Topics in High Dimensional Statistics

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Le Zhou

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Hui Zou, Adviser

December 2021

Acknowledgements

I would like to first express my deepest gratitude to my adviser, Professor Hui Zou, who has led me into the world of statistics and academia, encouraged and supported me continuously without hesitation. His deep insights in almost all branches of statistics have benefited me greatly. His wisdoms in not only research but also all aspects in life have enlightened me and shed lights on my future direction. I feel so lucky to have met him and become his student.

I would also like to express my sincere gratitude to my oral and defense committees. I thank Professor Yuhong Yang, Xiaoou Li, Saonli Basu, Lan Liu for kindly serving on my committee. Special thanks go to Professor Yuhong Yang, who has always been patient and offered generous help. I also want to genuinely thank Professor Dennis Cook for his support in my job application. Every lecture at University of Minnesota is a treasure to me, and I am grateful to every professor who has taught me here.

My heartfelt gratitude also goes to my academic brothers, sisters and fellow graduate students. I thank Qing Mai, Yuwen Gu, Boxiang Wang, Chenglong Ye, Yunan Wu, Yiyi Yin, Wenjun Lang, Mingxuan Han for their help within and outside research. Special thanks go to my most beloved brother, Boxiang Wang, for whom my appreciation goes beyond words. Special thanks also go to my dearest friends, Jiawei Zhang, Lin Zhang and Fan Yang. I owe you enormous debt that I could never repay. I also thank my beloved friends Gongting Peng and Rui Peng. The days spent with you were the happiest days in my past five years.

I must have saved the world in my past life to meet and know you all.

Last but foremost, I thank my beloved mom and dad, Li Liu and Guangsheng

Zhou. I have truly been blessed to be born as your child, to have your unconditional love and support. Without you, I never would have made it here.

Dedication

To the loving memory of my grandfather and grandmother, Caixian Liu and Zenglan Bi.

Abstract

In the era of big data, statistical application fields have been frequently encountering data sets where features measured for each sample are more than sample size. In these data sets, which we call high-dimensional data, traditional statistical tools are not feasible, imposing challenges from both theoretical and computational perspectives. This thesis is devoted to discuss several novel high dimensional methodologies with both solid theoretical justification and computational efficiency to cope with the new challenges in big data era.

Chapter 2 of the thesis systematically studies the estimation of a high dimensional heteroscedastic regression model. In particular, the emphasis is on how to detect and estimate the heteroscedasticity effects reliably and efficiently. To this end, we propose a cross-fitted residual regression approach and prove the resulting estimator is selection consistent for heteroscedasticity effects and establish its rates of convergence. Efficient algorithm is developed such that our method can be solved extremely fast. Chapter 3 introduces a novel methodology called sparse convoluted rank regression. The method is shown to maintain the good theoretical property of rank regression, a very popular alternative to the least squares. Moreover, it avoids the computational burden of rank regression caused by non-smooth loss, by adopting a smooth objective function, which is derived from a statistical point of view. Chapter 4 proposes a method called density convoluted support vector machine (DCSVM) for high dimensional classification. Theoretical error bound is established, and numerical examples demonstrate that our method outperforms SVM and other competitors in terms of both prediction accuracy and computational speed.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Due to the advanced technology for data collection over the past decades, scientific community has witnessed a surge of data complexity in many research fields such as genomics, genetics and finance, among others. As a result, it is very common for the number of predictors in the dataset to be far larger than the number of observations (Donoho et al., 2000). Under such high dimensionality in data, traditional statistical methods are often infeasible, posing new challenges for statisticians from both theoretical and computational perspectives.

To handle high dimensionality, many progress have been made during the past two decades. The most popular method is the sparse regularized estimation, which typically adopts the objective function with the form of an empirical loss plus an penalty term. Popular regularization methods include the ℓ_1 penalization (the lasso) (Tibshirani, 1996), the linearly constrained ℓ_1 minimization (the Dantzig selector) (Candes and Tao, 2007), and folded-concave penalization (Fan and Li, 2001), among others.

While existing regularization methods achieve success in many real applications, many limitations are still present and may potentially harm the performance when data does not behave as expected. One example is the homoscedasticity assumption, which is the common starting point for most theories supporting sparse regression

methods. Homoscedasticity refers to the constant variance of error, and is widely adopted merely for theoretical convenience. Nevertheless, there are many real applications where this assumption easily fails, and one must consider the heteroscedasticity, or non-constant variance in data. In fact, high dimensional datasets often exhibits heteroscedasticity due to the fact that measurement error can accumulate during the data collection process. To this end, in Chapter 2, we systematically study the estimation of a high dimensional heteroscedastic regression model. In particular, the emphasis is on how to detect and estimate the heteroscedasticity effects reliably and efficiently. We propose a novel approach named cross-fitted residual regression and prove the resulting estimator is selection consistent for heteroscedasticity effects and establish its rates of convergence. Our estimator is indexed by a set of tuning parameters to be determined by the data in practice. We further develop a high dimensional BIC criterion for tuning parameter selection and establish its consistency. Our method can be computed extremely fast, which takes only a few seconds on an ordinary dataset.

In many cases, the estimation efficiency and computational efficiency can not be shared by the same statistical method. For example, many high dimensional datasets exhibits heavy-tailedness, and it is well known that under heavy tailed error, classical least squares regression suffers from low estimation efficiency. An approach for achieving a higher estimation efficiency is to use the Wilcoxon rank regression (Hettmansperger and McKean, 2010; Wang and Li, 2009), whose high dimensional setting was investigated in Wang et al. (2020). However, due to the non-differentiability of the loss function in rank regression, the method can not be computed quickly, especially for high dimensional data. In Chapter 3, we resolve this problem by viewing the rank regression loss as a non-smooth empirical counterpart of a population level quantity. A smooth empirical counterpart is then derived by substituting a kernel density estimator for the true distribution in the expectation calculation. This view leads to

the convoluted rank regression loss and consequently the sparse penalized convoluted rank regression (CRR) for high-dimensional data. Under the same key assumptions for sparse rank regression, we establish the rate of convergence of the ℓ_1-penalized CRR for a tuning free penalization parameter and prove the strong oracle property of the folded concave penalized CRR. We further propose a high-dimensional Bayesian information criterion for selecting the penalization parameter in folded concave penalized CRR and prove its selection consistency. An efficient algorithm is developed for solving sparse convoluted rank regression, which scales well with high dimensions. Numerical examples demonstrate the promising performance of the sparse convoluted rank regression over the sparse rank regression.

In Chapter 4, we focus on high dimensional classification problem, where the state-of-the-art classification method, SVM, also suffers from computational issue. Adopting similar principle, we view the SVM objective function as the expectation of hinge loss function with respect to the non-smooth empirical measure corresponding to some true underlying measure, and a smooth counterpart is derived by substituting a kernel density estimator for the measure in the expectation calculation. This view leads to the density convoluted support vector machine (DCSVM) and consequently the penalized DCSVM for high-dimensional classification. We systematically study the rate of convergence of the DCSVM with elastic net penalty, and prove it has order $O_p(\sqrt{\frac{s\log p}{n}})$ under general random design setting. We further develop novel efficient algorithm for computing elastic-net DCSVM. Extensive numerical examples are used to demonstrate the superior performance of elastic-net DCSVM in both classification and computation.

Chapter 2

Cross-fitted Residual Regression for High Dimensional Heteroscedasticity Pursuit

There is a vast amount of work on high dimensional regression. The common starting point for the existing theoretical work is to assume the data generating model is a homoscedastic linear regression model with some sparsity structure. In reality the homoscedasticity assumption is often violated, and hence understanding the heteroscedasticity of the data is of critical importance. In this paper we systematically study the estimation of a high dimensional heteroscedastic regression model. In particular, the emphasis is on how to detect and estimate the heteroscedasticity effects reliably and efficiently. To this end, we propose a cross-fitted residual regression approach and prove the resulting estimator is selection consistent for heteroscedasticity effects and establish its rates of convergence. Our estimator has tuning parameters to be determined by the data in practice. We propose a novel high dimensional BIC for tuning parameter selection and establish its consistency. This is the first high dimensional BIC result under heteroscedasticity. The theoretical analysis is more involved in order to handle heteroscedasticity, and we develop a couple of interesting new concentration inequalities that are of independent interests.

2.1 Introduction

High dimensional linear regression has been extensively studied during the past two decades. Many fundamental developments have been made among which sparse regularized estimation plays an essential role including the ℓ_1 penalization (the lasso) (Tibshirani, 1996), the linearly constrained ℓ_1 minimization (the Dantzig selector) (Candes and Tao, 2007), and folded-concave penalization (Fan and Li, 2001), among others. For a comprehensive review on sparse regression, the readers are referred to Bühlmann and Van De Geer (2011), Fan et al. (2020), etc. The theories supporting the sparse regression methods typically begin with a common assumption that the data are generated from a homoscedastic linear model (see chapter 4 of Fan et al. (2020)):

$$y_i = \mathbf{x}_i^{\mathrm{T}} \beta^* + \epsilon_i, \ 1 \leq i \leq n \tag{2.1.1}$$

where ϵ_i's are independent and identically distributed model noise with mean zero and variance σ^2. While it is convenient to consider the homoscedastic noise assumption in theory, there are real applications where this assumption is easily violated and we must consider heteroscedasticity effects in the model. For example, financial time series often exhibits non-constant variance, due to varying volatility over the time. In genomic studies, outlying measurements can accumulate to cause heteroscedasticity, and it was shown by many studies (Wang, Wu and Li, 2012; Daye, Chen and Li, 2012) that not only the mean but also the scale of genetic responses could be affected by relevant genetic covariates. As mentioned above, most existing work on high dimensional analysis neglected the heteroscedasticity issue.

