Statistica Sinica Preprint No: SS-2019-0315	
Title	Elastic-net Regularized High-dimensional Negative Binomial Regression: Consistency and Weak Signal Detection
Manuscript ID	SS-2019-0315
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0315
Complete List of Authors	Huiming Zhang and Jinzhu Jia
Corresponding Author	Jinzhu Jia
E-mail	jzjia@math.pku.edu.cn
Notice: Accepted version subject to English editing.	

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

Huiming Zhang[1,3], Jinzhu Jia[2,3]

School of Mathematical Sciences[1]*, School of Public Health*[2] *and*

Center for Statistical Sciences[3]*,Peking University, Beijing, 100871, China*

Abstract: We study sparse negative binomial regression (NBR) for count data by showing non-asymptotic merits of the Elastic-net estimator. Two types of oracle inequalities are derived for the Elastic-net estimates of NBR by utilizing Compatibility Factor or Stabil Condition. The second-type oracle inequality is for random design which can be extended to many $\ell_1 + \ell_2$ regularized M-estimation with the corresponding empirical process having stochastic Lipschitz properties. To show some high probability events, we derive concentration inequality for suprema empirical processes for the weighted sum of negative binomial variables. For applications, we show the sign consistency provided that the non-zero components in sparse true vector are larger than a proper choice of the weakest signal detection threshold; and the second application is that we show the grouping effect inequality with high probability; thirdly, under some assumptions of design matrix, we can recover the true variable set with high probability if the weakest signal detection threshold is large than the turning parameter up to a known constant; at last, we briefly discuss the de-biased Elastic-net estimator and numerical studies are given to support the proposal.

Key words: high-dimensional count data regression, oracle inequalities, stochastic Lipschitz condition, mpirical processes, sign consistency, de-biased Elastic-net.

1 Introduction

In this paper, we focus on regression problems of count data (sometimes called categorical data). The responses are denoted as $\{Y_i\}_{i=1}^n$ each of which follows a discrete distribution. The expectation of Y_i will be related to $\boldsymbol{X}_i^T\boldsymbol{\beta}$ after a transformation by a link function. Poisson regression is a well-known example. Here the covariates $\boldsymbol{X}_i := (x_{i1}, \cdots, x_{ip})^T, (i = 1, 2, \cdots, n)$ are supposed to be a deterministic or random variable; if it is random we could deal with the model by conditioning on design matrix $\mathbf{X} := (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^T$. Covariates in count data regression may take discrete or continuous values, and important exemplified regressions includes: logistic regression, Poisson regression, negative binomial regression, etc. There are many monographs on statistical models for counting data, see Hilbe (2011); Tutz (2011).

A commonly used regression model for count data is the Poisson generalized linear model, which is of frequent occurrence in economic, social, and biological science, see Tutz (2011). Poisson regression considers that the response variables are nonnegative integers and follow the Poisson distribution. The sample responses Y_i's obey the Poisson distributions $P(Y_i = y_i|\lambda_i) = \frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i}, (i = 1, 2, \cdots, n)$, where the expectation of Y_i is $\lambda_i := \mathrm{E}(Y_i)$. We require that the positive parameter λ_i is related to a linear combination of p covariate variables. And the assumption of Poisson regression considers the logarithmic link function $\eta(\lambda_i) =: \log\lambda_i = \boldsymbol{X}_i^T\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$. According to the nature of Poisson distribution, the variance is equal to the expectation: $\mathrm{E}(Y_i\,|\,\boldsymbol{X}_i) = \mathrm{Var}(Y_i\,|\,\boldsymbol{X}_i) = \lambda_i$.

However, count data in practice often encounter the situation where the variance

is greater than the mean comparing to Poisson count data, in technical terms, called overdispersion. For example in RNA-Seq gene expression data, the negative binomial (NB) distribution provides a good choice for modelling a count variable and related high-dimensional sets of quantitative or binary variables are of interest, i.e. $p \gg n$. It is often shown the evidence of over-dispersion that the variance of the response variable is greater than its mean, see Rauschenberger et al. (2016), Qiu et al. (2018). To test whether the variance of a count data is greater than the expectation, a commonly used testing method is firstly proposed by Cameron and Trivedi (1990). It is called the Cameron-Trivedi test:

$$\mathrm{H}_0\colon \mathrm{Var}(Y_i\,|\boldsymbol{X}_i) = \mathrm{E}(Y_i\,|\boldsymbol{X}_i) =: \mu_i \quad \text{v.s.} \quad \mathrm{H}_1\colon \mathrm{Var}(Y_i\,|\boldsymbol{X}_i) = \mu_i + \alpha g(\mu_i),$$

where $g(\mu_i) = \mu_i$ or $g(\mu_i) = \mu_i^2$ and the constant α is the value to be tested. Therefore, the hypothesis test is alternatively written as $\mathrm{H}_0\colon\ \alpha = 0$ v.s. $\mathrm{H}_1\colon\ \alpha \neq 0$. For $\alpha \neq 0$, the count data is called over-dispersed if $\alpha > 0$ and it is called under-dispersed if $\alpha < 0$. Here the under-dispersion means that the variance of the data is less than the mean, which suggests that binomial regression (see Section 3.3.2 of Tutz (2011)) or COM-Poisson regression (see Sellers and Shmueli (2008)) should be suitable. More details on the overdispersion test can be found in Chapter 7 of Hilbe (2011).

When the data is tested to be over-dispersed, we have to correct the hypothetical distributions and then select a flexible distribution, such as some two-parameter models. As an overdispersed distribution, the negative binomial (NB) distribution is a special case of the discrete compound Poisson (DCP) family, which also belongs to the class of infinitely divisible distribution. For more details properties of NB and DCP distribution,

we refer readers to Section 5.9.3 of Johnson et al. (2005), Zhang et al. (2014).

In low and fixed dimensional regressions, it is often to use maximum likelihood estimator (MLE) of regression coefficients. The subsequent sections will frequently employ the average negative log-likelihood of NBR (i.e. a convex empirical process indexed by n):

$$\ell_n(\boldsymbol{\beta}) := -\frac{1}{n}\sum_{i=1}^{n}[Y_i\boldsymbol{X}_i^T\boldsymbol{\beta} - (\theta + Y_i)\log(\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}})],$$

see Section 2.1 below for details. The $\ell_n(\boldsymbol{\beta})$ is also termed as the empirical NBR loss function from the machine learning point of view. If it is given θ or treated as tuning parameter, the NBR actually belongs to the generalized linear models (GLMs) with non-canonical links. The coefficient of Y_i in the log-likelihood of common GLMs with canonical link function is linear in $\boldsymbol{X}_i^T\boldsymbol{\beta}$, while the coefficient of Y_i in log-likelihood of NBR is non-linear in $\boldsymbol{X}_i^T\boldsymbol{\beta}$ which is due to the non-canonical link function.

In high-dimensional setting, a powerful tool for remedying MLE is by adding the penalty function to the $\ell_n(\boldsymbol{\beta})$ to get the penalized (regularized) likelihood estimator. We prefer to study Elastic-net regularized MLE defined as follow.

Definition 1. *(Elastic-net method of NBR) For the empirical NB loss function $\ell_n(\boldsymbol{\beta})$, let $\lambda_1, \lambda_2 > 0$ be turning parameters, the Elastic-net estimates is defined as*

$$\hat{\boldsymbol{\beta}} =: \hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\{\ell_n(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\}. \tag{1}$$

where $\|\boldsymbol{\beta}\|_q := (\sum_{i=1}^p |\beta_i|^q)^{1/q}$ is the l_q-norm of β, $1 \le q < \infty$.

In the section below, we usually denote $\hat{\boldsymbol{\beta}}$ as $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ for simplicity.

Chapter 3 of Tutz (2011) begins with three golden criteria for penalized estimation

method for sparse of coefficient vector:

1°. *Existence of unique estimates - this is where maximum likelihood estimates often fail;*

2°. *Prediction accuracy - a model should be able to yield a decent prediction of the outcome;*

3°. *Sparseness and interpretation - the parsimonious model that contains the strongest effects is easier to interpret than a big model with hardly any structure.*

For 3°, the penalty function in this paper we study is Elastic-net estimate due to that Elastic-net enjoys the merit of both Lasso and Ridge, see Zou and Hastie (2005). Lasso can only select one variable in a group of highly related variables, but Elastic-net can choose more than one, this phenomenon is called a grouping effect. As for 1° and 2°, we concentrate on the non-asymptotic oracle inequalities of the Elastic-net penalized maximum likelihood estimator in NB regression, because asymptotic distribution of the high-dimensional penalized estimator is usually not available. Essentially, deriving oracle inequalities is a powerful mathematical skill which gives deep insight into the non-asymptotic fluctuation of an estimator as compared to that of an ideal unknown estimator (which is called an oracle). Wang et al. (2016) compared the negative binomial regression and Poisson regression models based on the Elastic-net, MCP-net and SCAD-net penalty functions by using the hospitalization days in hospitalized pediatric cardiac surgery and the associated covariates for variable selection analysis. Massaro (2016) constructed the Elastic-net penalized NBR to analyze the over-dispersed count data: time-to-death (in days), the Elastic-net selected functional characteristics of genes that increased or decreased the survival time in the high-dimensional scenario $p \gg n$. In practice, the covariates are habitually corrupted, since it contains measurement error. Recently, Sørensen et al. (2018) suggested that Elastic-net penalty (or generalized

Elastic-net penalty with higher-order terms, such as cubic, quadratic terms, etc.) can decorrupted the corrupted covariates in high-dimensional GLMs with the natural link, the idea is by specifically choosing the second tuning parameter in Elastic-net.

Contributions:

- For GLMs, Bunea (2008) investigated oracle inequalities in the setting of logistic and linear regression models for the Elastic-net penalization schemes under the Stabil Condition. However, by extending the proofs from Bunea (2008), Blazere et al. (2014) kept a watchful eye on deriving oracle inequalities for GLMs with canonical link function which does not contain NBR. Empirical processes technique is utilized by Blazere et al. (2014) to get oracle inequalities for Elastic-net in GLMs, but their assumption of GLMs does not contain the NBR. Even under the fixed design, the Hessian matrix of the NB log-likelihood contains the random responses, this complex phenomenon is substantially distinct from the canonical link GLMs. More treatments about the concentration of random Hessian matrix is needed to deal with. To show the KKT-like event with high probability, we proposed a new concentration inequality for superama of multiplier NB empirical processes.

- Moreover, van de Geer (2008) mainly studied oracle inequalities for high-dimensional GLMs with Lipschitz loss functions, but the loss of NBR is not Lipschitz owing to the unbounded responses. To handle the non-Lipschitz loss, we have to make sure the stochastic Lipschitz property (see Chi (2010)) of the NB loss with high probability. Thus we enable to derive oracle inequalities for Elastic-net estimates for NBR under the compatibility factor and Condition Stabil, and this is different

from conditions in van de Geer (2008).

- Except the ℓ_1-consistency, it is worth noting that, the sign consistent (Zhao and Yu (2006)) of the Elastic-net type estimators are not frequently studied in documental records, see Jia and Yu (2010) for the linear model and Yu (2010) for the Cox model. Based on bounded covariates assumption, we study the sign consistency of Elastic-net regularized NBR without using Irrepresentable Condition.

This paper aims to study the theoretical properties of the Elastic-net methods for sparse estimator in NBR within the framework of the non-asymptotics theory. Section 2.1 and Section 2.2 present a review of NBR and KKT conditions. In Section 2.3,2.4, we showed that, two types of oracle inequalities can be derived for ℓ_1 estimation and prediction error bound under the assumption of compatibility factor condition or Stabil Condition with measurement error, respectively. The remaining sections are byproducts of our proposed oracle inequalities. Typically phenomenon for Elastic-net, we establish a uniform bound for the grouping effect in the Section 3.1. To obtain sign consistency in Section 3.2, the requirement of uniform signal strength that we can detect coefficients larger than a constant multiplied by the tuning parameter of the ℓ_1 penalty is needed. Using the weakest signal condition, in Section 3.3, we arrive at that, the probability of correct inclusion for all true variables in the selected set $\hat{H}$ and the probability of corrected subset selection is high. We discuss the de-biased Elastic-net regularized M-estimators for low-dimensional parameters in Section 3.4. The simulation study is provided in Appendix A. All proofs of main theorems, lemmas and propositions are given in Appendix A, and assisted lemmas are postponed Appendix B.

2 High-dimensional Negative Binomial Regression

In following two subsections, we review the negative binomial GLMs and the corresponding mathematical optimization problems.

2.1 Negative Binomial Regression

The probability mass function of the negative binomial distribution random variable is $p_n =: P(Y = n) = \frac{\Gamma(n+\theta)}{\Gamma(\theta)n!}(1-p)^\theta p^n, (p \in (0,1), n \in \mathbb{N})$. The expectation and variance of the NB distribution are $\frac{\theta p}{1-p}$ and $\frac{\theta p}{(1-p)^2}$. If θ is positive integer, it is called Pascal distribution. This special case of NB is modeled as the number of failures $Y = n$ before the θ-th success in repeated mutually independent Bernoulli trials (with success probability $1-p$). The θ is positive integer or real number sometimes.

In the regression setting, one type of negative binomial regressions (NBR) assumes that the count data response obeys the NB distribution (denoted as $Y \sim \mathrm{NB}(\mu_i, \theta)$) with over-dispersion:

$$P(Y_i = y_i|\boldsymbol{X}_i) =: f(y_i, \theta, \mu_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(\theta)y_i!}(\frac{\mu_i}{\theta+\mu_i})^{y_i}(\frac{\theta}{\theta+\mu_i})^\theta, (i = 1, 2, \cdots, n)$$

Here $\mathrm{E}(Y_i\,|\boldsymbol{X}_i) = \mu_i, \mathrm{Var}(Y_i\,|\boldsymbol{X}_i) = \mu_i + \frac{\mu_i^2}{\theta}$. The θ is a qualification of level of overdispersion that underlies in a count data set, and θ is the known dispersion parameter which can be estimated (see Section 8 of Hilbe (2011)). When the mean parameter μ_i and the covariates are linked by $\log\mu_i = \boldsymbol{X}_i^T\boldsymbol{\beta}$, we have a NBR. When $\theta \to +\infty$, it implies $\mathrm{Var}(Y_i\,|\boldsymbol{X}_i) \to \mu_i = \mathrm{E}(Y_i\,|\boldsymbol{X}_i)$. Thus the Poisson regression is a limiting case of NBR when the dispersion parameter tends to infinite. As overdispersion occurs in real data, NBR can be more powerful and interpretable than Poisson regression.

