu_i)]|}{4}$ or 0.

Note that

$$\begin{aligned}\psi''(y) &:= \left.\tfrac{\partial^2 \log p(y)}{\partial y^2}\right|_{y=y_i} = \tfrac{d}{dy_i}\log\tfrac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)} = \sum_{k=1}^{\infty}(\tfrac{1}{k+1} - \tfrac{1}{k+\theta+y_i}) - \sum_{k=1}^{\infty}(\tfrac{1}{k+1} - \tfrac{1}{k+y_i+1})\\ &= \sum_{k=1}^{\infty}(\tfrac{1}{k+y_i+1} - \tfrac{1}{k+\theta+y_i}) \geq \inf_{y_i\in Z}\sum_{k=1}^{\infty}(\tfrac{1}{k+y_i+1} - \tfrac{1}{k+\theta+y_i}) = C_{\psi''}.\end{aligned}$$

Now, we get

$$\Delta(x,y) \geq |x-y|^2\left\{\frac{\psi''(a_1)+\psi''(a_2)}{16} - C_\psi\right\} \geq |x-y|^2\left(\frac{C_{\psi''}}{8} - C_\psi\right)$$

which gives $\gamma =: \frac{C_{\psi''}}{8} - C_\psi > 0$ from (H.4), if $C_\psi \approx \frac{|[\log(\theta+\mu_i)]|}{4}$ is small.

S4 Simulation Studies

In practice, the nuisance parameter θ is often unknown. We need a proper estimation for θ in the NB regression, although it is a nuisance parameter. Many dispersion estimators and their algorithms for non-penalized NBR are available; see section 8.4.2 of Hilbe (2011), Robinson and Smyth (2007) and references therein. Here we prefer to use two subproblem iteratively algorithms, which is applied by Wang et al. (2016). Firstly, we fit an NB regression by MLE with dispersion parameter θ and mean μ_i without considering covariates information. Secondly, we optimize the penalized log-likelihood (1) and estimate $\boldsymbol{\beta}$ with the θ being estimated in the first step. Thirdly, and estimating θ with the current estimates fixed (1). Repeated iteration when the desired stopping criteria are attained.

Well-chosen tuning parameters is also crucial in the NBR optimization problem. The BIC criterion (an adjusted AIC criterion) is employed to determine tuning parameters by the principal proposed by Zou et al. (2007). The negative likelihood with ridge terms is considered as our modified likelihood, thus the BIC criterion for elastic-net regularized NBR is defined as

$$\mathrm{BIC}_{\hat{\beta}(\lambda_1,\lambda_2)} := \frac{-1}{n}\sum_{i=1}^{n}[Y_i\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} - (\theta+Y_i)\log(\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}})] - \lambda_2\|\hat{\boldsymbol{\beta}}\|_2^2 + \frac{\log n}{n}\hat{\mathrm{df}}(en) \tag{S4.51}$$

where $\hat{\mathrm{df}}(en) := ||\hat{\beta}(\lambda_1,\lambda_2)||_0$ is the number of estimated nonzero coefficients. We use the BIC to find nearly optimal tuning parameters and then further tune the λ_1 such that the support recovery rate is high, and not all coefficients are penalized to zero.

A simulated comparison by elastic-net and Lasso estimator for NBR is performed by using `R`, and we also give the confidence intervals for both de-biased Lasso and de-biased elastic-net estimator. The package `mpath` is employed to estimate the solution path based on a sequence of turning parameters. The function `rnegbin()` is used to generate negative binomial r.v. with mean μ_i and variance $\mu_i + \frac{\mu_i^2}{\theta}$ in the package `MASS`, and its also includes the estimation of the the dispersion parameter θ by the function `fitdistr()`.

In confidence intervals based on de-biased estimators, the package `fastclime` is adopted for computing a high-dimensional precision matrix (i.e., the inverse Hessian matrix of NBR). It contains an efficient and fast algorithm for solving a family of regularized linear programming problems, see Pang et al. (2014).

In Table 1 and 2, we simulate responses via the model $Y_i \sim \mathrm{NB}(e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}, \theta)$ with $\theta = 5$ and true regression vector

$$\beta^* = (\underbrace{10|N(0,1)| + 0.2, \cdots, 10|N(0,1)| + 0.2}_{10}, \underbrace{0, \cdots, 0}_{p-10}).$$

Thus $H = \{1, 2, \cdots, 10\}$ and $d^* = 10$. The $\{X_{ij}\}$ are i.i.d. simulated from $N(0,1)$ and then do standardization (3.25) which renders that $\{X_{ij}\}$ are approximately bounded.