There are only a couple of papers that deal with the heteroscedasticity issue in high dimensional regression. Sparse quantile regression (Wang, Wu and Li (2012)) generalizes the quantile regression (Koenker and Bassett, 1978) by adopting folded-

concave penalization for sparsity. It assumes the conditional $100\tau\%$ quantile of $y_i|\mathbf{x}_i$ is $\mathbf{x}_i^{\mathrm{T}}\beta$, and the sparse quantile estimator is defined by

$$\hat{\beta}^{\mathrm{QR}} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\beta) + \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

where $\rho_\tau(t) = t(\tau - I_{\{t<0\}})$ is the quantile check loss, and $p_\lambda(\cdot)$ is some penalty function with tuning parameter λ. Wang, Wu and Li (2012) recommended to fit several quantile functions for different τ values and then one can see the heteroscedasticity effects by comparing the fitted quantiles functions. The main idea is that under the model ((2.1.1)) the theoretical quantile functions should be parallel to each other. Therefore, non-parallel fitted quantile functions indicate a violation of the model ((2.1.1)) and hence heteroscedasticity effects. The underlying model for the sparse quantile regression approach is

$$y_i = \mathbf{x}_i^{\mathrm{T}}\gamma^* + (\mathbf{x}_i^{\mathrm{T}}\omega + \omega_0)\epsilon_i,\ 1 \le i \le n \tag{2.1.2}$$

under which the τ quantile function is linear in $\mathbf{x}$ for every $\tau \in (0,1)$. Due to the non-differentiability of the check loss, sparse quantile regression can be computationally very expensive. Only recently, a fast and scalable algorithm based on ADMM was proposed and implemented in R package `FHDQR` (Gu et al., 2018). Moreover, sparse quantile regression does not directly show which variables are important in mean effects and which are important in the scale. To handle both issues, Gu and Zou (2016) proposed COSALES based on linear expectile regression (Newey and Powell, 1987; Efron, 1991):

$$(\hat{\gamma}, \hat{\boldsymbol{\varphi}}) = \arg\min_{\gamma,\boldsymbol{\varphi}} S_n(\gamma, \boldsymbol{\varphi}) + \sum_{j=1}^{p} p_{\lambda_1}(|\gamma_j|) + \sum_{j=1}^{p} p_{\lambda_2}(|\boldsymbol{\varphi}_j|),$$

where $S_n(\gamma, \boldsymbol{\varphi}) = \frac{1}{n}\sum_{i=1}^{n}\{\Psi_{0.5}(y_i - \mathbf{x}_i^{\mathrm{T}}\gamma) + \Psi_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\gamma - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\varphi})\}$, and $\Psi_\tau(u) = |\tau - I(u < 0)|u^2$ is the asymmetric square error loss. Theoretical justifications of COSALES were established under the model ((2.1.2)). COSALE is computationally very efficient because the asymmetric square error loss is convex and differentiable.

The model ((2.1.2)) is more or less a mathematical model for studying heteroscedacity. It may be difficult to apply model ((2.1.2)) in real applications. Note that the conditional scale of $y_i|\mathbf{x}_i$ can not be directly interpreted as $\mathbf{x}_i^{\mathrm{T}}\omega$, but its absolute value. A general model for heteroscedasticity effects would be

$$y_i = f(\mathbf{x}_i) + \mathrm{e}^{g(\mathbf{x}_i)}\epsilon_i, \tag{2.1.3}$$

where f and g are some unknown functionals that belong to some function classes $\mathcal{F}_1$ and $\mathcal{F}_2$, and ϵ_i is model noise with mean zero. In this paper, we focus on the linear version of model ((2.1.3)),

$$y_i = \mathbf{x}_i^{\mathrm{T}}\beta^* + \mathrm{e}^{\gamma_0^* + \mathbf{x}_i^{\mathrm{T}}\gamma^*}\epsilon_i,\ i = 1, \ldots, n, \tag{2.1.4}$$

where both f and g are modeled as linear functions of $\mathbf{x}$. Indeed, low dimensional version of this linear model has been widely considered in the literature (Feigl and Zelen, 1965; Cox and Snell, 1968; Cook and Weisberg, 1983; Carroll and Ruppert, 1988). The interpretation for model ((2.1.4)) is straightforward. The parameter γ^* in the variance structure can be directly interpreted as follows: with other covariates being fixed, if the jth covariate increases by one unit, then the conditional standard deviation of response given covariates increases by $100(\mathrm{e}^{\gamma_j} - 1)\%$. Note that the conditional mean of $y|\mathbf{x}$ is still a linear function of $\mathbf{x}$. Thus, the usual sparse mean regression methods should be able to estimate β^* well and recover its support under sparsity assumption on the mean effects. Let p be the dimension of covariates. It is directly implied by Fano's Lemma that even we know β^*, unless $\log p/n \to 0$, there

is no consistent estimator of γ^*. In many real applications, researchers are interested in knowing which variables are important for heteroscedasticity effects. For that, it is natural to assume γ^* is sparse. The problem of estimating γ^* is called high dimensional heteroscedasticity pursuit.

Quantile regression and COSALES do not work under the model ((2.1.4)) because the conditional quantile and expectile functions are no longer linear. Assuming the error distribution is normal, a penalized maximum likelihood estimator (MLE) was considered in Daye, Chen and Li (2012):

$$\min \sum_{i=1}^{n} [\exp(-\gamma_0 - \mathbf{x}_i^{\mathrm{T}}\gamma)(y_i - \mathbf{x}_i^{\mathrm{T}}\beta)^2 + (\gamma_0 + \mathbf{x}_i^{\mathrm{T}}\gamma)] + \sum_{j=1}^{p} p_{\lambda_1}(|\beta_j|) + \sum_{j=1}^{p} p_{\lambda_2}(|\gamma_j|). \quad (2.1.5)$$

Note that ((2.1.5)) is nonconvex. Daye, Chen and Li (2012) observed that ((2.1.5)) is bi-convex in that given γ solving β is a convex problem and given β solving γ is a convex problem. They proposed an alternating algorithm between solving β and γ, and for each iteration, they used coordinate descent to handle the computation. They implemented their method in an R package whose main function is written in Fortran. We tried their code in simulations and found the computation efficiency is too low to be practically useful: it took 9 hours to compute for $n = 700, p = 5000$. Moreover, no theoretical justification was given for the estimator. Note that the non-convexity of the log-likelihood makes its analysis technically nontrivial.

In light of the above discussions, we now formally state the objective as follows:

Heteroscedasticity pursuit: Can we design an explicit and efficient estimation procedure for estimating γ^* in the model ((2.1.4)) under the sparsity assumption on γ^* and $\log p/n \to 0$ without knowing the error distribution?

By “explicit" we mean that the whole procedure is clearly defined without any ambiguity such as the choice of starting value in an iterative algorithm. In this paper,

we propose a novel cross-fitted penalized residual regression method to estimate γ^* in model ((2.1.4)). The method is shown to consistently estimate the heteroscedasticity effect in ultra-high dimension. From computational perspective, our algorithm is very fast since it reduces to computing several weighted Lasso-type regressions. Thus our method can take advantage of the computational efficiency of ℓ_1-penalized least squares (Efron et al., 2004). We also develop the first information criterion for consistent model selection under heteroscedasticity and high dimensions. The theoretical analysis of the new procedure is highly nontrivial compared to the homoscedastic regression model case. For that, we develop a couple of interesting new concentration inequalities that are of independent interests.

The rest of this paper is organized as follows. In Section 2.2, we introduce the basic setup. In Section 3.2, we introduce our methodology and propose our cross-fitted penalized residual regression procedure. In Section 3.3, we establish the theoretical properties for our estimators. In 7, we introduce a high dimensional BIC for selecting tuning parameters and establish its model selection consistency. In Section 3.5, we present some simulation studies and a real data example to demonstrate the performance of our methodology. Some discussion is given in Section 2.7. The technical proofs for the theoretical results are given in appendices.

2.2 Basic setup and notation

We introduce some notation first. For an arbitrary index set $\mathbb{A} \subset \{1, \ldots, p\}$, any vector $\mathbf{c} = (c_1, \ldots, c_p)$ and any $n \times p$ matrix $\mathbf{U}$, let $\mathbf{c}_{\mathbb{A}} = (c_i, i \in \mathbb{A})$, and let $\mathbf{U}_{\mathbb{A}}$ be the submatrix with columns of $\mathbf{U}$ whose indices are in $\mathbb{A}$. The complement of an index set $\mathbb{A}$ is denoted as $\mathbb{A}^c = \{1, \ldots, p\} \setminus \mathbb{A}$. For any finite set $\mathbb{B}$, let $|\mathbb{B}|$ be the number of elements in $\mathbb{B}$. For a vector $\mathbf{c} \in \mathbb{R}^p$ and $q \in [1, \infty)$, let $\|\mathbf{c}\|_{\ell_q}$ (or $\|\mathbf{c}\|_q$) $= (\sum_{j=1}^p |c_j|^q)^{\frac{1}{q}}$ be its ℓ_q norm, let $\|\mathbf{c}\|_\infty$ (or $\|\mathbf{c}\|_{\max}$) $= \max_j |c_j|$ be its ℓ_∞ norm, and

let $\|\mathbf{c}\|_{\min} = \min_j |c_j|$ be its minimum absolute value. For a symmetric matrix $\mathbf{M}$, let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ be its smallest and largest eigenvalue, respectively. This is the common notation for eigenvalues of a matrix, and $\lambda_{\min}$, $\lambda_{\max}$ should not be confused with the penalization parameter used in a penalty function. For symmetric matrices A and B, we denote $A \preceq B$ if $B - A$ is a positive semidefinite matrix. For any matrix $\mathbf{G}$, let $\|\mathbf{G}\| = \sqrt{\lambda_{\max}(\mathbf{G}^{\mathrm{T}}\mathbf{G})}$ be its spectral norm. In particular, for a vector $\mathbf{c}$, $\|\mathbf{c}\| = \|\mathbf{c}\|_{\ell_2}$. For $a, b \in \mathbb{R}$, let $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For two nonnegative sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \gtrsim b_n$ if there exists $C > 0$ such that $b_n \leq C a_n$ for all $n \geq 1$, and we write $a_n \asymp b_n$ if $a_n \gtrsim b_n$ and $b_n \gtrsim a_n$. Also, we use $b_n = o(a_n)$ to represent $\frac{b_n}{a_n} \to 0$.