The log-likelihood function of NB responses is:

$$\begin{aligned} L(\boldsymbol{Y};\boldsymbol{\beta}) &= \log[\prod_{i=1}^{n} f(Y_i,\theta,\mu_i)] = \sum_{i=1}^{n} \log\{\frac{\Gamma(\theta+Y_i)}{\Gamma(\theta)Y_i!}(\frac{\mu_i}{\theta+\mu_i})^{Y_i}(\frac{\theta}{\theta+\mu_i})^{\theta}\} \\ &= \sum_{i=1}^{n} \{\log\Gamma(\theta+Y_i) + Y_i\log\mu_i + \theta\log\theta - \log\Gamma(\theta) - \log Y_i! - (\theta+Y_i)\log(\theta+\mu_i)\} \\ &= c_0 + \sum_{i=1}^{n} [Y_i \boldsymbol{X}_i^T\boldsymbol{\beta} - (\theta+Y_i)\log(\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}})] \text{ with a constant } c_0. \end{aligned}$$

Then, we take the derivative of the vector $\boldsymbol{\beta}$. Let $\frac{\partial L(\boldsymbol{Y};\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \{\frac{\partial L(\boldsymbol{Y};\boldsymbol{\beta})}{\partial\beta_1},\cdots,\frac{\partial L(\boldsymbol{Y};\boldsymbol{\beta})}{\partial\beta_p}\}^T$, we get $\frac{\partial L(\boldsymbol{Y};\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{X}_i\theta[\frac{\theta+Y_i}{\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}} - 1] = \sum_{i=1}^{n} \frac{\boldsymbol{X}_i(Y_i-e^{\boldsymbol{X}_i^T\boldsymbol{\beta}})\theta}{\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}$. Besides, by setting score function to be 0, $\frac{\partial L(\boldsymbol{Y};\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 0$, we obtain $\hat{\boldsymbol{\beta}}_{mle}$. The second derivative is calculated by $\frac{\partial^2 L(\boldsymbol{Y};\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = -\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^T \frac{\theta(\theta+Y_i)e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{(\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}})^2}$ which is semi-negative, so that $\hat{\boldsymbol{\beta}}_{mle}$ makes the likelihood function take the maximum value.

2.2 KKT conditions

For generalized Lasso-type convex penalty (GLCP) criterion, Yu (2010) considers penalized likelihood for convex loss function $\ell(\boldsymbol{\beta})$

$$F(\boldsymbol{\beta};\lambda_1,\lambda_2) = \ell(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 g(\boldsymbol{\beta})$$

where $g(\boldsymbol{\beta})$ is a nonnegative convex function with $g(\mathbf{0}) = \mathbf{0}$, λ_1,λ_2 being positive turning parameters. The GLCP estimation problem for general log-likelihood is $\hat{\boldsymbol{\beta}}(\lambda_1,\lambda_2) = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} F(\boldsymbol{\beta};\lambda_1,\lambda_2)$. By sub-derivative technique in the optimization function, the corresponding Karush-Kuhn-Tucker conditions of GLCP estimator is

$$\begin{cases} \dot{\ell}_j(\hat{\boldsymbol{\beta}}) + \lambda_2\dot{g}_j(\hat{\boldsymbol{\beta}}) = -\lambda_1\text{sign}(\hat{\beta}_j) & \text{if } \hat{\beta}_j \neq 0, \\ |\dot{\ell}_j(\hat{\boldsymbol{\beta}}) + \lambda_2\dot{g}_j(\hat{\boldsymbol{\beta}})| \leq \lambda_1 & \text{if } \hat{\beta}_j = 0, \end{cases} \tag{2}$$

(KKT conditions, see page68 of Bühlmann and van de Geer (2011)). Thus the KKT conditions for the non-zero (or zero) Elastic-net estimate of the NBR.

Lemma 1. *(Necessary and Sufficient Condition) Let* $k \in \{1, 2, \cdots, p\}$ *and* $\lambda_2 > 0$. *Then, a necessary and sufficient condition for Elastic-net estimates of NBR to be a solution of* (1) *is*

1. $\hat{\beta}_k = \hat{\beta}_k \neq 0$ *if* $\frac{1}{n}\sum_{i=1}^{n} x_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-Y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}} = [\text{sign}\hat{\beta}_k](\lambda_1 + 2\lambda_2|\hat{\beta}_k|)$.
2. $\hat{\beta}_k = 0$ *if* $\left|\frac{1}{n}\sum_{i=1}^{n} x_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-Y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}\right| \leq \lambda_1$.

Zhou (2013) gives an elementary proof of KKT conditions for the Elastic-net penalized optimization problem merely in linear regression. It is worth noting that KKT conditions are a standard result of sub-differentiation techniques. But here to apply some identities to prove Lemma A.11 in Section 3.1, we give a detailed proof of the above lemma. The prerequisite $\lambda_2 > 0$ in Lemma 1 is indispensable. The reason is that we need $\lambda_2 > 0$ such that $F(\hat{\boldsymbol{\beta}} + \varepsilon\mathbf{e}_k; \lambda_1, \lambda_2) - F(\hat{\boldsymbol{\beta}}; \lambda_1, \lambda_2) > 0$ in the lines of proof, see Appendix B, and then $\hat{\boldsymbol{\beta}}$ is the unique locally minimum. The KKT conditions are crucial for all sections below.

2.3 ℓ_q-estimation error via Compatibility Factor

In this part, we are going to present that the sparse estimator for high-dimensional negative binomial regression by using Elastic-net regularization is asymptotically close to the true parameter under some suitable regularity conditions.

For fixed designs $\{\boldsymbol{X}_i\}_{i=1}^n$, let $\boldsymbol{\beta}^*$ be the true coefficients vector, which satisfies

$$\text{E}Y_i = e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}. \tag{3}$$

In some sense, we can never really know the expectation of the negative log-likelihood because $\boldsymbol{\beta}^*$ is the unknown parameter to be estimated. In high-dimension, we are interested in the sparse estimates defined in (1) by adding Elastic-net penalty. For the true coefficient vector $\boldsymbol{\beta}^* = (\beta_1^*, \cdots, \beta_p^*)^T$, let $H = \{j : \beta_j^* \neq 0, j = 1, \cdots, p\}, H^c = \{j : \beta_j^* = 0, j = 1, \cdots, p\}$ be the nonzero and zero components respectively, and let $d_H^* = |H|$ be the number of non-zero coefficients in $\boldsymbol{\beta}^*$. For any index set $H \in \{1, 2, \cdots, p\}$, define the sub-vector indexed by H as $\boldsymbol{b}_H = (\cdots, \tilde{b}_j, \cdots)^T \in \mathbb{R}^p$ with $\tilde{b}_j = b_j$ if $j \in H$ and $\tilde{b}_j = 0$ if $j \notin H$. We know that the Kullback-Leibler divergence measures that how one probability distribution is different from a second. Similarly, in order to measure the distance between two penalized log-likelihood index by its parameter, the *symmetric Bregman divergence* between $\ell(\hat{\boldsymbol{\beta}})$ and $\ell(\boldsymbol{\beta})$ is

$$D_g^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T [\dot{\ell}(\hat{\boldsymbol{\beta}}) - \dot{\ell}(\boldsymbol{\beta}) + \lambda_2(\dot{g}(\hat{\boldsymbol{\beta}}) - \dot{g}(\boldsymbol{\beta}))].$$

If $g = 0$, the symmetric Bregman divergence is $D^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T [\dot{\ell}(\hat{\boldsymbol{\beta}}) - \dot{\ell}(\boldsymbol{\beta})]$. In this case, the symmetric Bregman divergence is a type of generalized quadratic distances (Mahalanobis distances) which can been seen as the symmetric extension of Kullback-Leibler divergence. See Nielsen and Nock (2009), Huang et al. (2013) for more discussions about symmetric Bregman divergence. Since $g(\boldsymbol{\beta})$ is a nonnegative convex function, we deduce the quantitative relation: $D_g^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \geq D^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$. With the above definitions, let $z^* = \|\dot{\ell}(\boldsymbol{\beta}^*) + \lambda_2 \dot{g}(\boldsymbol{\beta}^*)\|_\infty$ and $\Delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, we now provide the lower and upper bounds for the symmetric Bregman divergence.

Lemma 2 (Theorem 1 in Yu (2010))**.** *For GLCP estimation, we have*

$$(\lambda_1 - z^*)||\Delta_{H^c}||_1 \leq D_g^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + (\lambda_1 - z^*)||\Delta_{H^c}||_1 \leq (\lambda_1 + z^*)||\Delta_H||_1. \tag{4}$$

If $z^* \leq \frac{\zeta-1}{\zeta+1}\lambda_1$ for some $\zeta > 1$, the inequalities and implies

$$\frac{2\lambda_1}{\zeta+1}||\Delta_{H^c}||_1 \leq D_{\mathrm{g}}^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + \frac{2\lambda_1}{\zeta+1}||\Delta_{H^c}||_1 \leq \frac{2\zeta\lambda_1}{\zeta+1}||\Delta_H||_1, \tag{5}$$

which is from the fact that $\lambda_1 - z^* \geq \frac{2\lambda_1}{\zeta+1}$ and $\lambda_1 + z^* \leq \frac{2\zeta\lambda_1}{\zeta+1}$.

By (5), we have

$$||\Delta_{H^c}||_1 \leq \zeta\|\Delta_H\|_1. \tag{6}$$

Hence we conclude that in the event $\mathcal{K}_\lambda := \left\{z^* \leq \frac{\zeta-1}{\zeta+1}\lambda_1\right\}$, the error of estimate $\Delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ belongs to the *cone set*:

$$\mathrm{S}(s, H) := \{\boldsymbol{b} \in \mathbb{R}^p : ||\boldsymbol{b}_{H^c}||_1 \leq s||\boldsymbol{b}_H||_1\}, \ (s \in \mathbb{R}). \tag{7}$$

with $s = \zeta$.

The key of deriving oracle inequalities also depends on the behaviour of empirical covariance matrix, namely, the weighted Gram matrix:

$$\ddot{\ell}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T \frac{\theta(\theta+Y_i)e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\left(\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\right)^2} := \frac{1}{n}\sum_{i=1}^n \tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^T, \ \text{where } \tilde{\boldsymbol{X}}_i := \boldsymbol{X}_i\left(\frac{\theta(\theta+Y_i)e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\left(\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\right)^2}\right)^{1/2}.$$

In increasing dimension setting $p = p(n)$, it is well-known that the Gram matrix $\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T$ (i.e. the correlation between the covariates) which is necessarily singular when $p > n$. In order to obtain oracle inequality with a fast rate as discussed in Bickel et al. (2009), the following versions of restricted eigenvalues is usually needed under the restriction (7).

The *compatibility factor* (denoted by $C(s, H, \boldsymbol{\Sigma})$, see van de Geer (2007)) of a $p \times p$ nonnegative-definite matrix $\boldsymbol{\Sigma}$, is defined by

$$C^2(s, H, \boldsymbol{\Sigma}) := \inf_{0 \neq \boldsymbol{b} \in \mathrm{S}(s,H)} \frac{d_H^*(\boldsymbol{b}^T\boldsymbol{\Sigma}\boldsymbol{b})}{\|\boldsymbol{b}_H\|_1^2} > 0, \quad (s \in \mathbb{R}). \tag{8}$$

For the sake of deriving ℓ_q-loss ($q > 1$) oracle inequalities for target coefficient vectors, we require the concept of *weak cone invertibility factors* (weak CIF, see (53) of Ye and Zhang (2010)),

$$C_q(s, H, \mathbf{\Sigma}) := \inf_{0 \neq \boldsymbol{b} \in \mathrm{S}(s,H)} \frac{{d_H^*}^{1/q}(\boldsymbol{b}^T \mathbf{\Sigma} \boldsymbol{b})}{||\boldsymbol{b}_H||_1 \cdot ||\boldsymbol{b}||_q} > 0, (s \in \mathbb{R}). \tag{9}$$

This constant generalizes the compatibility factor and is close to the restricted eigenvalue (see Bickel et al. (2009)). From the results in Ye and Zhang (2010) and Huang et al. (2013), we know that the compatibility factor and weak CIF can achieve a sharper upper bounds for the oracle inequalities since both of them are bigger than the restricted eigenvalue.

The condition $C^2(s, H, \mathbf{\Sigma}) > 0$ or $C_q(s, H, \mathbf{\Sigma}) > 0$ for Hessian matrix $\mathbf{\Sigma} = \ddot{\ell}(\boldsymbol{\beta})$ is an indispensable assumptions for deriving the targeted oracle inequalities.

Some additional regularity conditions are required.

- (C.1): Bounded covariates, $\max\{|x_{ij}|\,; 1 \leq i \leq n, 1 \leq j \leq p\} = L < \infty$.
- (C.2): We assume the identifiability condition that $\boldsymbol{X}_i^T(\boldsymbol{\beta} + \boldsymbol{\delta}) = \boldsymbol{X}_i^T \boldsymbol{\beta}$ implies $\boldsymbol{X}_i^T \boldsymbol{\delta} = 0$ for $\boldsymbol{\delta} \in \mathbb{R}^p$.
- (C.3): Suppose that $||\boldsymbol{\beta}^*||_1 \leq B$.

The bounded covariates C.1 is a common assumption in GLMs (see Example 5.40 of van der Vaart (1998)), it may be achieved by doing some bounded transformation of the covariates in real data. The identifiability condition C.2 and compact parameter space C.3 are common assumptions for obtaining consistency for general M-estimation, see section 5.5 and remark of Theorem 5.9 in van der Vaart (1998). Recently, Weißbach

and Radloff (2019) shows the consistency for the NBR with fixed covariates under the assumption that all possible parameters and regressor are in the compact space.

At first, we present the non-asymptotic upper bounds for Elastic-net regularized NBR in following two theorems.

Theorem 1. *Let* $C(\zeta, H) := C(\zeta, H, \ddot{\ell}(\boldsymbol{\beta}^*))$ *and* $C_q(\zeta, H) := C_q(\zeta, H, \ddot{\ell}(\boldsymbol{\beta}^*))$ *to be the compatibility factor and the weak cone invertibility factor defined above. Define* $\tau := \frac{L(\zeta+1)d^*\lambda_1}{2[C(\zeta,H)]^2} \leq \frac{1}{2}e^{-1}$. *Then, assume that (C.1), (C.2) and the event* $\mathcal{K}_\lambda$ *hold, we have*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{e^{2a_\tau}(\zeta+1)d_H^*\lambda_1}{2C^2(\zeta, H)} \quad \text{and} \quad \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q \leq \frac{2e^{2a_\tau}\zeta {d_H^*}^{1/q}\lambda_1}{(\zeta+1)\, C_q(\zeta, H)} \tag{10}$$

where $a_\tau \leq \frac{1}{2}$ *is the smaller solution of the equation* $ae^{-2a} = \tau$.

On the one hand, the Theorem 1 contains basic oracle inequalities conditioning on the random event, which needs further refinements. What remains to be done is to focus the probability upper bound of event $\mathcal{K}_\lambda$. With assumption (C.3), we have $z^* \leq \|\dot{\ell}(\boldsymbol{\beta}^*)\|_\infty + 2\lambda_2 B$. Our aim of proof is to have

$$P(\mathcal{K}_\lambda^c) \leq P(||\dot{\ell}(\boldsymbol{\beta}^*)||_\infty + 2\lambda_2 B \geq \frac{\zeta-1}{\zeta+1}\lambda_1) \to 0 \quad \text{as } n, p \to \infty. \tag{11}$$

That all we need is to apply some concentration inequality in terms of NB empirical processes, i.e. sum of independent weighted NB random variables, these types of concentration inequalities have been constructed recently, see Zhang and Wu (2019) and references therein. As the dispersion parameter θ is known, then NB random variables $\{Y_i\}_{i=1}^n$ belong to the exponential family $f(y_i; \eta_i) \propto \exp\{y_i\eta_i - \psi(\eta_i)\}$ with $\eta_i := \boldsymbol{X}_i^T\boldsymbol{\beta}^* + \log(\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}) \in \Theta$, where Θ is the compact parameter space.