In Table 1, let $\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_n := \frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_1$ and $\|\delta\|_H := \sum_{i\in H}|\delta_i|$. The de-biased estimator for elastic-net (or Lasso) is $\hat{\boldsymbol{b}} = \hat{\boldsymbol{\beta}} - \hat{\Theta}\dot{\ell}(\hat{\boldsymbol{\beta}})$. The true coefficient is simulated as

$$\beta^* = (0.242, 0.648, 0.676, 0.313, 0.602, 0.236, 0.851, 0.796, 0.531, 0.404, \cdots)^T,$$

with $\|\beta^*\|_1 = 5.300$.

Thus by our assumption $\lambda_2 \le \frac{\lambda_1}{8B}$ in Theorem 3, we put $\lambda_2 \approx 0.02\lambda_1$. By referring the BIC criterion (S4.51) and the oracle inequality (2.22) in Theorem 3, we set λ_1,or $\lambda \approx 10\sqrt{\frac{\log p}{n}}$ in elastic-net or Lasso.

Table 1 shows that the proposed elastic-net estimators for NBR are more accurate than the Lasso estimators. The ridge penalty's help reflects that elastic-net can improve the estimation accuracy in aspects of estimation and prediction errors due to the bias-variance tradeoff. We can also see that the increasing p will hinder the estimated accuracy by thinking about the curse of dimensionality. We should note that penalized estimations always have a bias, and de-biased procedures correct the bias. The de-biased estimators have fewer ℓ_1-estimation errors in the support H, and de-biased elastic-net outperforms the de-biased Lasso.

Table 1 The ℓ_1 prediction error and support recovery for elastic-net (Lasso) and its debiased version in NBR, $n = 500$.

	elastic-net					
p	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$ ($\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_H$)	$P(H = \hat{H})$	$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_n$	$\|\hat{\boldsymbol{b}} - \boldsymbol{\beta}^*\|_H$	λ_1	$\hat{\theta}$
400	1.491 (1.376)	1.000	0.222	0.723	0.12	2.927
600	1.749 (1.405)	1.000	0.326	0.731	0.13	2.350
700	1.767 (1.709)	1.000	0.340	0.955	0.14	2.952
	Lasso				λ	
400	1.505 (1.405)	1.000	0.230	0.730	0.12	2.836
600	1.779 (1.719)	1.000	0.341	0.896	0.13	2.262
700	1.784 (1.739)	1.000	0.351	0.966	0.14	2.862

Table 2 Confidence intervals for the de-biased estimates with 95% confidence level, $n = 500, p = 700$.

		elastic-net ($\lambda_1 = 0.11, \lambda_2 = 0.02\lambda_1$)			Lasso ($\lambda = 0.11$)		
j	β_j^*	$\hat{\beta}_j$	$\hat{b}_j$	$[\hat{b}_j^L, \hat{b}_j^U]$	$\hat{\beta}_j$	$\hat{b}_j$	$[\hat{b}_j^L, \hat{b}_j^U]$
1	0.828	0.753	0.810	[0.677,0.944]	0.758	0.783	[0.377,1.190]
2	1.218	1.059	1.119	[0.986,1.252]	1.077	1.103	[0.726,1.481]
3	0.321	0.098	0.122	[-0.010,0.253]	0.107	0.109	[-0.209,0.428]
4	0.991	0.829	0.891	[0.769,1.013]	0.839	0.860	[0.602,1.118]
5	1.052	0.934	1.000	[0.872,1.129]	0.947	0.972	[0.622,1.322]
6	0.268	0.231	0.265	[0.145,0.385]	0.235	0.246	[-0.023,0.516]
7	0.510	0.351	0.384	[0.260,0.509]	0.374	0.384	[0.075,0.693]
8	0.838	0.728	0.773	[0.641,0.905]	0.755	0.772	[0.421,1.124]
9	1.183	0.988	1.048	[0.925,1.172]	0.974	0.998	[0.661,1.336]
10	0.382	0.276	0.314	[0.193,0.435]	0.295	0.303	[0.018,0.588]
covering number		7			10		

Table 1 shows that the proposed elastic-net estimators for NBR are more accurate than the Lasso estimators. The ridge penalty's help reflects that elastic-net can improve the estimation accuracy in aspects of estimation and prediction errors due to the bias-variance tradeoff. We can also see that the increasing p will hinder the estimated accuracy by thinking about the curse of dimensionality. We should note that penalized estimations always have a bias, and de-biased procedures correct the bias. The de-biased estimators have fewer ℓ_1-estimation errors in the support H, and de-biased elastic-net outperforms the de-biased Lasso.