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p) \in \mathbb{R}^{n \times p}$ be the design matrix, with $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ containing observations for the jth variable, $j = 1, \ldots, p$. The ith row of $\mathbf{X}$ can be written as $\mathbf{x}_i^{\mathrm{T}}$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. Let $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ be the n-dimensional response vector. We consider the following regression model,

$$y_i = \beta_0^* + \mathbf{x}_i^{\mathrm{T}}\beta^* + \mathrm{e}^{\gamma_0^* + \mathbf{x}_i^{\mathrm{T}}\gamma^*}\epsilon_i, \ i = 1, \ldots, n. \tag{2.2.1}$$

where $\beta_0^* \in \mathbb{R}$, $\beta^* \in \mathbb{R}^p$ are unknown parameters that control the conditional mean, and $\gamma_0^* \in \mathbb{R}$, $\gamma^* \in \mathbb{R}^p$ are unknown parameters that control the conditional scale; $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. random variables that are independent of the covariates with $\mathbb{E}[\epsilon_i] = 0$.

Remark 1 We assume that $\mathbb{E}[|\log|\epsilon_i||] < \infty$, which is satisfied by most continuous probability distributions. Without loss of generality, we can assume $\mathbb{E}[\log|\epsilon_i|] = 0$. In fact, if $\mathbb{E}[\log|\epsilon_i|] = c_0 \neq 0$, let $\epsilon_i' = \epsilon_i \mathrm{e}^{-c_0}$. Then we have $\mathbb{E}[\epsilon_i'] = 0$, and $\mathbb{E}[\log|\epsilon_i'|] = \mathbb{E}[\log|\epsilon_i \mathrm{e}^{-c_0}|] = \mathbb{E}[\log|\epsilon_i|] - c_0 = 0$. Correspondingly, let $\gamma_0^{*'} = \gamma_0^* + c_0$, we have $\mathrm{e}^{\gamma_0^* + \mathbf{x}_i^{\mathrm{T}}\gamma^*}\epsilon_i = \mathrm{e}^{\gamma_0^{*'} + \mathbf{x}_i^{\mathrm{T}}\gamma^*}\epsilon_i'$, and therefore $y_i = \beta_0^* + \mathbf{x}_i^{\mathrm{T}}\beta^* + \mathrm{e}^{\gamma_0^{*'} + \mathbf{x}_i^{\mathrm{T}}\gamma^*}\epsilon_i'$ holds with

$\mathbb{E}[\epsilon_i'] = 0$ and $\mathbb{E}[\log|\epsilon_i'|] = 0$. □

Notice that one can always set the components of the first column of design matrix to be one, without loss of generality. We rewrite model ((2.2.1)) as follows, for notational convenience:

$$y_i = \mathbf{x}_i^{\mathrm{T}}\beta^* + \mathrm{e}^{\mathbf{x}_i^{\mathrm{T}}\gamma^*}\epsilon_i,\ i = 1, \ldots, n, \tag{2.2.2}$$

where $\mathbf{x}_i, \beta^*, \gamma^* \in \mathbb{R}^p$. We let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ be the design matrix and let $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ be its jth column.

Let $\mathbb{A}_1 = \{j : \beta_j^* \neq 0\}$ and $\mathbb{A}_2 = \{j : \gamma_j^* \neq 0\}$ be the true support set of β^* and γ^*, respectively. Assume that $|\mathbb{A}_1| = s_1$ and $|\mathbb{A}_2| = s_2$ are relatively of smaller order compared to n, while p is allowed to increase exponentially with n. Penalized methods are used for sparse recovery. We consider the general folded concave penalty (Fan et al., 2014b) in this paper; namely, $p_\lambda(\cdot)$ is a function defined on $(-\infty, \infty)$ satisfying:

(i) $p_\lambda(-z) = p_\lambda(z)$;

(ii) $p_\lambda(z)$ is increasing and concave in $z \in [0, \infty)$, and $p_\lambda(0) = 0$;

(iii) $p_\lambda(z)$ is differentiable in $z \in (0, \infty)$, and $p_\lambda'(0) := p_\lambda'(0+) \geq a_1\lambda$;

(iv) $p_\lambda'(z) \geq a_1\lambda$ for $z \in (0, a_2\lambda]$;

(v) $p_\lambda'(z) = 0$ for $z \in [a\lambda, \infty)$ with some pre-specified constant $a > a_2$,

where a_1 and a_2 are two fixed positive constants. Special cases of general folded concave penalty are SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). The SCAD

penalty has the form

$$p_\lambda(|t|) = \lambda|t| I(0 \le |t| < \lambda) + \frac{a\lambda|t| - (t^2+\lambda^2)/2}{a-1} I(\lambda \le |t| \le a\lambda) + \frac{(a+1)\lambda^2}{2} I(|t| > a\lambda), \text{ for some } a > 2,$$

which corresponds to $a_1 = a_2 = 1$. The MCP penalty function is defined as

$$p_\lambda(|t|) = \lambda \left(|t| - \frac{t^2}{2a\lambda} \right) I(0 \le |t| < a\lambda) + \frac{a\lambda^2}{2} I(|t| \ge a\lambda), \text{ for some } a > 1,$$

which corresponds to $a_1 = 1 - \frac{1}{a}, a_2 = 1$.

2.3 Methodology

2.3.1 Heuristics for penalized residual regression

Our goal is to produce an estimator $\hat{\gamma}$ that can efficiently perform the estimation and support recovery for γ^*. The mean coefficient β is a nuisance parameter in this problem, although estimation of β plays an important role in the estimation of γ^*.

To get some intuition, let us assume that β^* is fully known, then recall from ((2.2.2)) that

$$\log|y_i - \mathbf{x}_i^{\mathrm{T}}\beta^*| = \mathbf{x}_i^{\mathrm{T}}\gamma^* + \log|\epsilon_i|.$$

By $\mathbb{E}[\log|\epsilon_i|] = 0$ (see Remark 1), we can view this as a new linear model with $\{\log|\epsilon_i|\}_{i=1}^n$ being i.i.d. errors. We would consider estimating γ^* through minimizing

$$\frac{1}{2n}\sum_{i=1}^n \left(\log|y_i - \mathbf{x}_i^{\mathrm{T}}\beta^*| - \mathbf{x}_i^{\mathrm{T}}\gamma\right)^2 + \sum_{j=1}^p p_\lambda(|\gamma_j|)$$

where $p_\lambda(\cdot)$ is a folded concave penalty imposed on γ. To mimic the above idea, we may consider using

$$\frac{1}{2n}\sum_{i=1}^{n}\left(\log|y_i-\mathbf{x}_i^{\mathrm{T}}\hat{\beta}|-\mathbf{x}_i^{\mathrm{T}}\gamma\right)^2+\sum_{j=1}^{p}p_\lambda(|\gamma_j|) \tag{2.3.1}$$

where $\hat{\beta}$ is some good estimator of β^*. Note that $y_i-\mathbf{x}_i^{\mathrm{T}}\hat{\beta}$ is called fitted residual in regression, hence the above method for estimating γ^* is called penalized residual regression.

We must show that we do have some appropriate estimator for β^* such that the above procedure leads to a good estimator of γ^*. Under high dimensional setting, we also need to keep the computational efficiency of the whole procedure in mind. Thus, a natural way of estimating β^* is to use penalized ordinary least squares, i.e., minimizing

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i-\mathbf{x}_i^{\mathrm{T}}\beta)^2+\sum_{j=1}^{p}p_{\lambda'}(|\beta_j|) \tag{2.3.2}$$

over $\beta\in\mathbb{R}^p$. We show that despite the heteroscedasticity effects the penalized OLS estimator of β^* still enjoys nice rates of convergence.

2.3.2 Cross-fitted penalized residual regression

It is very clear that the penalized residual regression is computationally very efficient. To handle the theoretical justification of the method, we need to carefully take care of the dependence between $\hat{\beta}$ and data which causes technical challenges in the analysis of penalized residual regression. We further propose using cross-fitted penalized residual regression to estimate the heteroscedasticity effect γ^*. The whole procedure is summarized in Figure 2.1. The cross-fitted technique allows us to handle a broad class of error distributions. Cross-fitting has been used for variance estimation in ho-

moscedastic high dimensional linear regression in Fan et al. (2012) who showed that in sparse linear regression, the usual mean squared error estimator underestimates the true variance, and they proposed a so-called refitted cross-validation technique, which was shown to consistently estimate the variance. The idea was extended to estimating constant variance in high dimensional sparse additive model in Chen et al. (2018). To our best knowledge, this work is the first to extend cross-fitting to heteroscedastic models.

Remark 2 Empirical experiments suggest that the penalized residual regression without cross-fitting may also work well under certain distributions. However, penalized residual regression without cross-fitting has no rigorous theoretical justification under general error distributions so far. Therefore, we do not present it in the present paper. □

We summarize the details of cross-fitted penalized residual regression. We use $Z = (y_i, \mathbf{x}_i)_{i=1}^n$ to represent the whole dataset. In cross-fitted penalized residual regression, we first randomly split the n random samples into two datasets with approximately same size, with the first dataset being $Z^{(1)} = (y_i, \mathbf{x}_i)_{i\in\mathcal{I}_1}$, and the second dataset being $Z^{(2)} = (y_i, \mathbf{x}_i)_{i\in\mathcal{I}_2}$, where $\mathcal{I}_1$ and $\mathcal{I}_2$ are disjoint subsets of $\{1, \ldots, n\}$ such that $\mathcal{I}_1 \cup \mathcal{I}_2 = \{1, \ldots, n\}$. Without loss of generality, we assume $\mathcal{I}_1 = \{1, \ldots, \frac{n}{2}\}$ and $\mathcal{I}_2 = \{\frac{n}{2} + 1, \ldots, n\}$ throughout this paper. As can be seen from Figure 2.1, the big picture is as follows. We first construct initial estimators of β^* on $Z^{(1)}$ and $Z^{(2)}$ through penalized ordinary least squares, denoted as $\hat{\beta}(Z^{(1)})$ and $\hat{\beta}(Z^{(2)})$. Then, we do the penalized residual regression on $Z^{(2)}$ using residuals by $\hat{\beta}$ from $\hat{\beta}(Z^{(1)})$, and on $Z^{(1)}$ using residuals by $\hat{\beta}$ from $\hat{\beta}(Z^{(2)})$. The resulting estimators of γ^* are denoted as $\hat{\gamma}(Z^{(1)} \to Z^{(2)})$ and $\hat{\gamma}(Z^{(2)} \to Z^{(1)})$, respectively. The final estimator of γ^*, $\hat{\gamma}^{\text{ave}}$, is simply the average of these two.