On the other hand, the methods in term of compatibility factor and weak CIR that

we employ in this section are yet random constants. It contains the Hessian matrix of the true coefficient vector and thus it encapsulates the random quantities $\{Y_i\}_{i=1}^n$. We should note that deriving the lower bound for these random quantities will decrease the probability that oracle inequalities are true, but the loss is negligible in next theorem. Next, we successfully show by NB concentration inequality again with some tricks that there exist much reasonable non-random lower bounds of the compatibility factor (or the weak CIR) such that the upper bounds are constants with high probability, thus the rigorous convergence rate of $\hat{\boldsymbol{\beta}}$ is well established. It should be noted that Yu et al. (2020) directly assumes that the inverse of compatibility factor of $\ddot{\ell}(\boldsymbol{\beta}^*)$ for the Cox model is $O_p(1)$, they call it "a high-level condition". The Hessian matrix of the Cox model is also a random element.

Two events for truncating the compatibility factor and the weak CIR, is defined by $\mathcal{E}_c := \{C^2(\zeta, H) > C_t^2(\zeta, H)\}$ and $\mathcal{E}_w := \{C_q(\zeta, H) > C_{qu}(\zeta, H)\}$ respectively, where $C_t^2(\zeta, H)$ and $C_{qu}(\zeta, H)$ are non-random constants defined in the proof.

Theorem 2. *Under assumptions of Theorem 1, we further assume (C.2). Let B_1 be the constant satisfying $C_{\xi,B_1} := \frac{\zeta-1}{\zeta+1} - 2B_1 > 0$. Let $\lambda_1 = \frac{C_{LB}L}{C_{\xi,B_1}}\sqrt{\frac{2r\log p}{n}}$ where $C_{LB}^2 := e^{LB} + \frac{e^{2LB}}{\theta}$ is a variance-depending constant and $r > 1$ is a constant. Put $\lambda_2 = B_1\lambda_1/B$. On the event $\mathcal{K} \cap \mathcal{E}_c$ (or $\mathcal{K} \cap \mathcal{E}_w$), we have:*

$$P\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{e^{2a_\tau}(\zeta+1)d_H^*\lambda_1}{2C_t^2(\zeta, H)}\right) \geq 1 - \frac{2}{p^{r-1}} - 2p^2 e^{-\frac{nt^2}{2[d_H^* C_{LB}(1+\varsigma)L^2]^2}} \tag{12}$$

$$or\ P\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q \leq \frac{2e^{2a_\tau}\zeta {d_H^*}^{1/q}\lambda_1}{(\zeta+1)\,C_{qu}(\zeta, H)}\right) \geq 1 - \frac{2}{p^{r-1}} - 2p^2 e^{-\frac{nu^2}{2[d_H^* C_{LB}(1+\varsigma)L^2]^2}}. \tag{13}$$

If we presume case $d_H^* = O(1)$ in Theorem 2, which implies that the error bounds is of the order $\sqrt{\frac{\log p}{n}}$ so the Elastic-net estimates have consistent property for the ℓ_1-

error when dimension of covariates could increase with order $e^{o(n)}$. As we know, the MLE has the convergence rate $\frac{1}{\sqrt{n}}$. Nevertheless, in order to pay the price in high-dimensional condition, we have to magnify $\sqrt{\log p}$ to the convergence rate of MLE. If we assume $d_H^* = o(\sqrt{\frac{n}{\log p}})$, i.e. $p = e^{o(n/d_H^*)}$, thus $d_H^*\lambda = o(1)$ and it also implies the consistent property. Under the random designs, the purpose of next section is to give a new approach that avoids the random upper bound for the ℓ_1 or ℓ_2 estimation error and provides square prediction error oracle inequality.

2.4 The prediction error via stabil condition with random design

In this section, we focuses on the prediction error. We presume that the $n \times p$ design matrix $\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^T$ is random. In applications, the test data set is a new design $\mathbf{X}^*$ which is an independent copy of $\mathbf{X}$, thus it requires the randomness assumption of the design matrix. This section aims to predict the response Y_{n+1} by the new random covariates $\boldsymbol{X}_{n+1}$ by resorting Elastic-net estimator $\hat{\boldsymbol{\beta}}$ to estimates the unknown Y_{n+1}.

The $\boldsymbol{Y} \in \mathbb{R}^n$ contains n independently (ind. in short) responses $\{Y_i\}_{i=1}^n$, thus the covariates and responses are considered as a pair of random vectors ($\mathbf{X}$, $\boldsymbol{Y}$). When $\{\boldsymbol{X}_i\}_{i=1}^n$ is degenerate distributed, it is just a case of fixed design and hence the result in this part also holds for fixed design. Through the paper, we denote the element in design matrix $\{x_{ij}\}$ as fixed design, $\{X_{ij}\}$ as random design. The conditional distribution of a single observations is $Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i$ is assumed to be conditional NB distributed with $\mathrm{E}(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) = e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}$.

Let $\boldsymbol{\beta}^*$ be the true coefficients vector, which is defined by the minimiser

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\, \mathrm{E} l(Y, \boldsymbol{X}, \boldsymbol{\beta}) \tag{14}$$

where $l(Y, \boldsymbol{X}, \boldsymbol{\beta}) = Y\boldsymbol{X}^T\boldsymbol{\beta} - (\theta + Y)\log(\theta + e^{\boldsymbol{X}^T\boldsymbol{\beta}})$ is the NB loss.

To derive non-asymptotical bounds for ℓ_1-estimation and square prediction error, we have to focus on the empirical process for any possible $\boldsymbol{\beta}$ [on NB loss function in (14) with random $\mathbf{X}$]

$$\mathbb{P}_n l(\boldsymbol{X}, Y, \boldsymbol{\beta}) := -\frac{1}{n}\sum_{i=1}^{n}[Y_i\boldsymbol{X}_i^T\boldsymbol{\beta} - (\theta + Y_i)\log(\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}})]$$

where $\mathbb{P}_n$ is the empirical measure of samples $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n \overset{\text{ind.}}{\sim} (\boldsymbol{X}, Y)$.

The analysis of concentration and fluctuation of the empirical process is crucial to evaluate the consistent properties of the estimates. For simplicity, we use the symbol language of the empirical process in this section. We need some assumptions such that $\hat{\boldsymbol{\beta}}$ could be consistent in high-dimensional NBR.

- (H.1): All the variables $\boldsymbol{X}_i$ are bounded: there exists a constant $L > 0$ such that

$$|||\mathbf{X}|||_\infty := \sup_{1\le i\le\infty} \|\boldsymbol{X}_i\|_\infty \le L \text{ a.s.}$$

- (H.2): Assume that $||\boldsymbol{\beta}^*||_1 \le B$.

- (H.3): There exists a large constant M_0 such that $\hat{\boldsymbol{\beta}}$ is in the ℓ_1-ball:

$$\hat{\boldsymbol{\beta}} \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \le M_0\}.$$

- (H.4): Let $\theta > 1$. The negative log-density of n independent NB responses $\psi(\boldsymbol{y}) := -\log p_{\boldsymbol{Y}}(\boldsymbol{y})$ for $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ satisfies the *strongly midpoint log-*

convex properties for some $\gamma > 0$

$$\psi(\boldsymbol{x}) + \psi(\boldsymbol{y}) - \psi(\lceil \frac{1}{2}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y} \rceil) - \psi(\lfloor \frac{1}{2}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y} \rfloor) \geq \frac{\gamma}{4}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n. \quad (15)$$

Remark 1. *The (H.1) and (H.2) are mentioned in Blazere et al. (2014), and the (H.3) is a high technique condition due to the non-canonical link GLMs. The constraint in the optimization is equivalent to* $\alpha \|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2 \leq t$ *with unknown* $\alpha \in [0,1]$ *and* $t \in \mathbb{R}$*, it lead to* $\|\hat{\boldsymbol{\beta}}\|_1 \leq M_0$ *if we suppose that* $t/\alpha \leq M_0$*. There is a constant* $K > 0$ *such that* $\max_{1 \leq i \leq n} \left| \boldsymbol{X}_i^T \boldsymbol{\beta}^* \right| \leq K$ *a.s. for all* n*. A convex function* F *is called strongly convex if the Hessian matrix of* F *has a (uniformly) lower bounded eigenvalue. In learning exponential families in high-dimensions, Kakade et al. (2010) assumed that continuous exponential families is a strongly convex log-likelihood function with* η_i *in a sufficiently small neighborhood. For fixed dimensional MLE, Balabdaoui et al. (2013) shown that discrete log-concave maximum likelihood estimator is strongly consistent under some settings. Our assumption (H.4) is a technique condition which makes sure the suprema of the multiplier empirical processes of* n *independent responses have sub-Gaussian concentration phenomenon. For the case of fixed design in Section 2.3, we do not require (H.4) to derive oracle inequalities.*

In this section, we give sharp bounds for ℓ_1-estimation and square prediction errors for NBR models by looking for a weaker condition which is analogous to the restricted eigenvalue condition (RE) proposed by Bickel et al. (2009) and the weak CIF and compatibility factor conditions presented in Section 3.2. Here we borrow a condition which is from the Stabil Condition introduced by Bunea (2008) for ℓ_1 and $\ell_1 + \ell_2$ penalized logistic regressions.

For c, $\varepsilon > 0$, we define the *fluctuated cone set* for some bias vector $\boldsymbol{b}$ as

$$\mathrm{V}(c, \varepsilon, H) := \{\boldsymbol{b} \in \mathbb{R}^p : ||\boldsymbol{b}_{H^c}||_1 \leq c||\boldsymbol{b}_H||_1 + \varepsilon\}. \tag{16}$$

which is a fluctuated (or measurement error) version of the cone set $\mathrm{S}(s, H) := \{\boldsymbol{b} \in \mathbb{R}^p : ||\boldsymbol{b}_{H^c}||_1 \leq s||\boldsymbol{b}_H||_1\}$ mentioned in (7).

We will plugging $\boldsymbol{b} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ in the proof. For real data, let $\hat{\boldsymbol{\beta}}$ be the estimator based on true covariates and $\hat{\boldsymbol{\beta}}_{me}$ be the the estimator from covariates with measurement error. Note that under the cone condition $||\boldsymbol{b}_{H^c}||_1 \leq c||\boldsymbol{b}_H||_1$ for $\boldsymbol{b} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, we get

$$\begin{aligned} ||(\hat{\boldsymbol{\beta}}_{me} - \boldsymbol{\beta}^*)_{H^c}||_1 - ||(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{me})_{H^c}||_1 \leq ||(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{H^c}||_1 &\leq c||(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_H||_1 \\ &\leq c||(\hat{\boldsymbol{\beta}}_{me} - \boldsymbol{\beta}^*)_H||_1 + c||(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{me})_H||_1 \end{aligned}$$

Then,

$$||\boldsymbol{b}^{me}_{H^c}||_1 \leq c||\boldsymbol{b}^{me}_H||_1 + \varepsilon \quad \text{for } \boldsymbol{b}^{me} := \hat{\boldsymbol{\beta}}_{me} - \boldsymbol{\beta}^*$$

where $\varepsilon = c||(\hat{\boldsymbol{\beta}}_{me} - \boldsymbol{\beta}^*)_H||_1 + ||(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{me})_{H^c}||_1$. This argument indicates that the fluctuated cone set quantifies the level of measurement error if $\hat{\boldsymbol{\beta}}_{me}$ is misspecified as $\hat{\boldsymbol{\beta}}$.

On fluctuated cone set, we assume that the $p \times p$ matrix $\boldsymbol{\Sigma}$ fulfills the Stabil condition as below. For example, the $\boldsymbol{\Sigma} = \mathrm{E}\boldsymbol{X}\boldsymbol{X}^T$ is the expected empirical covariance matrix.

Definition 2. *(Stabil Condition with measurement error) For given $c, \varepsilon > 0$, the matrix $\boldsymbol{\Sigma}$ satisfies the Stabil condition $S(c, \varepsilon, k)$ if there exists $0 < k < 1$ such that*

$$\boldsymbol{b}^T\boldsymbol{\Sigma}\boldsymbol{b} \geq k||\boldsymbol{b}_H||_2^2 - \varepsilon$$

for any $\boldsymbol{b} \in V(c, \varepsilon, H)$. Here the restriction $0 < k < 1$ can be attained by scaling the $\boldsymbol{X}$.

Let $l_1(\boldsymbol{\beta}) := l_1(\boldsymbol{\beta}, \boldsymbol{X}, Y) := -Y[\boldsymbol{X}^T\boldsymbol{\beta} - \log(\theta + \exp\{\boldsymbol{X}^T\boldsymbol{\beta}\})]$ which is a linear function of response, and let $l_2(\boldsymbol{\beta}) := l_2(\boldsymbol{\beta}, \boldsymbol{X}) := \theta\log(\theta + \exp\{\boldsymbol{X}^T\boldsymbol{\beta}\})$ which is free of

response. The NB loss function $l(\boldsymbol{\beta}, \boldsymbol{X}, Y) = l_1(\boldsymbol{\beta}, \boldsymbol{X}, Y) + l_2(\boldsymbol{\beta}, \boldsymbol{X})$ is thus decomposed into two part. Let $\mathbb{P}l(\boldsymbol{\beta}) := \mathrm{E}l(\boldsymbol{\beta}, \boldsymbol{X}, Y)$ be the expected risk function, where the expectation is under the randomness of $(\boldsymbol{X}, Y)$. We are fond of the centralized empirical loss $(\mathbb{P}_n - \mathbb{P})\, l(\boldsymbol{\beta})$ which represents the fluctuation between the expected loss and sample loss, rather than the loss itself. We break down the empirical process into two parts:

$$(\mathbb{P}_n - \mathbb{P})\, l(\boldsymbol{\beta}) = (\mathbb{P}_n - \mathbb{P})\, l_1(\boldsymbol{\beta}) + (\mathbb{P}_n - \mathbb{P})\, l_2(\boldsymbol{\beta}). \tag{17}$$

In the following, we give upper bounds for the first and second part of the empirical process: $(\mathbb{P}_n - \mathbb{P})(l_m(\boldsymbol{\beta}^*) - l_m(\hat{\boldsymbol{\beta}}))$, for $m = 1, 2$. It will be shown that $(\mathbb{P}_n - \mathbb{P})(l_m(\boldsymbol{\beta}^*) - l_m(\hat{\boldsymbol{\beta}}))$ has *stochastic Lipschitz properties* (see Chi (2010)) with respect to $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$. Let the ℓ_1-ball be $\mathcal{S}_{M_0}(\boldsymbol{\beta}^*) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \le M_0\}$ which is referred as the **local set**. Then,

Proposition 1. *Let the centred responses be* $\{Y_i^c := Y_i - \mathrm{E}Y_i\}_{i=1}^n$ *and the (H.1)-(H.4) are satisfied. If* $\lambda_1 \ge 4L(2\tilde{C}_{LB} + A\sqrt{2\gamma})\sqrt{\frac{2\log 2p}{n}}, (A \ge 1, \tilde{C}_{LB}^2 := e^{LB} + \frac{(1+\theta)e^{2LB}}{\theta})$, *define the event* $\mathcal{A}$ *for suprema of the multiplier empirical processes by*

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

we have $P(\mathcal{A}) \ge 1 - (2p)^{-A^2}$. *Moreover,*

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

This proposition indicates the discrepancy between the first part of the empirical process and its expectation is bounded from above by the tuning parameter multiplied by the ℓ_2 norm of the difference between the estimated vector and the target vector.