Table 3 Simulation for grouping effect.

$\hat{\boldsymbol{\beta}}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$
elastic-net	2.025	0.421	0.422	1.00	0	0	0	0	0	0
Lasso	1.838	0	0	1.469	0	0	0	0	0	0
Ridge	2.059	0.861	0.861	0.664	-0.199	-0.035	-0.164	0.027	-0.006	0.170
MLE	2.620	2.783	NA	NA	-0.142	0.083	-0.092	0.076	-0.063	0.180
$\boldsymbol{\beta}^*$	2	0.5	0.5	1	0	0	0	0	0	0

A numerically demonstration of the grouping phenomenon (see Theorem 4) is given in Table 3. The covariates are correlated simulated as: $X_1 \sim U[0,1], X_2 \sim U[0,1], X_3 = X_2, X_4 = 0.7X_3 + X_2 + 0.3X_1$. The true coefficient vector is $\beta^* = (2, 0.5, 0.5, 1, \underbrace{0, \cdots, 0}_{6})^T$. We consider the elastic-net ($\lambda_1 = 0.3, \lambda_2 = 0.3\lambda_1$), Lasso ($\lambda = 0.3$), Ridge ($\lambda = 0.3$), MLE. The results show that the elastic-net successfully select both X_2 and X_3 together into the model and the MLE the estimated coefficients fit better than other methods. Except X_5 to X_{10}, the Lasso shrinkages the coefficients of X_2, X_3 to zero, and MLE performs worst due to the correlated covariates X_2, X_3, X_4. The results indicate that the the elastic-net can select the strongly related variables X_2, X_3 into the model, reflecting the grouping effect.

References

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via l1 and l1+ l2 penalization. Electronic Journal of Statistics, 2, 1153-1194.

Bühlmann, P., van de Geer, S. A. (2011). Statistics for high-dimensional data: methods, theory and applications. Springer.

Blazere, M., Loubes, J. M., Gamboa, F. (2014). Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension. IEEE Transactions on Information Theory, 60(4), 2303-2318.

Giné, E., Nickl, R. (2015). Mathematical foundations of infinite-dimensional statistical models. Cambridge University Press.

Halikias, D., Klartag, B. A., & Slomka, B. A. (2019). Discrete variants of Brunn-Minkowski type inequalities. arXiv preprint arXiv:1911.04392.

Hilbe, J. M. (2011). Negative binomial regression, 2ed. Cambridge University Press.

Johnson, N. L., Kemp, A. W., Kotz S. (2005). Univariate Discrete Distributions, 3ed. Wiley.

Klartag, B. A., & Lehec, J. (2019). Poisson processes and a log-concave Bernstein theorem. Studia Mathematica, 247, 85-107.

Moriguchi, S., Murota, K., Tamura, A., & Tardella, F. (2020). Discrete midpoint convexity. Mathematics of Operations Research, 45(1), 99-128.

Wegkamp, M. (2007). Lasso type classifiers with a reject option. Electronic Journal of Statistics, 1, 155-168.

Pang, H., Liu, H., & Vanderbei, R. (2014). The fastclime package for linear programming and large-scale precision matrix estimation in R. The Journal of Machine Learning Research, 15(1), 489-493.

Robinson, M. D., Smyth, G. K. (2007). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics, 9(2), 321-332.

Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint (Vol. 48). Cambridge University Press.

Wang, Z., Ma, S., Zappitelli, M., Parikh, C., Wang, C. Y., & Devarajan, P. (2016). Penalized count data regression with application to hospital stay after pediatric cardiac surgery. Statistical methods in medical research, 25(6), 2685-2703.

Mahoney, M. W., Duchi, J. C., & Gilbert, A. C. (Eds.). (2018). The Mathematics of Data (Vol. 25). American Mathematical Society.

Maurer, A., & Pontil, M. (2021). Some Hoeffding-and Bernstein-type Concentration Inequalities. arXiv preprint arXiv:2102.06304.

van der Vaart, A. W., & Wellner, J. A. (1996). Weak convergence and empirical processes: with applications to statistics, Springer.

Rigollet, P. (2012). Kullback-Leibler aggregation and misspecified generalized linear models.

The Annals of Statistics, 40(2), 639-665.

Yüceer, Ü. (2002). Discrete convexity: convexity for functions defined on discrete spaces. Discrete Applied Mathematics, 119(3), 297-304.

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the degrees of freedom of the lasso. The Annals of Statistics, 35(5), 2173-2192.