Next, we explain in detail the steps $(\mathcal{O}, \mathcal{R})$ in Figure 2.1.

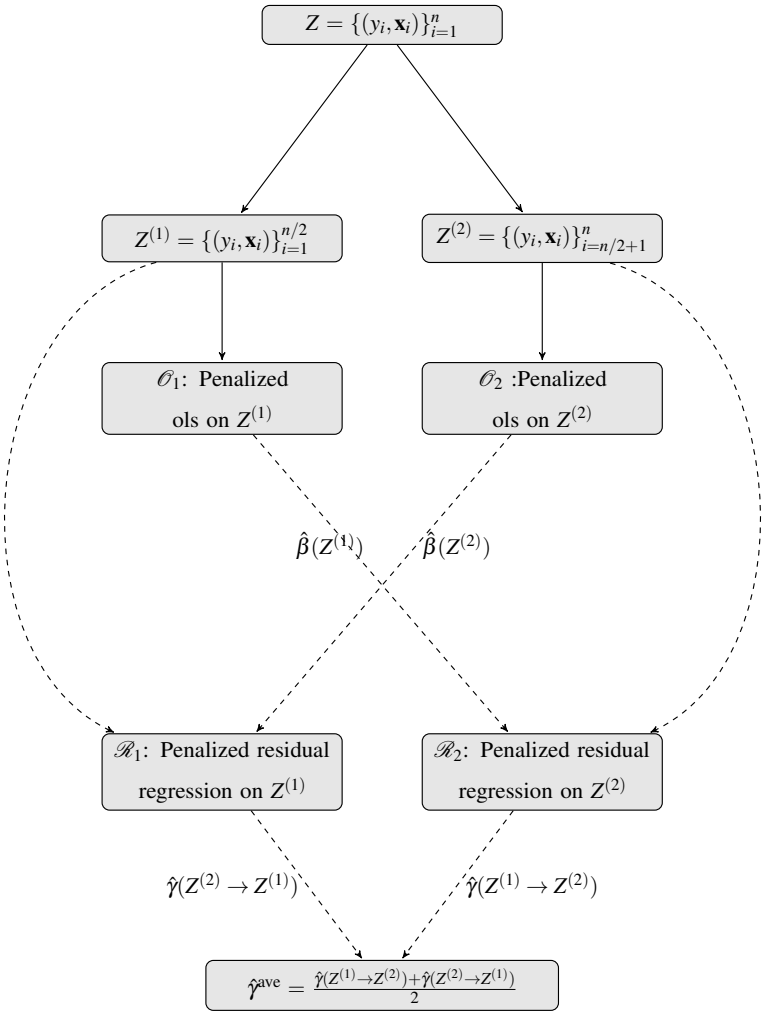

Figure 2.1. *Cross-fitted penalized residual regression*

Penalized OLS ($\mathcal{O}$) in Figure 2.1

We take penalized OLS on $Z^{(1)}$ as an example (i.e. $\mathcal{O}_1$) and illustrate how we get the initial estimator $\hat{\beta}(Z^{(1)})$. $\hat{\beta}(Z^{(2)})$ is computed through exactly the same procedure on a different dataset (with different tuning parameters), as indicated from Figure 2.1. We use folded-concave penalty instead of lasso penalty for estimating β^* in penalized OLS, because the latter induces some unnecessary bias which is propagated into penalized residual regression. We use

$$\frac{1}{n}\sum_{i=1}^{\frac{n}{2}}(y_i - \mathbf{x}_i^{\mathrm{T}}\beta)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{2.3.3}$$

with $p_\lambda(\cdot)$ being a folded concave penalty and $\lambda > 0$ being the tuning parameter. Following Fan et al. (2014b), we use the local linear approximation (LLA) algorithm (Zou and Li, 2008) for solving ((2.3.3)). We adopt zero vector as the initial value for the LLA algorithm. $\hat{\beta}(Z^{(1)})$ is defined as the solution that the LLA algorithm gives after convergence. Similarly, we can define $\hat{\beta}(Z^{(2)})$ and let the corresponding tuning parameters be $\tilde{\lambda}$.

Penalized residual regression ($\mathcal{R}$) in Figure 2.1

We take the penalized residual regression on $Z^{(2)}$ as an example (i.e. $\mathcal{R}_2$) and explain how we get $\hat{\gamma}(Z^{(1)} \to Z^{(2)})$. The procedure of getting $\hat{\gamma}(Z^{(2)} \to Z^{(1)})$ is similar. Let $\hat{z}_i = \log|y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}(Z^{(1)})|$. From the heuristics that we have provided, we consider minimizing

$$\frac{1}{n}\sum_{i=\frac{n}{2}+1}^{n}\left(\hat{z}_i - \mathbf{x}_i^{\mathrm{T}}\gamma\right)^2 + \sum_{j=1}^{p} p_{\lambda_1}(|\gamma_j|) \tag{2.3.4}$$

over $\gamma \in \mathbb{R}^p$. Here, λ_1 is some tuning parameter that could differ from the previous ones. Again, we adopt LLA algorithm for solving ((2.3.4)), and we use zero vector as the initial value for the LLA algorithm. The $\hat{\gamma}(Z^{(1)} \to Z^{(2)})$ is defined as the solution we get after the LLA algorithm converges. Similarly, we can define $\hat{\gamma}(Z^{(2)} \to Z^{(1)})$ and let $\tilde{\lambda}_1$ be the corresponding tuning parameters. The final estimator of γ^*, as can be seen from figure Figure 2.1, is $\hat{\gamma}^{\text{ave}} = \frac{\hat{\gamma}(Z^{(1)} \to Z^{(2)}) + \hat{\gamma}(Z^{(2)} \to Z^{(1)})}{2}$. The non-zero elements of $\hat{\gamma}^{\text{ave}}$ form the estimated subset of variables that impact the scale function.

2.4 Theory

In this section we provide the theoretical justifications for the estimators.

We make the following assumption for the distribution of ϵ_i's:

($\mathbf{A}_0$) $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. sub-Gaussian(σ) random variables with some fixed positive constant σ, i.e. $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\exp(t\epsilon_i)] \leq \exp(\sigma^2 t^2/2)$. Also, $\mathbb{E}[\log|\epsilon_i|] = 0$. Moreover, the distribution of ϵ_i has a density f on $\mathbb{R}$ with respect to Lebesgue measure which satisfies $|f(x) - f(y)| \leq L|x - y|, \forall x, y \in \mathbb{R}$, for some constant $L > 0$. Consequently, we have $C_0 \coloneqq \sup_{x \in \mathbb{R}} f(x) < \infty$.

Remark 3 The assumption of sub-Gaussian random variable is common in the literature on high dimensional statistics. The condition $\mathbb{E}[\log|\epsilon_i|] = 0$ has been justified in Remark 1. The assumption of error having Lipschitz continuous density is satisfied by most continuous probability distributions. □

For any index set $\mathbb{A} \subset \{1, \ldots, p\}$, let $\mathcal{S}_{\mathbb{A}} \coloneqq \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}_{\mathbb{A}^c}\|_{\ell_1} \leq 3\|\mathbf{u}_{\mathbb{A}}\|_{\ell_1} \neq 0\}$. Let $\mathbf{y}^{(1)} = (y_1, \ldots, y_{\frac{n}{2}})^{\mathrm{T}}$, $\mathbf{y}^{(2)} = (y_{\frac{n}{2}+1}, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{X}^{(1)} = (\mathbf{x}_1, \ldots, \mathbf{x}_{\frac{n}{2}})^{\mathrm{T}}$, and $\mathbf{X}^{(2)} = (\mathbf{x}_{\frac{n}{2}+1}, \ldots, \mathbf{x}_n)^{\mathrm{T}}$. Recall $\mathbb{A}_1 = \{j : \beta_j^* \neq 0\}$, $\mathbb{A}_2 = \{j : \gamma_j^* \neq 0\}$, $s_1 = |\mathbb{A}_1|$ and $s_2 = |\mathbb{A}_2|$, and let $s = \max(s_1, s_2)$. We impose the following assumptions $\mathbf{C}_1$, $\mathbf{C}_2$ or $\mathbf{C}_2'$, and $\mathbf{C}_3$ on the design matrix:

($\mathbf{C}_1$) There exist $M \in (0,\infty)$ such that $|x_{ij}|^2 \le M, \forall 1 \le i \le n, 1 \le j \le p$. And $\exists 0 < \Psi \le \Omega < \infty$ such that $\Psi \le \mathrm{e}^{\mathbf{x}_i^{\mathrm{T}}\gamma^*} \le \Omega, \forall i$.

($\mathbf{C}_2$) $\kappa := \min_{\mathbf{u}\in\mathcal{S}_{\mathbb{A}_1}} \frac{\|\mathbf{X}^{(1)}\mathbf{u}\|_{\ell_2}^2}{\frac{n}{2}\|\mathbf{u}\|_{\ell_2}^2} \wedge \frac{\|\mathbf{X}^{(2)}\mathbf{u}\|_{\ell_2}^2}{\frac{n}{2}\|\mathbf{u}\|_{\ell_2}^2} \in (0,\infty)$, and $\kappa' := \min_{\mathbf{u}\in\mathcal{S}_{\mathbb{A}_2}} \frac{\|\mathbf{X}^{(1)}\mathbf{u}\|_{\ell_2}^2}{\frac{n}{2}\|\mathbf{u}\|_{\ell_2}^2} \wedge \frac{\|\mathbf{X}^{(2)}\mathbf{u}\|_{\ell_2}^2}{\frac{n}{2}\|\mathbf{u}\|_{\ell_2}^2} \in (0,\infty)$. These imply $\min_{\mathbf{u}\in\mathcal{S}_{\mathbb{A}_1}} \frac{\|\mathbf{X}\mathbf{u}\|_{\ell_2}^2}{n\|\mathbf{u}\|_{\ell_2}^2} \ge \kappa$, $\min_{\mathbf{u}\in\mathcal{S}_{\mathbb{A}_2}} \frac{\|\mathbf{X}\mathbf{u}\|_{\ell_2}^2}{n\|\mathbf{u}\|_{\ell_2}^2} \ge \kappa'$.