The $\frac{\lambda_1}{4}$ can be seen as Lipschitz constant of the first part of the centralized empirical process.

Similar to $\mathcal{A}$, we provide a crucial lemma which is to bound the second part of the empirical process with responses. Let $\nu_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*) := \frac{(\mathbb{P}_n - \mathbb{P})(l_2(\boldsymbol{\beta}^*) - l_2(\boldsymbol{\beta}))}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 + \varepsilon_n}$ the normalized second part of the empirical process which is a random variable index by $\boldsymbol{\beta}$, then we define the *local stochastic Lipschitz constant* for a certain $M > 0$

$$Z_M(\boldsymbol{\beta}^*) := \sup_{\boldsymbol{\beta} \in \mathcal{S}_M(\boldsymbol{\beta}^*)} |\nu_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*)|, \text{ and a random event } \mathcal{B} := \{Z_M(\boldsymbol{\beta}^*) \leq \frac{\lambda_1}{4}\}$$

which is by bounding the local stochastic Lipschitz constant with the rescaled tuning parameter $\frac{\lambda_1}{4}$. Moreover, by definition we have $|\nu_n(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)| \leq \sup_{\mathcal{S}_M(\boldsymbol{\beta}^*)} |\nu_n(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)| \leq \frac{\lambda_1}{4}$, which gives following bound.

$$|(\mathbb{P}_n - \mathbb{P})(l_2(\hat{\boldsymbol{\beta}}) - l_2(\boldsymbol{\beta}^*))| \leq \frac{\lambda_1}{4}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \varepsilon_n) \quad \text{on } \mathcal{B}. \tag{18}$$

provided that $\hat{\boldsymbol{\beta}} \in \mathcal{S}_M(\boldsymbol{\beta}^*)$.

The following lemma in accordance with the phenomenon that, in the event $\mathcal{A} \bigcap \mathcal{B}$, the estimator $\hat{\boldsymbol{\beta}}$ lies in a known neighborhood of the true coefficient vector $\boldsymbol{\beta}^*$.

Lemma 3. *Under (H.2), let* $8B\lambda_2 + 4M = \lambda_1$, *we have*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 16||\boldsymbol{\beta}^*|| + 2\varepsilon_n \quad \text{on } \mathcal{A} \bigcap \mathcal{B}.$$

The proof of Lemma 3 is rely on the optimization (1) and the definition of the minimizer $\boldsymbol{\beta}^*$ from the expected loss (14). By Lemma 3, on the event $\mathcal{A} \bigcap \mathcal{B}$ we immediately get $\hat{\boldsymbol{\beta}} \in \mathcal{S}_{16B+2\varepsilon_n}(\boldsymbol{\beta}^*)$. Note that we assume that $\hat{\boldsymbol{\beta}} \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)$ for some finite $M_0 > M = 16B + 2\varepsilon_n$ in (H.3). That is to say the Lemma 3 sharpen the $\hat{\boldsymbol{\beta}}$ in the ℓ_1-ball $\mathcal{S}_M(\boldsymbol{\beta}^*)$, while $\hat{\boldsymbol{\beta}}$ is originally assumed in the ℓ_1-ball $\mathcal{S}_{M_0}(\boldsymbol{\beta}^*)$. Therefore, the follow

probability analysis of the event $\mathcal{A}\bigcap\mathcal{B}$ is indispensable. The the event $\mathcal{A}\cap\mathcal{B}$ associated with empirical loss functions play an important role in deriving the oracle inequalities for general loss functions, since we could bound the ℓ_1-estimation error conditioning on event $\mathcal{A}\bigcap\mathcal{B}$. We now give the result that the event $\mathcal{A}\cap\mathcal{B}$ occurs with high probability.

Proposition 2. *Let* $M = 16B + 2\varepsilon_n$. *Suppose that* $\hat{\boldsymbol{\beta}} \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)$ *for* $\infty > M_0 > M$, *and (H.1)-(H.4) is true. If*

$$\lambda_1 \geq \max\left(\frac{20\theta AML}{M+\varepsilon_n}\sqrt{\frac{2\log 2p}{n}}, 4L(2\tilde{C}_{LB} + A\sqrt{2\gamma})\sqrt{\frac{2\log 2p}{n}}\right), \ A \geq 1, \tag{19}$$

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

The proof of Theorem 3 is based on some lemmas in Appendix A which show that the event $\mathcal{A}\cap\mathcal{B}$ holds with high probability.

On the back of the above probability analysis, now we can formulate the main result of this section that gives bounds for the estimation and prediction error as the target model is sparse and $\log p$ is tiny as compared to n. Especially, the oracle inequality of estimation error is useful in the following sections.

Theorem 3. *Assume condition* $S(3.5, \varepsilon_n, k)$, *(H1)-(H4) is fulfilled. Let* λ_1 *be chosen by* (19) *and* $\lambda_2 \leq \frac{\lambda_1}{8B}$. *Then, in the event* $\mathcal{A}\cap\mathcal{B}$, *we have* $P(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \mathrm{V}(3.5, \frac{\varepsilon_n}{2}, H)) \geq 1 - 2(2p)^{-A^2}$ *and*

$$P\left\{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{2.25^2\lambda_1 d_H^*}{ak + 2\lambda_2} + (1 + \frac{a}{\lambda_1})\varepsilon_n\right\} \geq 1 - 2(2p)^{-A^2}. \tag{20}$$

Moreover, let the test data $(\boldsymbol{X}^*, Y^*)$ *be an independent copy of the the training data* $(\boldsymbol{X}, Y)$, *and denote* $\mathrm{E}^*(\cdot) := \mathrm{E}(\cdot|\boldsymbol{X}^*)$. *Conditioning on the event* $\mathcal{A}\cap\mathcal{B}$, *the square*

prediction error is

$$\mathrm{E}^*[\boldsymbol{X}^{*T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 \leq \frac{17.71875 d_H^* \lambda_1^2}{a(ak+2\lambda_2)} + (\frac{4\lambda_1}{a}+3.5)\varepsilon_n \tag{21}$$

where $a := \min_{\{|x|\leq LM+K, |y|\leq K\}} \{\frac{1}{2}\frac{\theta \mathrm{e}^x(\mathrm{e}^y+\theta)}{[\theta+\mathrm{e}^x]^2}\} > 0.$

Comparing to the upper bounds under compatibility factor condition in Section 2.3, in much the same fashion, we observe that when $d^* = O(1)$ and the number of covariates increases as large as $o(\exp(n))$. Then the bound on the estimation error is of the order $o(1)$ and the Elastic-net estimator ensures the consistent property. The Theorem 3 is also an improvement of Lemma 3 from a big neighbourhood of $\boldsymbol{\beta}^*$ to the desired small neighbourhood of $\boldsymbol{\beta}^*$.

Remark 2. *Discussion of the measurement error* ε_n *when* $d_H^* < \infty$*:*

- *1. If* $\varepsilon_n = o(\sqrt{\frac{\log p}{n}})$*, then* $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 \leq O(\sqrt{\frac{\log p}{n}})$, $\mathrm{E}^*[\boldsymbol{X}^{*T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 \leq O(\frac{\log p}{n})$*;*
- *2. If* $\varepsilon_n = O(\sqrt{\frac{\log p}{n}})$*, then* $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 \leq O(1)$*, but* $\mathrm{E}^*[\boldsymbol{X}^{*T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 \leq O(\sqrt{\frac{\log p}{n}})$*;*

More typical examples for ε_n *are* $\frac{1}{n}$ *or even 0. Under the restricted condition* $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^* \in V(3.5, \frac{\varepsilon_n}{2}, H)$*, Case 2 tells us that if the order of fluctuations* ε_n *is sightly lower than the order of tuning parameter, Elastic-net with* $\lambda_2 \leq \frac{\lambda_1}{8B}$ *guarantees that the square prediction error is asymptotical being zero with a lower rate* $O(\sqrt{\frac{\log p}{n}})$*.*

3 Applications of oracles results

The previous sections above pave the way for the non-asymptotic or asymptotic results in the consecutive sections. In this section, the applications of oracles results are derived

from oracle inequalities about ℓ_1-estimation error, and we assume that the design matrix is fixed for simplicity.

3.1 Grouping effect from oracle inequality

Zou and Hastie (2005) shows that the Elastic-net has a grouping effect, which asserts that strongly correlated predictors tend to be in or out of the model together when the coefficients have the same sign. Zhou (2013) proves the grouping effect of the elastic-net estimates holds without the assumption of the sign. Yu (2010) derives asymptotical result of grouping effect for Elastic-net estimates of the Cox models. Based on oracle inequalities we put forward, we provide an asymptotical version of grouping effect inequality as $p, n \to \infty$ for fixed design case.

Theorem 4. *Under the assumption of Theorem 2 with $d_H^* < \infty$, suppose that the covarates (non-random) are standardized as*

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1, \ \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \ \ \textit{for} \ \ j = 1, 2, \cdots, p. \tag{22}$$

Denote $\rho_{kl} = \frac{1}{n}\sum_{i=1}^{n} x_{ik}x_{il}$ as the correlation coefficient. For any constant $E_s > 0$, with probability at least $1 - \frac{2}{p^{r-1}} - 2p^2 e^{-\frac{nt^2}{2[d_H^* C_{LB}(1+\zeta)L^2]^2}} - \frac{\sigma_n^2}{nE_s^2}$, we get

(i). $|\hat{\beta}_k - \hat{\beta}_l|^2 \le (1-\rho_{kl})[Ke^{2LM}O(1) + \frac{1}{\lambda_2^2}(E_s + \mu_s)]$;

(ii). If the asymptotic correlation between two random predictors are asymptotically up to 1, i.e. $\rho_{kl} = 1 - o(\lambda_2^2)$ with $\lambda_2^2 = O(\frac{\log p}{n}) \to 0$, thus

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

This grouping effect oracle inequality asserts that if ρ_{kl} tends to 1 then with high-probability elastic-net is able to select covariates $k, l \in \{1, 2, \ldots, p\}$ together. Combining with Lasso sparse estimation, the $\ell_1 + \ell_2$ penalty enables that strongly correlated predictors are inclined to be in or out simultaneously. In addition to the sparse estimation, intuitively, highly related covariates should have similar regression coefficients, but lasso cannot select them at one time.

3.2 Sign Consistency

Sign consistency is another criteria to show if one estimate is good. A few researchers have studied the sign consistency property of the Elastic-net. One condition for sign consistency is the Irrepresentable Condition (IC). Zhao and Yu (2006) explores the IC to enjoy the sign consistency for linear regression under Lasso penalty. Moreover, model selection consistency of Elastic-net are studied by Jia and Yu (2010), which follows the lines of Zhao and Yu (2006). Along the same line, for Elastic-net penalized Cox model, Yu (2010) investigates the selection consistency. Their basic idea is that KKT condition is the necessary and sufficient condition for global minimizer of target function. In similar fashion, we pay attention to selection consistency of Elastic-net penalized NBR model based some reasonable assumptions. It is interesting to see that under the bounded covariates assumption we do not need the IC which is assume in Yu (2010), Lv et al. (2018), we only rely on the assumptions in Theorem 2.

Uniform Signal Strength Condition.

$$\beta_* := \min_{j \in H} |\beta_j^*| \geq \frac{e^{2a_\tau}(\zeta + 1) d_H^* \lambda_1}{2C^2(\zeta, H)}$$

with $\lambda_1 = O(\sqrt{\frac{\log p}{n}}), B\lambda_2 = B_1\lambda_1$.

Assume $d_H^* < \infty$, Zhang (2014) pointed out that the selection consistency theory characteristically necessitates a *uniform signal strength condition* (or *beta-min condition*) that the smallest non-zero regression coefficients $\beta_* := \min\{|\beta_j| : j \in H\}$ should be greater in size than a thresholded level $O(\sqrt{\frac{\log p}{n}})$. When β_* is less than the level, the presence of weak signals cannot be detected by statistical inferences procedures.

Theorem 5. *Suppose that the Uniform Signal Strength Condition and assumptions of Theorem 2 hold. Let* $\lambda_1 = O(\sqrt{\frac{\log p}{n}}), d_H^* < \infty$. *Then, for* $\sqrt{\frac{\log p}{n}} = o(1)$ *and suitable tuning parameter* r *in Theorem 2, we have the sign consistency:*

$$\lim_{n,p\to\infty} P(\text{sign}\hat{\boldsymbol{\beta}} = \text{sign}\boldsymbol{\beta}^*) = 1. \tag{23}$$

3.3 Honest variable selection and detection of weak signals

As a special case of random design in Section 2.4, we focus on the fixed design in this section where the $\{\boldsymbol{X}_i\}_{i=1}^n$ are deterministic.

Recall that $\hat{H} := \{j : \hat{\beta}_j \neq 0\}$, so $\hat{H}$ is an estimator of the true variable set $H := \{j : \beta_j \neq 0\}$ (or the set of positives). Let δ_1, δ_2 be constants such that $P(\hat{H} \not\subset H) \leq \delta_1, P(H \not\subset \hat{H}) \leq \delta_2$, we have $P(H \neq \hat{H}) \leq P(\hat{H} \not\subset H) + P(H \not\subset \hat{H}) \leq \delta_1 + \delta_2$. Here, if we treat H as null hypothesis, the $P(\hat{H} \not\subset H)$ is often called false positive rate in the language of ROC curve (or Type I error in statistical hypothesis testing, the estimate is $\hat{H}$ but it makes decision $\hat{H} \subset H^c$); the $P(H \not\subset \hat{H})$ is often called false negative rate (or Type II error). Thus the probability of correct subset selection under

some random events W (the assumptions hold with probability $P(W)$) is

$$P(H = \hat{H}) \geq P(W) - \delta_1 - \delta_2. \tag{24}$$

From the ℓ_1-estimation error obtained in Theorem 3, we could easily bound the false negative rate $P(H \not\subset \hat{H})$ in Proposition 3. But the upper bound of false positive rate $P(\hat{H} \not\subset H)$ cannot be directly obtained, more addition assumptions on the covariates correlation is required.

Proposition 3. *Let $\delta \in (0,1)$ be a fixed number and the assumption of Theorem 3 is satisfied, and the weakest signal and strongest signal meet the condition: $B_0 := \frac{2.25^2\lambda_1 d_H^*}{ak+2\lambda_2} + (1+\frac{a}{\lambda_1})\varepsilon_n \leq \min_{j\in H}|\beta_j^*| \leq B$. If $p = \exp\{\frac{1}{A^2-1}\log\frac{2^{1-A^2}}{\delta}\}$ with $A > 1$, then*

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

It is worth noting here that the lower bound we derived may be too large in some setting. For example, if d_H^* is as large as $\lambda_1 d_H^* = O(1)$, and $\min_{j\in H}|\beta_j^*| \geq \frac{2.25^2 O(1)}{ak+2\lambda_2} =: D$ where D is also a moderate large constant compared to the strongest signal threshold B. Then we can only detect few parts of the whole signals. To deal with this problem, we will use a new approach (inspired by Section 3.1.2 in Bunea (2008)) to find constant-free weakest signal detection threshold which only relies on the tuning parameter λ_1. There are no free lunch for getting a desirable results in statistics. Under some mild condition on design matrix, we will illustrate that the lower bounds could be considerably sharpen.