($\mathbf{C}_2'$) $\rho := \min_{\mathbf{u}\in\mathcal{S}_{\mathbb{A}_1}} \frac{\|\mathbf{X}^{(1)}\mathbf{u}\|_{\ell_2}^2}{\frac{n}{2}\|\mathbf{u}_{\mathbb{A}_1}\|_{\ell_1}\|\mathbf{u}\|_\infty} \wedge \frac{\|\mathbf{X}^{(2)}\mathbf{u}\|_{\ell_2}^2}{\frac{n}{2}\|\mathbf{u}_{\mathbb{A}_1}\|_{\ell_1}\|\mathbf{u}\|_\infty} \in (0,\infty)$, and
$\rho' := \min_{\mathbf{u}\in\mathcal{S}_{\mathbb{A}_2}} \frac{\|\mathbf{X}^{(1)}\mathbf{u}\|_{\ell_2}^2}{\frac{n}{2}\|\mathbf{u}_{\mathbb{A}_2}\|_{\ell_1}\|\mathbf{u}\|_\infty} \wedge \frac{\|\mathbf{X}^{(2)}\mathbf{u}\|_{\ell_2}^2}{\frac{n}{2}\|\mathbf{u}_{\mathbb{A}_2}\|_{\ell_1}\|\mathbf{u}\|_\infty} \in (0,\infty)$.
These imply that $\min_{\mathbf{u}\in\mathcal{S}_{\mathbb{A}_1}} \frac{\|\mathbf{X}\mathbf{u}\|_{\ell_2}^2}{n\|\mathbf{u}_{\mathbb{A}_1}\|_{\ell_1}\|\mathbf{u}\|_\infty} \ge \rho$ and $\min_{\mathbf{u}\in\mathcal{S}_{\mathbb{A}_2}} \frac{\|\mathbf{X}\mathbf{u}\|_{\ell_2}^2}{n\|\mathbf{u}_{\mathbb{A}_2}\|_{\ell_1}\|\mathbf{u}\|_\infty} \ge \rho'$.

($\mathbf{C}_3$) $\varphi := \lambda_{\min}(\frac{1}{n/2}\mathbf{X}_{\mathbb{A}_1}^{(1)^{\mathrm{T}}}\mathbf{X}_{\mathbb{A}_1}^{(1)}) \wedge \lambda_{\min}(\frac{1}{n/2}\mathbf{X}_{\mathbb{A}_1}^{(2)^{\mathrm{T}}}\mathbf{X}_{\mathbb{A}_1}^{(2)}) \in (0,\infty)$, and also
$\varphi' := \lambda_{\min}(\frac{1}{n/2}\mathbf{X}_{\mathbb{A}_2}^{(1)^{\mathrm{T}}}\mathbf{X}_{\mathbb{A}_2}^{(1)}) \wedge \lambda_{\min}(\frac{1}{n/2}\mathbf{X}_{\mathbb{A}_2}^{(2)^{\mathrm{T}}}\mathbf{X}_{\mathbb{A}_2}^{(2)}) \in (0,\infty)$. These imply that $\lambda_{\min}(\frac{1}{n}\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1}) \ge \varphi$ and $\lambda_{\min}(\frac{1}{n}\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2}) \ge \varphi'$.

Condition ($\mathbf{C}_1$) keeps the magnitude of design and the variances from blowing up. Condition ($\mathbf{C}_2$) and ($\mathbf{C}_2'$) are known as the restricted eigenvalue condition (RE) and the generalized invertability factor (GIF) condition respectively, which are commonly used in the literature to study the estimation accuracy of the Lasso and Dantzig estimators in high dimensional setting. We refer to Bickel et al. (2009), Meier et al. (2009), Ye and Zhang (2010), Huang and Zhang (2012) and Negahban et al. (2012) for discussions on these conditions and other relevant conditions. Condition ($\mathbf{C}_3$) is used to rule out the case where the important covariates with indices in set $\mathbb{A}_1$(or $\mathbb{A}_2$) are linearly dependent.

Let $a_0 = \min\{1, a_2\}$ where a_2 is the constant associated with the folded-concave penalty function. For SCAD or MCP, $a_2 = 1$ and hence $a_0 = 1$. Let $\hat{\mathbb{A}}^{(1)} = \{i : \hat{\gamma}(Z^{(1)} \to Z^{(2)})_i \ne 0, i = 1,\ldots,p\}$ and $\hat{\mathbb{A}}^{(2)} = \{i : \hat{\gamma}(Z^{(2)} \to Z^{(1)})_i \ne 0, i = 1,\ldots,p\}$ be the support sets of $\hat{\gamma}(Z^{(1)} \to Z^{(2)})$ and $\hat{\gamma}(Z^{(2)} \to Z^{(1)})$. And let $\hat{\mathbb{A}} = \{i : \hat{\gamma}_i^{\mathrm{ave}} \ne 0, i = 1,\ldots,p\}$ be the support of our final estimator. The following theorem demonstrates the oracle properties of the estimators.

Theorem 1 Consider the SCAD or MCP as the penalty function. Let assumptions $(\mathbf{A}_0)$, $(\mathbf{C}_1)$, $(\mathbf{C}_2)$ or $(\mathbf{C}_2')$ and $(\mathbf{C}_3)$ hold. Assume that $a_0\kappa \geq 3s_1^{\frac{1}{2}}$ and $a_0\kappa' \geq 3s_2^{\frac{1}{2}}$ hold under $(\mathbf{C}_2)$, or $a_0\rho \geq 3$ and $a_0\rho' \geq 3$ hold under $(\mathbf{C}_2')$. Choose the tuning parameters so that $\|\beta^*_{\mathbb{A}_1}\|_{\min} > (a+1)(\lambda \vee \tilde{\lambda})$ and $\|\gamma^*_{\mathbb{A}_2}\|_{\min} > (a+1)(\lambda_1 \vee \tilde{\lambda}_1)$. Then we have

(i) $\hat{\mathbb{A}}^{(1)} = \mathbb{A}_2$ holds true with probability at least $1 - \xi(\lambda, \lambda_1)$, where

$$\begin{aligned}\xi(x,y) \coloneqq\ & 2p\exp\big(-\frac{nx^2}{16M\sigma^2\Omega^2}\big) + 2(p-s_1)\exp\big(-\frac{a_1^2nx^2}{4\sigma^2\Omega^2M}\big)\\ & + 2s_1\exp\big(-\frac{n\varphi(\|\beta^*_{\mathbb{A}_1}\|_{\min}-ax)^2}{4\sigma^2\Omega^2}\big)\\ & + 2p\exp(-d_1n) + 2(p-s_2)\exp(-d_2n) + 2s_2\exp(-d_3n)\\ & + n\exp\big(-\frac{(K\wedge\frac{f_1}{(4L+2)G_1}\wedge\frac{f_2}{(4L+2)G_2}\wedge\frac{f_3}{(4L+2)G_3})^2\Psi^2\varphi}{4\sigma^2\Omega^2s_1M}n\big),\end{aligned}$$

where $d_1 = d_1(y) = \frac{y^2}{64\eta_0^2M} \wedge \frac{y}{16\eta_0\sqrt{M}}$, $d_2 = d_2(y) = \frac{a_1^2y^2}{16\eta_0^2G_2^2} \wedge \frac{a_1y}{8\eta_0G_2}$, $d_3 = d_3(y) = \frac{\varphi'^2(\|\gamma^*_{\mathbb{A}_2}\|_{\min}-ay)^2}{16\eta_0^2s_2M} \wedge \frac{\varphi'(\|\gamma^*_{\mathbb{A}_2}\|_{\min}-ay)}{8\eta_0\sqrt{s_2M}}$, $f_1 = f_1(y) = \frac{y}{2}$, $f_2 = f_2(y) = a_1y$, $f_3 = f_3(y) = \|\gamma^*_{\mathbb{A}_2}\|_{\min} - ay$, $G_1 \coloneqq \sqrt{M}$, $G_2 \coloneqq \sqrt{M}(\frac{2s_2M}{\varphi'}+1)$, $G_3 \coloneqq \frac{\sqrt{s_2M}}{\varphi'}$, and η_0, K are some fixed positive constants whose definitions can be found in Lemma 3.

(ii) $\hat{\mathbb{A}}^{(2)} = \mathbb{A}_2$ holds true with probability at least $1 - \xi(\tilde{\lambda}, \tilde{\lambda}_1)$.

(iii): Assume $\frac{s_1\log n}{n} \to 0$. Suppose we suitably choose $\lambda \asymp \tilde{\lambda} \asymp \sqrt{\frac{\log p}{n}}$ and $\lambda_1 \asymp \tilde{\lambda}_1 \asymp (s_2\sqrt{\frac{\log(p-s_2)}{n}} \vee \frac{s_2\log(p-s_2)}{n} \vee s_2\sqrt{\frac{s_1\log n}{n}})$. Then we have

$$\begin{aligned}&\|\hat{\gamma}(Z^{(1)} \to Z^{(2)}) - \gamma^*\|_{\ell_2} = O_p(\sqrt{s/n}),\\ &\|\hat{\gamma}(Z^{(2)} \to Z^{(1)}) - \gamma^*\|_{\ell_2} = O_p(\sqrt{s/n}),\end{aligned}$$

and therefore $\|\hat{\gamma}^{\text{ave}} - \gamma^*\|_{\ell_2} = O_p(\sqrt{s/n})$. Moreover, $\mathrm{P}(\hat{\mathbb{A}} \neq \mathbb{A}_2) \to 0$. □

2.5 Consistency of BIC tuning

Theorem 1 shows that the proposed estimator is good in principle. In practice, we also need to specify the tuning parameters used in the procedure. For homoscedastic regression models, the commonly used tuning methods include cross-validation and information criteria. However, the theory for CV or information criteria is not well understood under heteroscedastic regression models. In this section, we propose a new BIC for selecting the tuning parameters in our estimator and prove its selection consistency. Note that the study on high dimensional BIC has been reported in several papers such as Wang et al. (2013) and Fan and Tang (2013) where the underlying model is the standard homoscedastic regression model. To the best of our knowledge, our theory of BIC for heteroscedastic regression is the first in the literature.