First, we assume that the covariates are centered and standardized like (22). This crucial method of processing covariates is also employed similarly in studying the grouping effect in Section 3.1. Second, let $\rho_{kl} = \frac{1}{n}\sum_{i=1}^{n} X_{ik}X_{il}$, $k, l \in \{1, 2, \cdots, p\}$ be the correlation constants between covarates k and l. For constant $h \in (0,1)$, we pose the

Identifiable Condition: $\max_{k,l\in H,k\neq l}|\rho_{kl}| \leq \frac{h}{\theta d_H^*}$, $\frac{\theta}{n}\sum_{i=1}^{n} X_{ik}^2 = 1$.

This assumption of maximal correlation constant of two distinct covariates on the true set H, measures the dependence structure by a constant h in the whole predictor. The less h is, the more degree of separation is, and the easier to detect weak signals. Bunea (2008) explained the intuition that:" If the signal is very weak and the true variables are highly correlated with one another and with the rest, one cannot hope to recover the true model with high probability". Interestingly, the grouping effect in previous, says that the Elastic-net is able to simultaneously estimate highly correlated true variables, and this grouping effect is valid without the premise that the signal is enough strong. If both faint signals under the level of detection bounds, then the Elastic-net estimates are both zero, and grouping effect is also true.

Additionally, we require two technical conditions as we have to build some connections between $P(H \not\subset \hat{H}), P(\hat{H} \not\subset H)$ and the ℓ_1-estimation error in Theorem 3. Let a_i (b_i) is the intermediate point between $\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_i^T\boldsymbol{\beta}^*$ by the first order Taylor expansion of the function $f(t) = \frac{e^t}{\theta+e^t}$ ($g(t) = \frac{1}{\theta+e^t}$), and $L_1, L_2 \in [1,\infty)$. By (H.1)-(H.3), it leads to for all i

$$|a_i| \text{ or } |b_i| \leq |\boldsymbol{X}_i^{*T}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*| + |\boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*| \leq |\boldsymbol{X}_i^{*T}\hat{\boldsymbol{\beta}} - \boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*| + |\boldsymbol{X}_i^{T}\boldsymbol{\beta}^*| \leq L(M+B).$$

We pose some weighted correlation conditions (WCC in short):

Weighted Correlation Condition (1):

$$\sup_{\substack{k,j\in H,\\ |a_i|\leq L(M+B)}} \frac{1}{n}\left(|\sum_{i=1}^{n} X_{ij}X_{ik}\frac{\theta^2 e^{a_i}}{(\theta+e^{a_i})^2}| \vee |\sum_{i=1}^{n}\theta X_{ij}X_{ik}(1-\frac{\theta e^{a_i}}{(\theta+e^{a_i})^2})|\right) \leq \frac{hL_1}{d_H^*}.$$

Weighted Correlation Condition (2) holds with high probability:

$$P\left(\sup_{\substack{k,j\in H,\\ |b_i|\le L(M+B)}} |\frac{1}{n}\sum_{i=1}^{n}\frac{X_{ik}X_{ij}Y_i\cdot\theta^2 e^{b_i}}{(\theta+e^{b_i})^2}| \le \frac{hL_2}{d_H^*}\right) = 1-\varepsilon_{n,p},$$

where $\varepsilon_{n,p}$ is a constant satisfying $\lim_{n,p\to\infty}\varepsilon_{n,p}=0$.

By (H.1) and (H.2), a_i, b_i are uniformly bounded random variables and they are viewed as ignorable constant in asymptotic analysis, so do $\frac{\theta e^{a_i}}{(\theta+e^{a_i})^2}$ and $(1-\frac{\theta e^{a_i}}{(\theta+e^{a_i})^2})$. We can check WCC(2) by the similar approach in the concentration phenomenon for suprema of the multiplier empirical processes, see the proof of Lemma A.6. The conditions above can be obtained by make a linear transformation of the covariates, i.e. scaling the covariates. The WCC(1) is a technical condition which has been used by Bunea (2008) for the case of logistic regression. This assumption means that the maximum weighted-correlation version of $\rho_{kl}, (k\neq l)$ is less than $\frac{hL_1}{\theta d_H^*}$. However, NBR is more complex than logistic regression since its Hessian matrix dependents on random responses, thus WCC(2) should be assumed with high probability.

Together with above ingredients, we have the following constant-free weakest signal detection threshold for correct subset selection.

Theorem 6. *If assumptions in Theorem 3 hold with $\varepsilon_n = 0$, under the Identifiable Condition, WCC(1,2) with $h \le \frac{a+2\lambda_2}{20.25L_i+8a}\wedge\frac{1}{8}$ for $i=1,2$. Let $p=\exp\{\frac{1}{1-A^2}\log(2^{A^2-1}\delta)\}$,*

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

provided that the minimal signal condition $\min_{j\in H}|\beta_j^*| \ge 2\lambda_1$ is satisfied.

3.4 De-biased Elastic-net and confidence interval

Introduced by Zhang (2014), the de-biased Lasso was further studied in van de Geer et al. (2014) and Janková and van de Geer (2016) within some generalized linear models. Following the the de-biasing idea, we deal with the de-biased estimator $\hat{\boldsymbol{b}} =: \hat{\boldsymbol{\beta}} - \hat{\Theta}\dot{\ell}(\hat{\boldsymbol{\beta}})$, which is asymptotic normality based on the established oracle inequality in Section 2. Let $\hat{\boldsymbol{\beta}}$ be defined in optimization problem (1). Let $\hat{\Theta}$ be an approximated estimator of the inverse of the Hessian $-\ddot{\ell}(\boldsymbol{\beta}^*)$(for example, the CLIME or nodewise Lasso estimator for estimated Hessian matrix). If $\dot{\ell}(\hat{\boldsymbol{\beta}})$ is continuously differentiable, by Taylor's expansion of vector-valued functions, we have

$$\begin{aligned}
\dot{\ell}(\boldsymbol{\beta}^*) &= \dot{\ell}(\hat{\boldsymbol{\beta}}) - \ddot{\ell}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - r(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2) \\
&= \ddot{\ell}(\boldsymbol{\beta}^*)[\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} - \ddot{\ell}(\boldsymbol{\beta}^*)^{-1}\dot{\ell}(\hat{\boldsymbol{\beta}})] - r(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2) \\
&= \ddot{\ell}(\boldsymbol{\beta}^*)[\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} + \hat{\Theta}\dot{\ell}(\hat{\boldsymbol{\beta}})] - \ddot{\ell}(\boldsymbol{\beta}^*)[\ddot{\ell}(\boldsymbol{\beta}^*)^{-1} + \hat{\Theta}]\dot{\ell}(\hat{\boldsymbol{\beta}}) - r(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2) \\
&=: \ddot{\ell}(\boldsymbol{\beta}^*)[\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} + \hat{\Theta}\dot{\ell}(\hat{\boldsymbol{\beta}})] + R_n
\end{aligned}$$

where $r(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2) = o_p(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2)$ is a vector-valued function.

Operate $\sqrt{n}\hat{\Theta}$ in the equation above, if $\sqrt{n}R_n = o_p(1)$, then

$$\sqrt{n}(\hat{\boldsymbol{b}} - \boldsymbol{\beta}^*) \approx \hat{\boldsymbol{\Theta}}[\sqrt{n}R_n - \sqrt{n}\dot{\ell}(\boldsymbol{\beta}^*)] \xrightarrow{d} N(0, \hat{\boldsymbol{\Theta}}\boldsymbol{\Sigma}\hat{\boldsymbol{\Theta}}^T)$$

where the notation $\approx$ means the asymptotic equivalence under some regular conditions. Here $\boldsymbol{\Sigma}$ is asymptotic variance of $\sqrt{n}\dot{\ell}(\boldsymbol{\beta}^*)$, where $\mathrm{Var}\dot{\ell}(\boldsymbol{\beta}^*) = \frac{1}{n}\sum_{i=1}^{n} \frac{\theta e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}}{\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}}\boldsymbol{X}_i\boldsymbol{X}_i^T$. We could plug in a consistent estimator for $\boldsymbol{\Sigma}$ in high-dimensional case.

The asymptotic a confidence level of $1-\alpha$ for β_j^* is then given by

$$\big[\hat{b}_j - c(\alpha, n, \sigma), \hat{b}_j + c(\alpha, n, \sigma)\big], \quad c(\alpha, n, \sigma) := \Phi^{-1}(1-\alpha/2)\sqrt{(\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Theta}}^T)_{j,j}/n},$$

where $\Phi(\cdot)$ denotes the c.d.f. of $N(0,1)$.

By KKT conditions in Lemma 1, the de-biased Elastic-net estimator is written as

$$\hat{\boldsymbol{b}} = \hat{\boldsymbol{\beta}} - \hat{\Theta}\dot{\ell}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}(\mathrm{I}_p - 2\lambda_2\hat{\Theta}) - \hat{\Theta}\lambda_1 \mathrm{sign}(\hat{\boldsymbol{\beta}}).$$

The theoretical analysis of de-biased Elastic-net estimator (includes precision matrix estimation, confidence interval and hypothesis testing) is beyond the length and scope of the this paper, and the interested reader could refer to the proofs in Janková and van de Geer (2016) for some technical details.

A simulation study for the de-biased Elastic-net is presented in Appendix A, which illustrates that de-biased Elastic-net has less bias than de-biased Lasso. When do simulation, it is of paramount importance to estimate the nuisance parameter θ and the estimation of the inverse of the Hessian.

4 Conclusions and Discussions

In this technical paper, we thoroughly study sparse high-dimensional negative binomial regression problems via several consistency results such as prediction or ℓ_q-estimation error bounds. Negative binomial regressions are widely used in modeling count data. We show that under a few conditions, the Elastic-net estimator has oracle properties, which means that when sample size is large enough, our sparse estimator is very close to the true parameter if the tuning parameters are properly chosen. We also show the sign consistency property under beta-min condition. We discuss the detection of weak signals, and give a constant-free weakest signal threshold for correct subset selection under some correlation conditions of covariates. Asymptotic normality of the de-biased

Elastic-net estimator is also discussed and the further study is beyond the scope of this paper. These results provide theoretical understanding of the proposed sparse estimator and provide practical guidance for the use of the Elastic-net estimator.

It should be noticed that oracles inequalities in Section 2.4 and Section 3 could be extended to many ℓ_1 or $\ell_1 + \ell_2$ regularized M-estimation regression with the corresponding empirical process (17) has stochastic Lipschitz properties which is presented in Proposition 1. For example, the analysis of stochastic Lipschitz properties of the average negative log-likelihood empirical process can be employed to Elastic-net or Lasso penalized COM-Poisson regression (see Sellers and Shmueli (2008)).

As we can see in the simulation, it demonstrates that the two-step estimation of $\hat{\theta}$ is not behave well. Like the misspecified models in Example 5.25 of van der Vaart (1998), the θ which is nuisance parameter, is not an important estimate in the consistency results. It will be interesting and important to find a better estimator of dispersion parameter in the further research, since θ is a crucial quantization in constructing confidence interval.

Acknowledgements

In writing the manuscript had the kind assistance of Xiaoxu Wu, to whom warm thanks are due. The authors would like to thank the anonymous referees for their valuable comments which greatly improve the quality of our manuscript. The authors also thank Prof. Cun-Hui Zhang, Prof. Fang Yao and Dr. Sheng Fu for helpful discussions about this revision. This work is partially supported by National Science Foundation of China

(11571021).

Supplementary Materials

The supplementary material includes simulations, detailed proofs of main theorems, lemmas and propositions.

References

Balabdaoui, F., Jankowski, H., Rufibach, K., & Pavlides, M. (2013). Asymptotics of the discrete log-concave maximum likelihood estimator and related applications. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(4), 769-790.

Bickel, P. J., Ritov, Y. A., Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics, 1705-1732.

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.

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

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.

Cameron, A. C., Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. Journal of econometrics, 46(3), 347-364.

Chi, Z. (2010). A local stochastic Lipschitz condition with application to Lasso for high dimensional generalized linear models. arXiv preprint arXiv:1009.1052.

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

Huang, J., Sun, T., Ying, Z., Yu, Y., Zhang, C. H. (2013). Oracle inequalities for the lasso in the Cox model. Annals of statistics, 41(3), 1142-1165.

Janková, J., van de Geer, S. (2016). Confidence regions for high-dimensional generalized linear models under sparsity. arXiv:1610.01353.

Jia, J., Yu, B. (2010). On model selection consistency of the Elastic Net when $p \gg n$. Statistica Sinica, 595-611.

Kakade, S., Shamir, O., Sindharan, K., Tewari, A. (2010). Learning exponential families in high-dimensions: Strong convexity and sparsity. In International Conference on Artificial Intelligence and Statistics (pp. 381-388).

Lv, S., You, M., Lin, H., Lian, H., & Huang, J. (2018). On the sign consistency of the Lasso for the high-dimensional Cox model. Journal of Multivariate Analysis, 167, 79-96.

Massaro, T. J. (2016). Variable selection via penalized regression and the genetic algorithm using information complexity, with applications for high-dimensional-omics data. PhD Dissertations, University of Tennessee.

Nielsen, F., & Nock, R. (2009). Sided and symmetrized Bregman centroids. IEEE transactions on Information Theory, 55(6), 2882-2904.

Qiu, Y., Chen, S. X., & Nettleton, D. (2018). Detecting rare and faint signals via thresholding maximum likelihood estimators. The Annals of Statistics, 46(2), 895-923.

Rauschenberger, A., Jonker, M. A., van de Wiel, M. A., & Menezes, R. X. (2016). Testing for association between RNA-Seq and high-dimensional data. BMC bioinformatics, 17(1), 118.

Sørensen, ø., Hellton, K. H., Frigessi, A., & Thoresen, M. (2018). Covariate selection in high-dimensional generalized linear models with measurement error. Journal of Computational and Graphical Statistics, 27(4), 739-749.

Sellers, K. F., Shmueli, G. (2008). A flexible regression model for count data. The Annals of applied statistics, 4(2), 943-961.

Tutz, G. (2011). Regression for categorical data. Cambridge University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

van de Geer, S. A. (2007). The deterministic lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. The Annals of Statistics, 36(2), 614-645.

van de Geer, S. A., Bühlmann, P., Ritov, Y. A., Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics, 42(3), 1166-1202.

van der Vaart, A. W. (1998). Asymptotic statistics (Vol. 3). Cambridge university press.

Wang Z, Ma S, Zappitelli M, et al. Penalized count data regression with application to hospital stay after pediatric cardiac surgery. Statistical methods in medical research, 2016, 25(6): 2685-2703.

Weißbach, R., & Radloff, L. (2019). Consistency for the negative binomial regression with fixed covariate. Metrika, 1-15.

Ye, F., Zhang, C. H. (2010). Rate Minimaxity of the Lasso and Dantzig Selector for the lq Loss in lr Balls. Journal of Machine Learning Research, 11(Dec), 3519-3540.

Yu, Y. (2010). High-dimensional Variable Selection in Cox Model with Generalized Lasso-type Convex Penalty. https://people.maths.bris.ac.uk/~yy15165/index_files/Cox_generalized_convex.pdf

Yu, Y., Bradic, J., & Samworth, R.J. (2020). Confidence intervals for high-dimensional Cox models. Statistica Sinica. https://doi.org/10.5705/ss.202018.0247

Zhang, C. H., Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1), 217-242.

Zhang, H., Liu, Y., Li, B. (2014). Notes on discrete compound Poisson model with applications to risk theory. Insurance: Mathematics and Economics, 59, 325-336.

Zhang, H., Wu, X. (2019). Compound Poisson Point Processes, Concentration and Oracle Inequalities, Journal of Inequalities and Applications, 2019: 312.

Zhao, P., Yu, B. (2006). On model selection consistency of Lasso. Journal of Machine learning research, 7(Nov), 2541-2563.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

Zhou, D. X. (2013). On grouping effect of elastic net. Statistics & Probability Letters, 83(9), 2108-2112.

School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, 100871, China

E-mail: (zhanghuiming@pku.edu.cn)

School of Public Health and Center for Statistical Science, Peking University, Beijing, 100871, China

E-mail: (jzjia@pku.edu.cn)

Supplementary Materials:

Elastic-net Regularized High-dimensional Negative Binomial Regression

A Main Proofs

A.1 Proof of Theorem 1

With the aim of deriving the targeted oracle inequalities (10), we first prove the lower bound for symmetric Bregman divergence $D_g^s(\boldsymbol{\beta}+\boldsymbol{\delta},\boldsymbol{\beta})$ with $g=0$.

Lemma A.4. *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). 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

$$\frac{\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\theta+e^{\boldsymbol{X}_i^T(\boldsymbol{\beta}+\boldsymbol{\delta})}} \geq e^{-L\|\boldsymbol{\delta}\|_1}. \tag{A.25}$$

To show the (A.25), 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 (A.25), it implies by the expression of $\ddot{\ell}(\boldsymbol{\beta})$ that

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

□

Next, we give the proof of Theorem 1 based on Lemma A.4.

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 (5), we have

$$\boldsymbol{b}^T[\dot{\ell}(\boldsymbol{\beta}^* + \boldsymbol{b}x) - \dot{\ell}(\boldsymbol{\beta}^*)] \leq \frac{2\ \zeta\lambda_1}{\zeta+1}\|\boldsymbol{b}_H\|_1 - \frac{2\lambda_1}{\zeta+1}\|\boldsymbol{b}_{H^C}\|_1 \leq \frac{2\ \zeta\lambda_1}{\zeta+1}\|\boldsymbol{b}_H\|_1 \tag{A.26}$$

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

By the Lemma A.4, 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

$$\boldsymbol{b}^T[\dot{\ell}_n(\boldsymbol{\beta}^* + \boldsymbol{b}x) - \dot{\ell}_n(\boldsymbol{\beta}^*)] \geq xe^{-2Lx}\boldsymbol{b}^T\ddot{\ell}_n(\boldsymbol{\beta})\boldsymbol{b}. \tag{A.27}$$

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

$$\begin{aligned}
Lxe^{-2Lx}[C(\zeta,H)]^2\|\boldsymbol{b}_H\|_1^2/d_H^* &\leq Lxe^{-2Lx}\boldsymbol{b}^T\ddot{\ell}_n(\boldsymbol{\beta})\boldsymbol{b} \\
\text{(by (A.27))} \quad &\leq L\boldsymbol{b}^T[\dot{\ell}_n(\boldsymbol{\beta}^* + \boldsymbol{b}x) - \dot{\ell}_n(\boldsymbol{\beta}^*)] \\
\text{(by (A.26))} \quad &\leq L(\frac{2\ \zeta\lambda_1}{\zeta+1}\|\boldsymbol{b}_H\|_1 - \frac{2\lambda_1}{\zeta+1}\|\boldsymbol{b}_{H^C}\|_1) \\
&= L[\frac{2\ \zeta\lambda_1}{\zeta+1}\|\boldsymbol{b}_H\|_1 - \frac{2\lambda_1}{\zeta+1}(1-\|\boldsymbol{b}_H\|_1)] \\
&\leq L(2\lambda_1\|\boldsymbol{b}_H\|_1 - \frac{2\lambda_1}{\zeta+1}) \leq \frac{L(\zeta+1)\|\boldsymbol{b}_H\|_1^2\lambda_1}{2}.
\end{aligned}$$

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

$$Lxe^{-2Lx} \leq \frac{L(\zeta+1)d_H^*\lambda_1}{2[C(\zeta,H)]^2} =: \tau \tag{A.28}$$

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 (A.28) 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 (A.28) again, it implies $L\tilde{x}e^{-2L\tilde{x}} \leq \tau$. Then, for $\forall x \in [0, \tilde{x}]$, we have

$$\|\tilde{\boldsymbol{\beta}}\|_1 \leq \tilde{x} \leq \frac{a_\tau}{L} = \frac{e^{2a_\tau}\tau}{L} = \frac{e^{2a_\tau}(\zeta+1)d_H^*\lambda_1}{2[C(\zeta,H)]^2} \tag{A.29}$$

where the last equality is by the definition of τ.

Similarly, by the definition of weak CIF, we have

$$\begin{aligned} xe^{-2Lx} &\leq \frac{xe^{-2Lx}\boldsymbol{b}^T\ddot{\ell}_n(\boldsymbol{\beta})\boldsymbol{b}}{C_q(\zeta,H)(\|\boldsymbol{b}_H\|_1/(d_H^{*\,1/q})\|\boldsymbol{b}\|_q} \leq \frac{\boldsymbol{b}^T[\dot{\ell}_n(\boldsymbol{\beta}^*+\boldsymbol{b}x) - \dot{\ell}_n(\boldsymbol{\beta}^*)]}{C_q(\zeta,H)(\|\boldsymbol{b}_H\|_1/(d_H^{*\,1/q})\|\boldsymbol{b}\|_q} \\ &\quad \text{(by (A.26))} \quad \leq \frac{2\zeta d_H^{*\,1/q}\lambda_1}{(\zeta+1)\,C_q(\zeta,H)\|\boldsymbol{b}\|_q}. \end{aligned}$$

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 (A.29). □

A.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 A.5 (Lemma 6.1 in Rigollet (2012)). *Let $\{Y_i\}_{i=1}^n$ be a sequence of random variables whose distribution belongs to canonical exponential family with $f(y_i;\theta_i) = c(y_i)\exp(y_i\theta_i - \psi(\theta_i))$. We assume uniformly bounded variances condition: there exist compact set Ω and some constant C_ψ^2 such that $\sup_{\theta_i\in\Omega}\ddot{\psi}(\theta_i) \le C_\psi^2$ 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*

$$P\{|S_n^w - E(S_n^w)| > t\} \le 2\exp\{-\frac{t^2}{2C_\psi^2\|\boldsymbol{w}\|_2^2}\}. \tag{A.30}$$

Moreover, we have $\mathrm{E}|S_n^w - \mathrm{E}S_n^w|^k \le D_{k,C}\,\|w\|_2^k$ 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

$$\sup_{\theta_i\in\Omega}\ddot{\psi}(\theta_i) = \sup_{\mu_i}(\mu_i + \frac{\mu_i^2}{\theta}) = \sup_{|\boldsymbol{X}_i^T\boldsymbol{\beta}^*|\le LB}(e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*} + \frac{e^{2\boldsymbol{X}_i^T\boldsymbol{\beta}^*}}{\theta}) = e^{LB} + \frac{e^{2LB}}{\theta} := C_{LB}^2. \tag{A.31}$$

Now, we can apply concentration inequality Lemma A.5 to go on the proof. The first step is to evaluate the event $\mathcal{K} := \left\{z^* \le \frac{\zeta-1}{\zeta+1}\lambda_1\right\}$ from the inequality in (11). By assuming $B\lambda_2 := B_1\lambda_1$, we have

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

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

Therefore, by using Lemma A.5, we have

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

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,

$$P(z^*\geq\frac{\zeta-1}{\zeta+1}\lambda_1)\leq 2p\exp\{-\frac{C^2_{\xi,B_1}\lambda_1^2 n}{2C^2_{LB}L^2}\}=:\frac{2}{p^{r-1}},\quad r>1. \tag{A.32}$$

The expression of tuning parameter λ_1 is solved by the equality in (A.32), 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,

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

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

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

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

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

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

where the last inequality is from (6).

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 A.5. Then,

$$P\{|S_n^c(\boldsymbol{b}, Y) - \mathrm{E}S_n^c(\boldsymbol{b}, Y)| \geq t, \ \forall \boldsymbol{b} \in \Lambda\} \leq P\{\max_{j,k} |(\ddot{\ell}_n(\beta^*) - \mathrm{E}\ddot{\ell}_n(\beta^*))_{j,k}| \leq t/d_H^*(1+\zeta)^2\}$$
$$\leq p^2 P\{|(\ddot{\ell}_n(\beta^*) - \mathrm{E}\ddot{\ell}_n(\beta^*))_{j,k}| \leq t/d_H^*(1+\zeta)^2\}$$
$$\leq 2p^2 \exp\{-\frac{nt^2}{2C_{LB}^2[d_H^*(1+\zeta)L^2]^2}\} \qquad \text{(A.33)}$$

where the last inequality is by using Lemma A.5 with $||\boldsymbol{w}||_2^2 \leq L^4/n$.

We 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 (A.33) is free of $\boldsymbol{b}$, thus by the (A.33) for $\Lambda \ni \tilde{\boldsymbol{b}}_c$ we have

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

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 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 difference of $S_n^w(\boldsymbol{b}, Y)$ and $\mathrm{E}S_n^w(\boldsymbol{b}, Y)$ is bound by

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

where the second last inequality is from (6).

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

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

A.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

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

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

$$\mathbb{P}_n l(\hat{\boldsymbol{\beta}}) + \lambda_1\|\hat{\boldsymbol{\beta}}\|_1 + \lambda_2\|\hat{\boldsymbol{\beta}}\|_2^2 \leq \mathbb{P}_n l(\boldsymbol{\beta}^*) + \lambda_1\|\boldsymbol{\beta}^*\|_1 + \lambda_2\|\boldsymbol{\beta}^*\|_2^2. \tag{A.34}$$

By convexity of the optimization function (1), combined with (A.34), 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

$$\mathbb{P}(l(\tilde{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^*)) + \lambda_1|\tilde{\boldsymbol{\beta}}\| + \lambda_2|\tilde{\boldsymbol{\beta}}\|_2^2 \leq \frac{\lambda_1}{2}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \lambda_1\frac{\varepsilon_n}{4} + \lambda_1\|\boldsymbol{\beta}^*\|_1 + \lambda_2\|\boldsymbol{\beta}^*\|_2^2. \tag{A.35}$$

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

$$\begin{aligned}\lambda_1\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 &\leq \lambda_1\|\tilde{\boldsymbol{\beta}}\|_1 + \lambda_1\|\boldsymbol{\beta}^*\|_1 \leq [\mathbb{P}(l(\tilde{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^*)) + \lambda_1\|\tilde{\boldsymbol{\beta}}\|_1] + \lambda_1\|\boldsymbol{\beta}^*\|_1 \\ [\text{by (A.35)}] \quad &\leq \frac{\lambda_1}{2}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \frac{\lambda_1\varepsilon_n}{4} + 2\lambda_1\|\boldsymbol{\beta}^*\|_1 + (\lambda_2\|\boldsymbol{\beta}^*\|_2^2 - \lambda_2\|\tilde{\boldsymbol{\beta}}\|_2^2). \end{aligned} \tag{A.36}$$

From the assumption that $8B\lambda_2 + 4M = \lambda_1$ 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, (A.36) 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 $\frac{M\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1}{M + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1} \leq \frac{M}{2}$. It derives $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq M$. □

A.4 Proof of Proposition 1 and 2

We deduce Proposition 2 by showing the following key lemma.

Lemma A.6. *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 A.6 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 A.6 and and Proposition 1 used in the proof of Proposition 2.

A.4.1 Proof of Lemma A.6

The proof rests on the following lemma.

Lemma A.7. *Given $M > 0$, if $A \geq 1$, under (H.1), we have*

$$P(Z_M(\boldsymbol{\beta}^*) \geq \frac{5\theta AML}{(M+\varepsilon_n)}\sqrt{\frac{2\log(2p)}{n}}) \leq (2p)^{-A^2}. \tag{A.37}$$

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 A.7 (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

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

with $\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}$ ($\boldsymbol{X}'^T_i\tilde{\boldsymbol{\beta}}$) being an intermediate point between $\boldsymbol{X}_i^T\boldsymbol{\beta}$ ($\boldsymbol{X}'^T_i\boldsymbol{\beta}$) 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 $|f(x)|-\sup_x|g(x)|\le|f(x)-g(x)|$ (and take suprema over x 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

$$P(Z_M(\boldsymbol{\beta}^*)-\mathrm{E}Z_M(\boldsymbol{\beta}^*)\ge\lambda)\le(2p)^{-A^2}. \tag{A.38}$$

The next step is to estimate the sharper upper bounds of $\mathrm{E}Z_M(\boldsymbol{\beta}^*)$ by the symmetrization theorem and the contraction theorem below. It can be found in van der Vaart (1998), Bühlmann and van de Geer (2011).

Lemma A.8 (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*

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n[f(X_i)-\mathrm{E}\{f(X_i)\}]\right|\right]\le2\mathrm{E}\left[\mathrm{E}_\epsilon\left\{\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\epsilon_if(X_i)\right|\right\}\right].$$

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 A.9 (Contraction Theorem). *Let $x_1, ..., x_n$ be the non-random elements of $\mathcal{X}$ and $\varepsilon_1, ..., \varepsilon_n$ be Rademacher sequence. 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 f and h in $\mathcal{F}$, 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 $(\mathbb{P}_n - \mathbb{P})\{l_2(\boldsymbol{\beta}^*) - l_2(\boldsymbol{\beta})\} = \mathbb{P}_n\{l_2(\boldsymbol{\beta}^*) - l_2(\boldsymbol{\beta})\} - \mathrm{E}\{l_2(\boldsymbol{\beta}^*) - l_2(\boldsymbol{\beta})\}$, 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 A.9 with $\mathcal{F} = \mathbb{R}^p$. Let $f(x_i) = \frac{x_i^T\boldsymbol{\beta}}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1+\varepsilon_n}$ 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

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

where $\tilde{t} \in [-LB, LB]$ is an intermediate point between t and s given by applying Lagrange mean value theorem for function $g_i(t)$.

Via the symmetrization theorem and the contraction theorem we have

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

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

From Proposition 4, 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

$$\mathrm{E}Z_M(\boldsymbol{\beta}^*)\le\frac{4\theta ML}{(M+\varepsilon_n)}\sqrt{\frac{2\log 2p}{n}}\le\frac{4\theta AML}{(M+\varepsilon_n)}\sqrt{\frac{2\log 2p}{n}}. \tag{A.39}$$

So we can conclude from (A.38) and (A.39) that

$$P(Z_M(\boldsymbol{\beta}^*)\ge\frac{5\theta AML}{(M+\varepsilon_n)}\sqrt{\frac{\log 2p}{n}})\le P(Z_M(\boldsymbol{\beta}^*)\ge\lambda+\mathrm{E}Z_M(\boldsymbol{\beta}^*))\le(2p)^{-A^2}. \tag{A.40}$$

Finally, we complete the proof of Lemma A.6 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 $Z_M(\boldsymbol{\beta}^*)$.