To highlight the importance of tuning parameters, we use notation $\hat{\beta}^{\lambda}(Z^{(1)})$ and $\hat{\gamma}^{\lambda_1}(Z^{(1)} \to Z^{(2)})$ to denote the estimator of β^* and γ^* in $\mathcal{O}_1$ and $\mathcal{R}_2$ that correspond to the tuning parameter λ and λ_1. Similarly, we define $\hat{\beta}^{\tilde{\lambda}}(Z^{(2)})$ and $\hat{\gamma}^{\tilde{\lambda}_1}(Z^{(2)} \to Z^{(1)})$ in $\mathcal{O}_2$ and $\mathcal{R}_1$. The high dimensional Bayesian information criteria in $\mathcal{O}_1$ and $\mathcal{O}_2$ are defined as

$$\mathrm{HBIC}^{(1)}(\lambda) = \log\left(\frac{2}{n}\sum_{i=1}^{\frac{n}{2}}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}^{\lambda}(Z^{(1)})\right)^2\right) + |M_{\lambda}^{(1)}|\frac{C_{n,p}^{(1)}\log p}{n},$$
$$\mathrm{HBIC}^{(2)}(\lambda) = \log\left(\frac{2}{n}\sum_{i=\frac{n}{2}+1}^{n}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}^{\lambda}(Z^{(2)})\right)^2\right) + |M_{\lambda}^{(2)}|\frac{C_{n,p}^{(2)}\log p}{n},$$

where $M_{\lambda}^{(1)} = \{j : \hat{\beta}_j^{\lambda}(Z^{(1)}) \neq 0\}$ and $M_{\lambda}^{(2)} = \{j : \hat{\beta}_j^{\lambda}(Z^{(2)}) \neq 0\}$, and the choice of $C_{n,p}^{(i)}, i = 1, 2$ is discussed in Proposition 1. The corresponding tuning parameters for $\mathcal{O}_1$ and $\mathcal{O}_2$ are chosen by minimizing the proposed $\mathrm{HBIC}^{(1)}$ and $\mathrm{HBIC}^{(2)}$, respectively.

Proposition 1 Let $\hat{\lambda} = \arg\min_{\lambda\in\Lambda}\mathrm{HBIC}^{(1)}(\lambda)$ and $\hat{\tilde{\lambda}} = \arg\min_{\lambda\in\tilde{\Lambda}}\mathrm{HBIC}^{(2)}(\lambda)$,

where $\Lambda = \{\lambda : |M_\lambda^{(1)}| \leq K_n^{(0)}\}$ and $\tilde{\Lambda} = \{\lambda : |M_\lambda^{(2)}| \leq K_n^{(0)}\}$, and $K_n^{(0)} > s_1$ is allowed to diverge to infinity.

Under the conditions of Theorem 1, assume that there exists a positive constant c_0 such that for $i = 1, 2$,

$$\liminf_{n\to\infty} \min_{\mathbb{A}\not\supseteq\mathbb{A}_1, |\mathbb{A}|\leq K_n^{(0)}} \left\{ \frac{1}{n} \left\| \left(\mathbf{I}_{n/2} - \mathbf{P}_{\mathbb{A}}^{(i)} \right) \mathbf{X}_{\mathbb{A}_1}^{(i)} \boldsymbol{\beta}_{\mathbb{A}_1}^* \right\|^2 \right\} > c_0, \tag{2.5.1}$$

where $\mathbf{P}_{\mathbb{A}}^{(i)} = \mathbf{X}_{\mathbb{A}}^{(i)} (\mathbf{X}_{\mathbb{A}}^{(i)\mathrm{T}} \mathbf{X}_{\mathbb{A}}^{(i)})^{-1} \mathbf{X}_{\mathbb{A}}^{(i)\mathrm{T}}$ is the projection matrix onto the column space of $\mathbf{X}_{\mathbb{A}}^{(i)}$, $i = 1, 2$. If $C_{n,p}^{(i)} \to \infty$, $\frac{C_{n,p}^{(i)} s_1 \log p}{n} = o(1), i = 1, 2$, and $\frac{K_n^{(0)} \log p}{n} = o(1)$, then we have

$$\mathrm{P}(M_{\hat{\lambda}}^{(1)} = \mathbb{A}_1) \to 1, \quad \text{and} \quad \mathrm{P}(M_{\hat{\tilde{\lambda}}}^{(2)} = \mathbb{A}_1) \to 1,$$

as $n, p \to \infty$. □

Remark 4 The conditions ((2.5.1)) is the asymptotic model identifiability condition used in Wang et al. (2013). Proposition 1 is for the selection consistency of BIC for mean regression under heteroscedasticity. As expected, it is similar to the previous results on BIC for mean regression under homoscedasticity. We only need to bound the influences due to heteroscedasticity, and the rest proof of Proposition 1 are similar to Wang et al. (2013). Thus, we omit its proof for the sake of space. □

Proposition 1 is an intermediate step in our BIC theory because our goal is to develop BIC for model selection with regard to the heteroscedasticity pursuit problem. We need to first construct a BIC for our estimator of γ^* and then prove its selection consistency. To gain some intuition, suppose the true β^* is known, then recall we have $\log|y_i - \mathbf{x}_i^{\mathrm{T}}\beta^*| = \mathbf{x}_i^{\mathrm{T}}\gamma^* + \log|\epsilon_i|$, which can be treated as a homoscedastic regression. The BIC for the homoscedastic case can be applied. However, β^* is unknown, so we

can use its estimators to replace it. Following our cross-fitted procedure in Figure 2.1, we define the HBIC in $\mathcal{R}_1$ and $\mathcal{R}_2$ as

$$\begin{aligned}
&\mathrm{HBIC}^{(3)}(\lambda) \\
&= \log\left(\frac{2}{n}\sum_{i=1}^{\frac{n}{2}}\left(\log|y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}(Z^{(2)})| - \mathbf{x}_i^{\mathrm{T}}\hat{\gamma}^{\lambda}(Z^{(2)}\to Z^{(1)})\right)^2\right) + |M_{\lambda}^{(3)}|\frac{C_{n,p}^{(3)}\log p}{n}, \\
&\mathrm{HBIC}^{(4)}(\lambda) \\
&= \log\left(\frac{2}{n}\sum_{i=\frac{n}{2}+1}^{n}\left(\log|y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}(Z^{(1)})| - \mathbf{x}_i^{\mathrm{T}}\hat{\gamma}^{\lambda}(Z^{(1)}\to Z^{(2)})\right)^2\right) + |M_{\lambda}^{(4)}|\frac{C_{n,p}^{(4)}\log p}{n}
\end{aligned}$$

where $M_{\lambda}^{(3)} = \{j : \hat{\gamma}_j^{\lambda}(Z^{(2)} \to Z^{(1)}) \neq 0\}$ and $M_{\lambda}^{(4)} = \{j : \hat{\gamma}_j^{\lambda}(Z^{(1)} \to Z^{(2)}) \neq 0\}$. Note that $\hat{\beta}(Z^{(2)})$ in $\mathrm{HBIC}^{(3)}$ is the $\mathrm{HBIC}^{(2)}$ tuned estimator from step $\mathcal{O}_2$ in Figure 2.1. Likewise, $\hat{\beta}(Z^{(1)})$ in $\mathrm{HBIC}^{(4)}$ is the $\mathrm{HBIC}^{(1)}$ tuned estimator from step $\mathcal{O}_1$ in Figure 2.1. The tuning parameters in $\mathcal{R}_1$ and $\mathcal{R}_2$ are then chosen by minimizing $\mathrm{HBIC}^{(3)}$ and $\mathrm{HBIC}^{(4)}$, respectively. The choice of $C_{n,p}^{(i)}, i = 3, 4$ is discussed in Theorem 2 in order to achieve model selection consistency.

Theorem 2 Let $\hat{\lambda}_1 = \arg\min_{\lambda\in\Lambda_1}\mathrm{HBIC}^{(4)}(\lambda)$ and $\hat{\tilde{\lambda}}_1 = \arg\min_{\lambda\in\tilde{\Lambda}_1}\mathrm{HBIC}^{(3)}(\lambda)$, where $\Lambda_1 = \{\lambda : |M_{\lambda}^{(4)}| \leq K_n\}$ and $\tilde{\Lambda}_1 = \{\lambda : |M_{\lambda}^{(3)}| \leq K_n\}$, and $K_n > s = \max(s_1, s_2)$ is allowed to diverge to infinity. Under the conditions of Proposition 1, assume there exists a positive constant c_0' such that

$$\liminf_{n\to\infty}\min_{\mathbb{A}\not\supseteq\mathbb{A}_2, |\mathbb{A}|\leq K_n}\left\{\frac{1}{n}\left\|\left(\mathbf{I}_{n/2} - \mathbf{P}_{\mathbb{A}}^{(i)}\right)\mathbf{X}_{\mathbb{A}_2}^{(i)}\boldsymbol{\gamma}_{\mathbb{A}_2}^*\right\|^2\right\} > c_0', \tag{2.5.2}$$

where $\mathbf{P}_{\mathbb{A}}^{(i)} = \mathbf{X}_{\mathbb{A}}^{(i)}(\mathbf{X}_{\mathbb{A}}^{(i)\mathrm{T}}\mathbf{X}_{\mathbb{A}}^{(i)})^{-1}\mathbf{X}_{\mathbb{A}}^{(i)\mathrm{T}}$ is the projection matrix onto the column space of $\mathbf{X}_{\mathbb{A}}^{(i)}$, $i = 1, 2$. Also assume that $\phi := \min_{i\in\{1,2\}}\min_{|\mathbb{A}|\leq K_n}\lambda_{\min}(\frac{2\mathbf{X}_{\mathbb{A}}^{(i)\mathrm{T}}\mathbf{X}_{\mathbb{A}}^{(i)}}{n}) > 0$.

If $C_{n,p}^{(i)} \to \infty$, $\frac{s_1\log n}{C_{n,p}^{(i)}\log p} = o(1)$, $\frac{C_{n,p}^{(i)}s_2\log p}{n} = o(1), i = 3, 4$, and $\frac{K_n^2\log p}{n} = o(1)$, then

we have

$$\mathrm{P}(M^{(3)}_{\hat{\hat{\lambda}}_1} = \mathbb{A}_2) \to 1, \quad \text{and} \quad \mathrm{P}(M^{(4)}_{\hat{\lambda}_1} = \mathbb{A}_2) \to 1,$$

as $n, p \to \infty$. □

Remark 5 Similar to ((2.5.1)), ((2.5.2)) is the asymptotic model identifiability condition for the scale effects. □

2.6 Numeric Results

We demonstrate the sparsity recovery and estimation accuracy of our procedure through several simulation examples and a real data example. For the nonconvex penalty function, we use the SCAD penalty.

2.6.1 Simulation

In all simulations, we generated data ($n = 700$ and $p = 5000$) from the following heteroscedastic regression model $y = \mathbf{x}^{\mathrm{T}}\beta^* + \mathrm{e}^{\mathbf{x}^{\mathrm{T}}\gamma^*}\epsilon$, where in the fifth and tenth examples ϵ follows mixture of normal distributions, and in the other examples ϵ follows the standard normal distribution. In example 1–5, β^* is less sparse compared to γ^*, and in example 6–10, γ^* is less sparse compared to β^*. We let $(x_1, \ldots, x_p) \sim N(\mathbf{0}, \Sigma)$ with $\Sigma_{ij} = \rho^{|i-j|}, 1 \le i, j \le p$.

- Example 1.
  - $\rho = 0.5$; $y = 6(x_6 + x_{12} + x_{15} + x_{20}) + \mathrm{e}^{0.9x_1}\epsilon$, where $\epsilon \sim N(0, 1)$.
- Example 2.

- $\rho = 0.85$; $y = 6(x_6 + x_{12} + x_{15} + x_{20}) + \mathrm{e}^{0.9x_1}\epsilon$, where $\epsilon \sim N(0,1)$.

- Example 3.
  - $\rho = 0.5$; $y = 6(x_6 + x_{12} + x_{15} + x_{20}) + \mathrm{e}^{0.9x_1+0.9x_{12}}\epsilon$, where $\epsilon \sim N(0,1)$.

- Example 4.
  - $\rho = 0.85$; $y = 6(x_6 + x_{12} + x_{15} + x_{20}) + \mathrm{e}^{0.9x_1+0.9x_{12}}\epsilon$, where $\epsilon \sim N(0,1)$.

- Example 5.
  - $\rho = 0.5$; $y = 6(x_6 + x_{12} + x_{15} + x_{20}) + \mathrm{e}^{0.9x_1+0.9x_{12}}\epsilon$, where $\epsilon \sim \frac{1}{2\sqrt{10}}N(3,1) + \frac{1}{2\sqrt{10}}N(-3,1)$.

- Example 6.
  - $\rho = 0.5$; $y = 6x_{25} + \mathrm{e}^{0.5x_1+0.5x_{12}+0.5x_{25}+0.5x_{46}}\epsilon$, where $\epsilon \sim N(0,1)$.

- Example 7.
  - $\rho = 0.85$; $y = 6x_{25} + \mathrm{e}^{0.5x_1+0.5x_{12}+0.5x_{25}+0.5x_{46}}\epsilon$, where $\epsilon \sim N(0,1)$.

- Example 8.
  - $\rho = 0.5$; $y = 6x_{25} + 6x_{59} + \mathrm{e}^{0.5x_1+0.5x_{12}+0.5x_{25}+0.5x_{46}}\epsilon$, where $\epsilon \sim N(0,1)$.

- Example 9.
  - $\rho = 0.85$; $y = 6x_{25} + 6x_{59} + \mathrm{e}^{0.5x_1+0.5x_{12}+0.5x_{25}+0.5x_{46}}\epsilon$, where $\epsilon \sim N(0,1)$.

- Example 10.
  - $\rho = 0.5$; $y = 6x_{25} + \mathrm{e}^{0.5x_1+0.5x_{12}+0.5x_{25}+0.5x_{46}}\epsilon$, where $\epsilon \sim \frac{1}{2\sqrt{19}}N(-4,1) + \frac{1}{4\sqrt{19}}N(2,\frac{1}{2}) + \frac{1}{4\sqrt{19}}N(6,\frac{3}{2})$.

We used the first $\frac{n}{2}$ rows of the data in $\mathcal{O}_1$ and $\mathcal{R}_1$ steps in Figure 2.1 and the remaining $\frac{n}{2}$ rows of the data in $\mathcal{O}_2$ and $\mathcal{R}_2$ steps in Figure 2.1. We used $\mathbf{0}$ as the initial values for all the LLA algorithms involved in our procedure. The tuning parameters were selected by using the high dimensional BIC method as shown in 7. We let $C_{n,p}^{(i)} = \log\log n, i = 1, 2, 3, 4$. Let $\hat{\beta}^{\text{ave}}$ be the average of two estimators of β^* from $\mathcal{O}_1$ and $\mathcal{O}_2$. Let $\hat{\mathbb{A}}_1 = \{j : \hat{\beta}_j^{\text{ave}} \neq 0\}$, $\hat{\mathbb{A}}_2 = \{j : \hat{\gamma}_j^{\text{ave}} \neq 0\}$ be our selected submodel. After model selection, we refit γ^* by the same cross-fitted residual regression on the selected submodel without using any penalty. This extra step does not change model selection results but may further reduce estimation bias due to penalization, which is related to the relaxed lasso (Meinshausen, 2007; Hastie et al., 2017).

We evaluated the performance of our final estimator $\hat{\gamma}$ based on 400 replications. We evaluated the performance by the following quantities:

- ℓ_1: the average ℓ_1 risk $\|\hat{\gamma} - \gamma^*\|_{\ell_1}$.
- ℓ_2: the average ℓ_2 risk $\|\hat{\gamma} - \gamma^*\|_{\ell_2}$.
- **CP**: coverage probability, which is the proportion of replicates where $\mathbb{A}_2 \subset \hat{\mathbb{A}}_2$, $\mathbb{A}_2 = \{j : \hat{\gamma}_j^* \neq 0\}$ is the active set of γ^*, and $\hat{\mathbb{A}}_2 = \{j : \hat{\gamma}_j \neq 0\}$ is the active set of $\hat{\gamma}$.
- **TP**: true positives, which is the average size of the set $\mathbb{A}_2 \cap \hat{\mathbb{A}}_2$.
- **FP**: false positives, which is the average size of the set $\mathbb{A}_2^{\text{c}} \cap \hat{\mathbb{A}}_2$.

In general, estimating the scale function is different from estimating the mean function. In our model, the error in estimating β^* can be transferred into the estimation of γ^*. To highlight the impact of estimating β^* in the estimation of γ^*, we consider an idealized estimators of γ^*, denoted by $\hat{\gamma}^{\text{oracle}}$, by assuming the true support set of β^* is known in our estimation procedure. In other words, we replace the estimators of β^* in our original procedure by their corresponding oracle estimators

of β^*. This estimator, $\hat{\gamma}^{\text{oracle}}$, utilizes the information of the support set of β^*, which is not available in practice. Its performance provides a benchmark for comparing an actual estimator. Note that this estimator avoids high-dimensionality in estimating β^*.

The results for the idealized estimator ($\hat{\gamma}^{\text{oracle}}$) and our actual estimator ($\hat{\gamma}$) are summarized in 1 and Table 2.2. We first discuss the selection results. We can see that $\hat{\gamma}$ can cover the true support set $\mathbb{A}_2$ with high probability in all examples. Moreover, the average false positives for $\hat{\gamma}$ are very small in all examples, suggesting that our proposed method selects very few redundant variables. We can see that even under highly correlated design matrix ($\rho = 0.85$), the selection performance of our method is only slightly affected. Thus, in terms of variable selection, our proposed method is almost perfect. This confirms our theoretical results for the selection consistency of the proposed method. As for estimation accuracy, 1 and Table 2.2 show that the estimation error of $\hat{\gamma}$ is very close to the error of $\hat{\gamma}^{\text{oracle}}$ in all examples. This comparison demonstrates that our actual estimator $\hat{\gamma}$ is almost the same as the idealized estimator $\hat{\gamma}^{\text{oracle}}$, which also confirms our theory. Moreover, it suggests that there is little room for improvement for the estimation performance of our estimator. Finally, we can see that our method works very well in the cases where the error does not follow normal distribution. This is consistent with the fact that our theory does not require the error distribution to be normal.

Remark 6 The implemented HHR algorithm in Daye, Chen and Li (2012) was initially considered as a competing method of estimating γ^*. However, their algorithm costs over 9 hours for one replication. So we did not present their results here. □

Table 2.1. *Simulation results for example 1–5. The estimation accuracy is measured by ℓ_1 and ℓ_2. The sparsity recovery performance is measured by the last three columns. The ideal selection result would be $CP = 1$, $TP = 1$, $FP = 0$ for example 1–2, and $CP = 1$, $TP = 2$, $FP = 0$ for example 3–5 (shown in "Truth" rows). The standard errors listed in the parentheses are based on 400 replications.*

Example	Estimator	ℓ_1	ℓ_2	CP	TP	FP
1	$\hat{\gamma}^{\text{oracle}}$	0.113 (0.003)	0.112 (0.003)	100%	1.000 (0.000)	0.018 (0.007)
	$\hat{\gamma}$	0.111 (0.003)	0.109 (0.003)	100%	1.000 (0.000)	0.025 (0.008)
	Truth				1	0
2	$\hat{\gamma}^{\text{oracle}}$	0.125 (0.003)	0.123 (0.003)	100%	1.000 (0.000)	0.020 (0.008)
	$\hat{\gamma}$	0.124 (0.003)	0.123 (0.003)	100%	1.000 (0.000)	0.013 (0.006)
	Truth				1	0
3	$\hat{\gamma}^{\text{oracle}}$	0.705 (0.009)	0.488 (0.006)	99.25%	1.993 (0.004)	0.258 (0.027)
	$\hat{\gamma}$	0.711 (0.010)	0.493 (0.007)	99.5%	1.995 (0.004)	0.245 (0.026)
	Truth				2	0
4	$\hat{\gamma}^{\text{oracle}}$	0.953 (0.012)	0.641 (0.008)	96.25%	1.963 (0.010)	0.668 (0.046)
	$\hat{\gamma}$	0.940 (0.014)	0.635 (0.009)	94.25%	1.935 (0.014)	0.665 (0.045)
	Truth				2	0
5	$\hat{\gamma}^{\text{oracle}}$	0.468 (0.010)	0.335 (0.007)	99.75%	1.998 (0.003)	0.113 (0.018)
	$\hat{\gamma}$	0.482 (0.010)	0.344 (0.007)	100%	2.000 (0.000)	0.115 (0.017)
	Truth				2	0