A.4.2 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

$$\begin{aligned}(\mathbb{P}_n - \mathbb{P})(l_1(\boldsymbol{\beta}^*) - l_1(\hat{\boldsymbol{\beta}})) &= \frac{-1}{n}\sum_{i=1}^n (Y_i - \mathrm{E}Y_i)\boldsymbol{X}_i^T[(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) - \log(\frac{\theta+\exp\{\mathbf{X}_i^T\boldsymbol{\beta}^*\}}{\theta+\exp\{\mathbf{X}_i^T\hat{\boldsymbol{\beta}}\}})] \\ &= \frac{-1}{n}\sum_{i=1}^n (Y_i - \mathrm{E}Y_i)\boldsymbol{X}_i^T[(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) - \frac{\exp\{\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}\}\boldsymbol{X}_i^T(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})}{\theta+\exp\{\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}\}}] \\ &= \frac{-1}{n}\sum_{i=1}^n (Y_i - \mathrm{E}Y_i)\frac{\theta\boldsymbol{X}_i^T(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})}{\theta+\exp\{\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}\}}. \qquad \text{(A.41)}\end{aligned}$$

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

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

Note that the random sum in (A.41) 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. The Theorem 3.16 in Wainwright (2019) is for continuous case], it implies for $t > 0$

$$P(f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) - \mathrm{E}_{\boldsymbol{Y}} f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) \geq t | \mathbf{X}) \leq \exp\{-\frac{\gamma n t^2}{4|||\mathbf{X}|||_\infty^2}\}. \tag{A.42}$$

provided that (H.4) holds.

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

$$P(f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) - \mathrm{E} f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) \geq t) \leq \mathrm{E} \exp\{-\frac{\gamma n t^2}{4|||\mathbf{X}|||_\infty^2}\} \leq \exp\{-\frac{\gamma n t^2}{4L^2}\}. \tag{A.43}$$

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 A.10 (Symmetrization Theorem with difference functions)**.** *Let $\varepsilon_1, ..., \varepsilon_n$ be a Rademacher sequence with uniform distribution on $\{-1, 1\}$, independent of $X_1, ..., X_n$*

and $g_i \in \mathcal{G}_i$. *Then we have*

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

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 $\{X_i'\}_{i=1}^n$ be an independent copy of $\{X_i\}_{i=1}^n$. The E' denote the exportation w.r.t. $(X_i')_{i=1}^n$, then let $\mathcal{F}_n' = \sigma(X_1', \cdots, X_n')$. So

$$\begin{aligned}
&\mathrm{E}\left(\sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\sum_{i=1}^{n}\left[g_i(X_i)-\mathrm{E}\left\{g_i(X_i)\right\}\right]\right|\right)\\
&= \mathrm{E}\left(\sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\mathrm{E}'\sum_{i=1}^{n}[g_i(X_t)-g_i(X_i')]|\mathcal{F}_n'\right|\right)\\
&\le \mathrm{E}\left(\sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\mathrm{E}'\left|\sum_{i=1}^{n}[g_i(X_t)-g_i(X_i')]\right||\mathcal{F}_n'\right) \text{ (Jensen's inequality of absolute function)}\\
&\le \mathrm{E}\left(\mathrm{E}'\sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\sum_{i=1}^{n}[g_i(X_t)-g_i(X_i')]\right||\mathcal{F}_n'\right) \text{ (Jensen's inequality of max function)}\\
&= \mathrm{E}\left(\sup_{f_1,\cdots,f_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\sum_{i=1}^{n}[g_i(X_t)-g_i(X_i')]\right|\right),\\
&= \mathrm{E}\left(\sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\sum_{i=1}^{n}\varepsilon_i\left(g_i(X_i)-g_i(X_i')\right)\right|\right) \le 2\mathrm{E}\left[\mathrm{E}_\epsilon \sup_{g_1,\cdots,g_n\in\mathcal{G}_1,\cdots,\mathcal{G}_n}\left|\sum_{i=1}^{n}\epsilon_i g_i(X_i)\right|\right],
\end{aligned}$$

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

Then the symmetrization theorem implies

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

Now we are going to use a maximal inequality mentioned by Blazere et al. (2014), p2316. The proof is a consequence of Hoeffding's lemma (see Lemma 14.10 in Bühlmann and van de Geer (2011)) and we will give a proof in end of Appendix B.

Proposition 4 (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 $i = 1, ..., n$*

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

Then

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

By Proposition 4, with $\mathrm{E}[\epsilon_i Y_i X_{ij}|\mathbf{X}, \boldsymbol{Y}] = 0$ we get

$$\begin{aligned}
\frac{2}{n}\mathrm{E}\left(\mathrm{E}_\epsilon \max_{1\le j\le p}\left|\sum_{i=1}^{n}\epsilon_i Y_i X_{ij}\right|\right) &\le \frac{2}{n}\sqrt{2\log 2p}\mathrm{E}\left(\sqrt{\sum_{i=1}^{n} Y_i^2}|\mathbf{X}\right) \\
\text{[By Jensen's inequality]} &\le \frac{2L}{n}\sqrt{2\log 2p}\sqrt{\mathrm{E}\left(\sum_{i=1}^{n} Y_i^2|\mathbf{X}\right)} \\
&\le \frac{2L}{n}\sqrt{2\log 2p}\sqrt{n\tilde{C}_{LB}^2} = 2L\tilde{C}_{LB}\sqrt{\frac{2\log 2p}{n}}
\end{aligned}$$

where the last inequality stems from

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

using (H.1) and (H.2).

Thus, we get

$$\mathrm{E}f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) \le 2L\tilde{C}_{LB}\sqrt{\frac{2\log 2p}{n}}. \tag{A.44}$$

In equation (A.43), if we choose $t = AL\sqrt{2\gamma}\sqrt{\frac{2\log 2p}{n}}$ such that

$$P(f_n(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\beta}^*) \ge t + 2L\tilde{C}_{LB}\sqrt{\frac{2\log 2p}{n}}) \le \exp\{-\frac{\gamma n t^2}{4L^2}\} = (2p)^{-A^2}. \tag{A.45}$$

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 (A.45)

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

In (A.41), 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

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

A.5 Proofs of big Theorem 3.

The proof techniques follow the guidelines in Wegkamp (2007), Bunea (2008).

A.5.1 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

$$\mathbb{P}_n l(\hat{\boldsymbol{\beta}})+\lambda_1\sum_{j=1}^{p}|\hat{\beta}_j|+\lambda_2\sum_{j=1}^{p}|\hat{\beta}_j|^2\leq\mathbb{P}_n l(\boldsymbol{\beta}^*)+\lambda_1\sum_{j=1}^{p}|\beta_j^*|+\lambda_2\sum_{j=1}^{p}|\beta_j^*|^2. \tag{A.46}$$

So we obtain

$$\mathbb{P}(l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*))\leq(\mathbb{P}_n-\mathbb{P})(l(\boldsymbol{\beta}^*)-l(\hat{\boldsymbol{\beta}}))+\lambda_1\sum_{j=1}^{p}(|\beta_j^*|-|\hat{\beta}_j|)+\lambda_2\sum_{j=1}^{p}(|\beta_j^*|^2-|\hat{\beta}_j|^2). \tag{A.47}$$

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 18 implies.

$$\begin{aligned}&(\mathbb{P}_n-\mathbb{P})(l(\boldsymbol{\beta}^*)-l(\hat{\boldsymbol{\beta}}))=(\mathbb{P}_n-\mathbb{P})(l_1(\boldsymbol{\beta}^*)-l_1(\hat{\boldsymbol{\beta}}))+(\mathbb{P}_n-\mathbb{P})(l_2(\boldsymbol{\beta}^*)-l_2(\hat{\boldsymbol{\beta}}_n))\\&\leq\frac{\lambda_1}{4}\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|+\frac{\lambda_1}{4}(\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|+\varepsilon_n)=\frac{\lambda_1}{2}\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|+\frac{\lambda_1}{4}\varepsilon_n.\end{aligned} \tag{A.48}$$

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 (A.47), and combining with the inequality (A.48), it gives

$$\begin{aligned}&\frac{\lambda_1}{2}\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|+(\mathbb{P}l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*))+\lambda_2\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2\\&\leq\lambda_1\sum_{j=1}^{p}(|\hat{\beta}_j-\beta_j^*|+|\beta_j^*|-|\hat{\beta}_j|)+\frac{\lambda_1\varepsilon_n}{4}+\lambda_2(|\beta^*|_2^2-|\hat{\beta}|_2^2)+\lambda_2\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2.\end{aligned}\tag{A.49}$$

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 (A.49) 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 (A.49) is rewritten as

$$\begin{aligned}&\frac{\lambda_1}{2}\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|+\mathbb{P}(l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*))+\lambda_2\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|^2\\&\leq 2\lambda_1\sum_{j\in H}|\hat{\beta}_j-\beta_j^*|+\frac{\lambda_1\varepsilon_n}{4}+\frac{1}{4}\lambda_1\sum_{j\in H}|\beta_j^*-\hat{\beta}_j|.\end{aligned}\tag{A.50}$$

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 $\mathcal{A}\cap\mathcal{B}$.

A.5.2 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 5 (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)

$$|\boldsymbol{X}_i^{*T}\tilde{\boldsymbol{\beta}}| \leq |\boldsymbol{X}_i^{*T}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*| + |\boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*| \leq |\boldsymbol{X}_i^{*T}\hat{\boldsymbol{\beta}} - \boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*| + |\boldsymbol{X}_i^T\boldsymbol{\beta}^*| \leq L(M+B). \tag{A.51}$$

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

From Propositon 5 and (A.50) we deduce that

$$\lambda_1 \sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| + a\mathrm{E}^*[\boldsymbol{X}^{*T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 + 2\lambda_2 \sum_{j\in H} |\hat{\beta}_j - \beta_j^*|^2 \leq 4.5\lambda_1 \sum_{j\in H} |\hat{\beta}_j - \beta_j^*| + \frac{\lambda_1 \varepsilon_n}{2}. \tag{A.52}$$

A.5.3 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 (A.52),

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

By using Cauchy-Schwarz inequality, we get

$$\lambda_1 \sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| + (ak + 2\lambda_2) \sum_{j\in H} |\hat{\beta}_j - \beta_j^*|^2 \leq 4.5\lambda_1 \sqrt{d_H^*} \sqrt{\sum_{j\in H} |\hat{\beta}_j - \beta_j^*|^2} + \frac{\varepsilon_n(\lambda_1 + a)}{2}. \tag{A.53}$$

Apply the elementary inequality $2xy \le Tx^2 + y^2/T$ to (A.53) for all $t > 0$, it leads to

$$\lambda_1 \sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| + (a_n k + 2\lambda_2) \sum_{j \in H} |\hat{\beta}_j - \beta_j^*|^2 \le 2.25^2 T \lambda_1^2 d_H^* + \frac{1}{T} \sum_{j \in H} |\hat{\beta}_j - \beta_j^*|^2 + \frac{\varepsilon_n(\lambda_1 + a)}{2}. \tag{A.54}$$

We choice $T = \frac{1}{a_n k + 2\lambda_2}$ in (A.54), we obtain

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

For the square prediction error, we deduce from (A.52) by dropping the term $2\lambda_2 \sum_{j \in H} |\hat{\beta}_j - \beta_j^*|^2$

$$\lambda_1 \sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| + a\mathrm{E}^*[\boldsymbol{X}^{*T}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})]^2 \le 4.5\lambda_1 (\sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| - \sum_{j \in H^c} |\hat{\beta}_j - \beta_j^*|) + \frac{\lambda_1 \varepsilon_n}{2}. \tag{A.55}$$

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

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

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 (A.55), it is very small for $j \in H^c$. Thus we have

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

Finally we conclude the proof using Proposition 2.

A.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 A.11. *Let $\hat{\boldsymbol{\beta}}$ be the Elastic-net estimate of NBR defined in (1). Suppose that $\lambda_2 > 0$. Then for any $k, l \in \{1, 2, \ldots, p\}$,*

$$|\hat{\beta}_k - \hat{\beta}_l| \leq \frac{1}{2n\lambda_2} \sum_{i=1}^{n} \frac{\theta \, |x_{ik} - x_{il}| \, |e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}} - Y_i|}{\theta + e^{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}}}. \tag{A.56}$$

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 A.12. *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 $\mathrm{E}S_n \leq \mu$ for some constant μ;*

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

The proof of Lemma A.12 is straightforward which is given in Appendix B, and we present the proof of Theorem 4 in advance.

By Lemma A.11, Cauchy inequality, triangle inequality and Taylor expansion, we have

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

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

Under the assumption of oracle inequality (12), with probability $1-\frac{2}{p^{r-1}}-\exp\{-\frac{nt^2}{2C_{LB}^2 d_c^2 L^4}\}$, we have

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

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

A.7 Proof of Theorem 5

By KKT condition (see Lemma 3.1 and (2)), then we claim that $\mathrm{sgn}\hat{\boldsymbol{\beta}}=\mathrm{sgn}\boldsymbol{\beta}^*$ if

$$\begin{cases}\mathrm{sign}\hat{\beta}_j=\mathrm{sign}\beta_j^*, j\in H\\ \dot{\ell}_j(\hat{\beta})+2\lambda_2\hat{\beta}_j=-\lambda_1\mathrm{sign}\hat{\beta}_j, \hat{\beta}_j\neq 0\\ |\dot{\ell}_j(\hat{\beta})|\le\lambda_1, \hat{\beta}_j=0\end{cases}\tag{A.57}$$

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 $\mathrm{sign}\hat{\beta}_j=\mathrm{sign}\beta_j^*, j\in H$, then $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_H,0)^T$ is the solution of the KKT conditions. So, the (A.57)

holds if

$$\begin{cases} \text{sign}\hat{\beta}_j = \text{sign}\beta_j^*, j \in H \\ |\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H)| \le \lambda_1, j \notin H \end{cases} \Leftarrow \begin{cases} |\hat{\beta}_j - \beta_j^*| < |\beta_j^*|, j \in H \\ |\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H)| \le \lambda_1, j \notin H \end{cases} \tag{A.58}$$

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 (A.58) 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 $P(E) \ge P(E_1) + P(E_2) + P(E_3) - 2$. 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

$$P(E_1) \ge P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le \frac{e^{2a_\tau}(\zeta+1)d_H^*\lambda_1}{2C_t^2(\zeta,H)}) \ge 1 - \frac{2}{p^{r-1}} - 2p^2 e^{-\frac{nt^2}{2[d_H^* C_{LB}(1+\zeta)L^2]^2}}. \tag{A.59}$$

For E_2, thus we get

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

where we use $||\boldsymbol{w}^{(j)}||_2^2 \le \frac{L^2}{n}$ in Lemma A.5.

This implies that

$$P(E_2) \leq \eta\lambda_1) \geq 1 - 2p\exp\{-\frac{\eta^2\lambda_1^2 n}{2C_{LB}^2 L^2}\} = 1 - \frac{2}{p^{1-r\eta^2/C_{\xi,B_1}^2}}. \tag{A.60}$$

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

$$\begin{aligned}
\max_{j\notin H}|\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H) - \dot{\ell}_j(\boldsymbol{\beta}^*)| &= \max_{j\notin H}|\dot{\ell}_j(\hat{\boldsymbol{\beta}}_H) - \dot{\ell}_j(\boldsymbol{\beta}_H^*)| \\
&= \max_{j\notin H}\frac{1}{n}\left|\sum_{i=1}^n x_{ij}\theta[\frac{\theta+Y_i}{\theta+e^{\boldsymbol{X}_{iH}^T\hat{\boldsymbol{\beta}}_H}} - \frac{\theta+Y_i}{\theta+e^{\boldsymbol{X}_{iH}^T\boldsymbol{\beta}_H^*}}]\right| \\
&= \max_{j\notin H}\frac{1}{n}\left|\sum_{i=1}^n x_{ij}\frac{\theta(\theta+Y_i)e^{\boldsymbol{X}_{iH}^T\boldsymbol{\beta}_H^*}[e^{\boldsymbol{X}_{iH}^T(\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta}_H^*)}-1]}{(\theta+e^{\boldsymbol{X}_{iH}^T\hat{\boldsymbol{\beta}}_H})(\theta+e^{\boldsymbol{X}_{iH}^T\boldsymbol{\beta}_H^*})}\right| \\
&\leq \frac{L}{n}\sum_{i=1}^n\left|(\theta+Y_i)[e^{\boldsymbol{X}_{iH}^T(\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta}_H^*)}-1]\right| \\
&\leq \frac{L}{n}\sum_{i=1}^n\left|(\theta+Y_i)[\boldsymbol{X}_{iH}^T(\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta}_H^*) + o_p(|\boldsymbol{X}_{iH}^T(\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta}_H^*)|)]\right| \\
&\leq \frac{C_X L^2}{n}\sum_{i=1}^n|\theta+Y_i|||\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta}_H^*||_1.
\end{aligned} \tag{A.61}$$

where the second last inequality is by (12) and the boundedness of $|\boldsymbol{X}_{iH}^T(\hat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_H^*)|$, the last inequality (A.61) stems from $\|\boldsymbol{X}_i\|_\infty \leq L$ and C_X is determined by

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

Let $A_n := \frac{1}{n}\sum_{i=1}^n|\theta+Y_i|$, similar to the proof of Lemma A.12, we have

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

And we can find a constant $\mu(A) > 0$ such that $EA_n \leq \mu(A)$.

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

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

with probability at least $1-\frac{\sigma_n^2(A)}{nA^2}$.

Then (A.61) 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

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

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 (12) with high probability in Theorem 2 and (A.32), we conclude that

$$\begin{aligned}
P(E_3)&\ge P(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1\le\frac{e^{2a_\tau}(\zeta+1)d_H^*\tilde{\lambda}_1}{2C_t^2(\zeta,H)})\\
&\ge P(|A_n-\mathrm{E}A_n|\le A)-2p\exp\{-\frac{C_{\xi,B_1}^2\tilde{\lambda}_1^2 n}{2C_{LB}^2C_X L^2}\}-2p^2e^{-\frac{nt^2}{2[d_H^*C_{LB}(1+\zeta)L^2]^2}}\\
&\ge 1-\frac{\sigma_n^2(A)}{nA^2}-\frac{2}{p^{1-r\tilde{C}^2}}-2p^2e^{-\frac{nt^2}{2[d_H^*C_{LB}(1+\zeta)L^2]^2}}.
\end{aligned}\tag{A.62}$$

Combining (A.59),(A.60) and (A.62), we get

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

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

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

A.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 A.13. *(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 A.13. □

A.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 A.14 (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 A.14, we have

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

Let

$$
A_n^{(k)} = \frac{1}{n}\sum_{i=1}^{n} X_{ik}\theta\Big(\frac{e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}{\theta + e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}} - \frac{e^{X_i^T\boldsymbol{\beta}^*}}{\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}}\Big), \quad C_n^{(k)} = \frac{1}{n}\sum_{i=1}^{n} X_{ik}\theta\Big(\frac{Y_i}{\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}} - \frac{Y_i}{\theta + e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}\Big),
$$

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

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

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

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

$$|B_n^{(k)}| \geq \min_{j\in H} |\beta_j^*| - |\tilde{B}_n^{(k)}| \geq 2\lambda_1 - |\tilde{B}_n^{(k)}| \tag{A.63}$$

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

Lemma A.15 (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 A.15 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

$$P(|\tilde{B}_n^{(k)}| \geq \frac{\lambda_1}{4}) \leq P(\sum_{j\neq k}^{p} |\hat{\beta}_j - \beta_j^*||\frac{\theta}{n}\sum_{i=1}^{n} X_{ij}X_{ik}| \geq \frac{\lambda_1}{4})$$
$$\leq P(\sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| \geq \frac{\lambda_1 d_H^*}{4h})$$
$$\leq P(\sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| \geq \frac{2.25^2\lambda_1 d_H^*}{ak+2\lambda_2}) \leq \frac{\delta}{p}. \tag{A.64}$$

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: $\frac{1}{4L_1 h} \geq \frac{2.25^2}{ak+2\lambda_2}, k \leq 1-8h$, we get $h \leq \frac{ak+2\lambda_2}{20.25L_1+8a} \wedge \frac{1}{8}$.

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

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

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

$$P(|A_n^{(k)} - B_n^{(k)}| \geq \frac{\lambda_1}{4}) \leq P(\sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| \geq \frac{2.25^2\lambda_1 d_H^*}{ak+2\lambda_2}) \leq \frac{\delta}{p} \tag{A.65}$$

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: $\frac{1}{4L_2 h} \geq \frac{2.25^2}{ak+2\lambda_2}, k \leq 1-8h$, we have $h \leq \frac{ak+2\lambda_2}{20.25L_2+8a} \wedge \frac{1}{8}$. Then

$$P(|C_n^{(k)}| \geq \frac{\lambda_1}{4}) \leq P(\sum_{j=1}^{p} |\hat{\beta}_j - \beta_j^*| \geq \frac{2.25^2\lambda_1 d_H^*}{ak+2\lambda_2}) \leq \frac{\delta}{p}. \tag{A.66}$$

It remains to obtain the upper bound for the fourth term. This can adopt Lemma A.5 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 (A.31), we get

$$P(|D_n^{(k)}| \geq \frac{\lambda_1}{4}) = P(|\frac{1}{n}\sum_{i=1}^{n} \frac{X_{ik}\theta}{\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}}(Y_i - \mathrm{E}Y_i)| \geq \frac{\lambda_1}{4}\} \leq 2\exp\{-\frac{n\lambda_1^2}{32C_{LB}^2L^2}\}. \tag{A.67}$$

In summary, the four probabilities (A.64),(A.65),(A.66) and (A.67) 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 (A.66) and (A.67).

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 the inequality (24), it gives

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

B Assisted lemmas

We fix $Y_i = y_i$ in the proof of Lemma 1,A.11. Rewrite $\hat{\boldsymbol{\beta}}(\lambda_1,\lambda_2), \hat{\beta}_k(\lambda_1,\lambda_2), \hat{\beta}_l(\lambda_1,\lambda_2)$ as $\hat{\boldsymbol{\beta}}, \hat{\beta}_k, \hat{\beta}_l$ respectively.

B.1 Proof of Lemma 1

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

$$F(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}[(\theta+y_i)\log(\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}) - y_i\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}] + \lambda_1\sum_{i=1}^{p}|\hat{\beta}_i| + \lambda_2\sum_{i=1}^{p}|\hat{\beta}_i|^2. \quad \text{(B.68)}$$

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 (B.68) 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

$$\sum_{i=1}^{n}[x_{ik}\frac{\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}-y_i)}{\theta+e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}}]+\lambda_1\text{sign}\varepsilon=0,(k=1,2,\cdots,p).\tag{B.69}$$

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

B.2 Proof of Lemma A.11

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 (A.56) 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 (A.56) is also proved.

Case 3. When $\hat{\beta}_k \neq 0, \hat{\beta}_l = 0$. By the Case 1 in Lemma 1 and (B.69), 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 (A.56) 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 (A.56) is proved. If $\hat{\beta}_l \neq 0, \hat{\beta}_k = 0$, the proof is by the same method.

Case 4. When $\hat{\beta}_k = \hat{\beta}_l = 0$, (A.56) is obviously.

B.3 Proof of Lemma A.12

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

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

For (2), we obtain

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

So, it implies

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

B.4 Proof of Proposition 4

Lemma B.16. *Hoeffding's lemma Let $Y_1, \cdots, Y_n$ be independent centralized random variables on $\mathbb{R}$ satisfying bound condition*

$$\mathrm{E}Y_i = 0, |Y_i| \leq c_i \text{ for } i = 1, 2, \cdots, n. \tag{B.70}$$

holds. We have

$$\mathrm{E}\exp\{\lambda \sum_{i=1}^{n} Y_i\} \leq \exp\{\frac{1}{2}\lambda^2 \sum_{i=1}^{n} c_i^2\} \quad for \quad \lambda > 0.$$

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

have

$$\begin{aligned}
\mathrm{E}\max_{1\le j\le p}|V_j| &= \frac{1}{\lambda}\mathrm{E}\log \mathrm{e}^{\lambda\max_{1\le j\le p}|V_j|} \le \frac{1}{\lambda}\log \mathrm{E}\mathrm{e}^{\lambda\max_{1\le j\le p}|V_j|} \\
&\le \frac{1}{\lambda}\log\sum_{i=1}^{n}\mathrm{E}\mathrm{e}^{\lambda|V_j|} \le \frac{1}{\lambda}\log[\sum_{j=1}^{p} 2e^{\frac{1}{2}\lambda^2\sum_{i=1}^{n}a_{ij}^2}] \text{ [Apply Lemma B.16 to } \mathrm{E}\mathrm{e}^{\lambda|V_j|}] \\
&\le \frac{1}{\lambda}\log\left[2pe^{\frac{1}{2}\lambda^2\max_{1\le j\le p}\sum_{i=1}^{n}a_{ij}^2}\right] = \frac{1}{\lambda}\log(2p) + \frac{1}{2}\lambda\max_{1\le j\le p}\sum_{i=1}^{n}a_{ij}^2.
\end{aligned}$$

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

C The the proof of (A.42) and the value γ

C.1 The the proof of (A.42)

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, 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 C.17 (discrete Prékopa–Leindler inequality)**.** *Let* $\lambda \in [0,1]$ *and suppose* $f, g, h, k : \mathbb{Z}^n \to [0,\infty)$ *satisfy*

$$f(\boldsymbol{x})g(\boldsymbol{y}) \le h(\lfloor\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}\rfloor)k(\lceil(1-\lambda)\boldsymbol{x} + \lambda\boldsymbol{y}\rceil) \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n \tag{C.71}$$

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 $\gamma > 0$

$$\psi(\boldsymbol{x}) + \psi(\boldsymbol{y}) - \psi(\lceil \frac{1}{2}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y} \rceil) - \psi(\lfloor \frac{1}{2}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y} \rfloor) \geq \frac{\gamma}{4}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n. \qquad \text{(C.72)}$$

Let $\gamma = 1/2$. The (C.72) is a slightly extension *strongly convex with modulus of convexity* γ for continuous functions on $\mathbb{R}^n$

$$\lambda\psi(\boldsymbol{x}) + (1-\lambda)\psi(\boldsymbol{y}) - \psi(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \geq \frac{\gamma}{2}\lambda(1-\lambda)\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n \text{ and } \forall \alpha \in [0,1]$$

see Chapter 2 of Mahoney et al. (2018). Strongly discrete log-convex property requires the restricted behavior of continuous functions on lattice space. If $\gamma = 0$, (C.72) 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, it 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 C.8 (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*

$$P_\gamma\{|f(\boldsymbol{X}) - \mathrm{E}f(\boldsymbol{X})| \geq t\} \leq 2e^{-\frac{\gamma t^2}{4L^2}}. \tag{C.73}$$

The Theorem C.8 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 C.17 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 (C.74) in Lemma C.17 is satisfied with $\lambda = 1/2$, i.e.

$$e^{\frac{1}{2}[l(\boldsymbol{y})-h(\boldsymbol{x})-\psi(\boldsymbol{y})-\psi(\boldsymbol{x})]} \leq e^{-\frac{1}{2}\psi(\lceil \frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rceil)} \cdot e^{-\frac{1}{2}\psi(\lfloor \frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\rfloor)} \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n \tag{C.74}$$

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 (C.74).

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 C.17 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, i.e. $|h(\boldsymbol{x}) - h(\boldsymbol{Y})| \leq L\|\boldsymbol{x}-\boldsymbol{Y}\|_2$.

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 (C.73). □

C.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

$$\begin{aligned} \log p(\boldsymbol{y}) &=: \sum_{i=1}^n \log p_i(y_i) =: \sum_{i=1}^n \psi(y_i) \\ &= \sum_{i=1}^n \{\log\Gamma(\theta+y_i) + y_i\log\mu_i + \theta\log\theta - \log\Gamma(\theta) - \log y_i! - (\theta+y_i)\log(\theta+\mu_i)\}. \end{aligned}$$

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

$$\psi(y) = \psi(\lceil\tfrac{1}{2}x+\tfrac{1}{2}y\rceil) + \tfrac{1}{2}\psi'(\lceil\tfrac{1}{2}x+\tfrac{1}{2}y\rceil)(y-x) + \tfrac{1}{8}(y-x)^2\psi''(a_1)$$

$$\psi(x) = \psi(\lfloor\tfrac{1}{2}x+\tfrac{1}{2}y\rfloor) + \tfrac{1}{2}\psi'(\lfloor\tfrac{1}{2}x+\tfrac{1}{2}y\rfloor)(x-y) + \tfrac{1}{8}(y-x)^2\psi''(a_2)$$

where $a_1 = t_1y + (1-t_1)(x+y)/2, a_2 = t_2y + (1-t_1)(x+y)/2$ with $t_1,t_2\in[0,1]$.

So we have

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

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

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

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

Note that

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

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).

D 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 estima-

tors and its 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 a two subproblem iteratively algorithms which is applied by Wang et al. (2016). Firstly, we fit a 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 stoping criteria is 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

$$\mathrm{BIC}_{\hat{\beta}(\lambda_1,\lambda_2)} := -\frac{1}{n}\sum_{i=1}^{n}[Y_i \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} - (\theta + Y_i)\log(\theta + e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}})] - \lambda_2\|\hat{\boldsymbol{\beta}}\|_2^2 + \frac{\log n}{n}\hat{\mathrm{df}}(en) \quad \text{(D.75)}$$

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 a nearly optimal tuning parameters and then further tune the λ_1 such that 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 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 (22) 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 $\lambda_2 \leq \frac{\lambda_1}{8B}$ in Theorem 3, we put $\lambda_2 \approx 0.02\lambda_1$. By referring the BIC criterion (D.75) and the oracle inequality (20) 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, with help of the ridge penalty it reflects that Elastic-net is able to improve the accuracy of the estimation 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 bias, and the bias is corrected by de-biased procedures. The de-biased estimators have less 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 2 presents the de-biased Elastic-net and de-biased Lasso estimates of low dimensional coefficients in sparse high-dimensional NBR, and confidence intervals are given with 95% confidence level. We resort the package `fastclime` to get a sparse precision matrix estimate for $-\ddot{\ell}(\boldsymbol{\beta}^*)^{-1}$, the tuning parameter is assigned as 0.2 in CLIME method. The covering number is the number of true coefficients that are contained in the 95% confidence level. The de-biased lasso confidence intervals cover 10 true coefficients, while the covering number of de-biased Elastic-net is 7. The de-biased Elastic-net has shorter length of confidence interval than 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.

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

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.