2.6.2 A Real Data Example

In this section, we apply our procedure to a microarray data set used in Scheetz et al. (2006). The data set consists of summary gene expression values of 18986 probe sets on 120 twelve-week-old male offspring rats, which were analyzed on a logarithmic scale. The goal is to find the genes whose expression share strong association with the expression of gene TRIM32 (corresponding to probe 1389163_at). This gene is found to cause Bardet-Biedl syndrome (Chiang et al., 2006), which is a human genetic disorder that affects many body systems including the retina. We apply the fused Kolmogorov filter (Mai and Zou, 2015) to obtain a reduced set of 300 probes. The 300 probes are standardized so that they have mean zero and standard deviation one. The key motivation for using variable screening is to remove noise variables and

Table 2.2. *Simulation results for example 6–10. The estimation accuracy is measured by ℓ_1 and ℓ_2. The sparsity recovery performance is measured by the last three columns. The ideal selection result would be $CP = 1$, $TP = 4$, $FP = 0$ (shown in "Truth" rows). The standard errors listed in the parentheses are based on 400 replications.*

Example	Estimator	ℓ_1	ℓ_2	CP	TP	FP
6	$\hat{\gamma}^{\text{oracle}}$	0.379 (0.009)	0.192 (0.004)	98.5%	3.983 (0.007)	1.053 (0.055)
	$\hat{\gamma}$	0.405 (0.011)	0.208 (0.006)	96.25%	3.938 (0.018)	0.963 (0.049)
	Truth				4	0
7	$\hat{\gamma}^{\text{oracle}}$	0.704 (0.016)	0.321 (0.007)	91.25%	3.888 (0.021)	2.565 (0.086)
	$\hat{\gamma}$	0.737 (0.017)	0.326 (0.007)	90%	3.900 (0.015)	2.100 (0.057)
	Truth				4	0
8	$\hat{\gamma}^{\text{oracle}}$	0.418 (0.010)	0.211 (0.005)	98.75%	3.970 (0.015)	0.813 (0.050)
	$\hat{\gamma}$	0.438 (0.010)	0.221 (0.005)	96.25%	3.950 (0.013)	0.863 (0.044)
	Truth				4	0
9	$\hat{\gamma}^{\text{oracle}}$	0.773 (0.017)	0.346 (0.007)	90%	3.880 (0.020)	2.658 (0.094)
	$\hat{\gamma}$	0.802 (0.016)	0.359 (0.008)	90%	3.813 (0.033)	2.663 (0.089)
	Truth				4	0
10	$\hat{\gamma}^{\text{oracle}}$	0.169 (0.006)	0.094 (0.003)	100%	4.000 (0.000)	0.410 (0.034)
	$\hat{\gamma}$	0.138 (0.005)	0.078 (0.002)	100%	4.000 (0.000)	0.313 (0.035)
	Truth				4	0

boost the signal to noise ratio in data, which in practice can benefit the performance of statistical methods in further analysis. Many methods for variable screening has been proposed in the literature. The reader is referred to chapter 8 of Fan et al. (2020) for a comprehensive review of this topic. Here we choose the fused Kolmogorov filter as a part of our procedure because it is a nonparametric model-free method and its strong theoretical guarantee and excellent empirical performance (Mai and Zou, 2015).

We first compare our method with the standard penalized least square developed for homoscedastic regression. We randomly split the data to generate a training set with 100 samples, and the rest were treated as testing data on which we can compute the prediction error. The whole process was repeated 100 times. 6 summarizes the prediction errors as well as variable selection outcomes with the standard errors listed in the parentheses. For our method, we can list the number of variables selected for

the mean ($|\hat{A}_1|$) and the scale ($|\hat{A}_2|$), as well as the number of variables that affect both mean and scale ($|\hat{A}_1 \cap \hat{A}_2|$). For SCAD penalized least square $|\hat{A}_2|$ is non-applicable. The values along with their standard errors are reported in the last three columns of 6. We see that our method provides more accurate prediction, implying that the loss of prediction power of SCAD penalized least squares comes from the ignorance of possible heteroscedasticity of the data.

Table 2.3. *Results based on 100 replicates. The average prediction error under squared error loss is given in column 2. The average number of variables selected for the mean and scale (only for our method) are given in columns 3 to 5. Standard errors are given in the parentheses.*

Method	Prediction error	$\lvert\hat{A}_1\rvert$	$\lvert\hat{A}_2\rvert$	$\lvert\hat{A}_1 \cap \hat{A}_2\rvert$
SCAD-LS	0.017 (0.002)	5.13 (0.44)	N/A	N/A
Our method	0.013 (0.001)	12.34 (1.09)	4.67 (1.14)	0.17 (0.06)

We further examine the fitted model by our method. We fit our method on the whole dataset. In Table 2.4, we report the probes selected and their corresponding estimated coefficients, respectively. It is worth noting that probe X1396130_at is found to be related to both the mean and the scale.

Table 2.4. *Estimated mean and heteroscedasticity effect using the whole data set.*

Mean effect	
Probe	Estimated value
X1380486_at	0.017
X1389183_at	0.021
X1378901_at	-0.016
X1373709_at	0.027
X1378758_at	0.021
X1369143_at	0.008
X1379614_at	0.018
X1391529_at	-0.045
X1390539_at	0.028
X1396310_at	-0.011
X1369326_at	0.014
X1369756_a_at	0.011
X1395404_at	0.011
X1380123_at	-0.018

Heteroscedasticity effect	
Probe	Estimated value
X1372930_at	0.108
X1396310_at	0.155
X1381083_at	0.269
X1382278_at	-0.082

2.7 Discussion

In this paper we have proposed a log-linear model for modeling the heteroscedasticity in a high-dimensional regression model and have developed a cross-fitted penalized residual regression for estimating the heteroscedasticity effects with strong theoretical guarantee and excellent empirical performance. There are other ways to model heteroscedasticity. For example, it is known that the mixed-effect models can also be used to model heteroscedasticity. A high dimensional linear mixed effects model can be written as follows

$$y_i = \mathbf{x}_i^{\mathrm{T}}\beta + \mathbf{z}_i^{\mathrm{T}}\gamma_i + \epsilon_i,\ i = 1, \ldots, n,$$

where $\mathbf{x}_i, \beta \in \mathbb{R}^p$, $\mathbf{z}_i, \gamma_i \in \mathbb{R}^q$, $\{\epsilon_i\}_{i=1}^n$ are i.i.d. random errors with mean zero and variance σ^2, β is the unknown fixed effects, $\{\gamma_i\}_{i=1}^n$ are i.i.d. unknown random effects with mean zero and covariance matrix G, and $\{\gamma_i\}_{i=1}^n$ are independent from $\{\epsilon_i\}_{i=1}^n$. G is a $q \times q$ covariance matrix, which is typically unknown. In high dimensions, p and q are allowed to grow exponentially with n, and sparsity in both fixed effects and random effects is typically assumed. Under these settings, the conditional mean and variance of response given covariates are $\mathbb{E}[y_i|\mathbf{x}_i, \mathbf{z}_i] = \mathbf{x}_i^{\mathrm{T}}\beta^*$ and $\mathrm{var}(y_i|\mathbf{x}_i, \mathbf{z}_i) = \mathbf{z}_i^{\mathrm{T}}G\mathbf{z}_i + \sigma^2$. Therefore, the high dimensional linear mixed effects model also allows for high dimensional heteroscedasticity. The scale function is $\sqrt{\mathbf{z}_i^{\mathrm{T}}G\mathbf{z}_i + \sigma^2}$ or the log-scale function is $\frac{1}{2}(\mathbf{z}_i^{\mathrm{T}}G\mathbf{z}_i + \sigma^2)$. In our model the log-scale function is modeled as $\mathbf{x}_i^{\mathrm{T}}\gamma^*$. As mentioned before, it is natural to consider a linear model for the log-scale function, which has been considered in the literature (Feigl and Zelen, 1965; Cox and Snell, 1968; Cook and Weisberg, 1983; Carroll and Ruppert, 1988; Daye, Chen and Li, 2012). For a positive univariate quantity, it is often more natural to consider its multiplicative effects, which can be well handled by logarithm transformation (Cleveland, 1993).

Mathematically, it is hard to tell which model for the log-scale function is the best. We notice that existing popular works on high-dimensional mixed-effects model all focused on the mean function part and variable selection in mixed effects. It is shown that without knowing the true G matrix, such goals can be achieved (Fan and Li, 2012). On the other hand, the estimation of heteroscedasticity in the linear mixed-effects model requires the estimation of the true G matrix, which seems to be very much understudied in the high-dimensional literature. It would be an important topic for future study if one uses the linear mixed-effects model for heteroscedasticity.

Chapter 3

Sparse Convoluted Rank Regression in High Dimensions

Wang et al. (2020, JASA) studied the high-dimensional sparse penalized rank regression and established its nice theoretical properties. Compared with the least squares, rank regression can have a substantial gain in estimation efficiency while maintaining a minimal relative efficiency of 86.4%. However, the computation of penalized rank regression can be very challenging for high-dimensional data, due to the highly nonsmooth rank regression loss. In this work we view the rank regression loss as a non-smooth empirical counterpart of a population level quantity, and a smooth empirical counterpart is derived by substituting a kernel density estimator for the true distribution in the expectation calculation. This view leads to the convoluted rank regression loss and consequently the sparse penalized convoluted rank regression (CRR) for high-dimensional data. Under the same key assumptions for sparse rank regression, we establish the rate of convergence of the ℓ_1-penalized CRR for a tuning free penalization parameter and prove the strong oracle property of the folded concave penalized CRR. We further propose a high-dimensional Bayesian information criterion for selecting the penalization parameter in folded concave penalized CRR and prove its selection consistency. We derive an efficient algorithm for solving sparse convoluted rank regression that scales well with high dimensions. Numerical examples

demonstrate the promising performance of the sparse convoluted rank regression over the sparse rank regression. Our theoretical and numerical results suggest that sparse convoluted rank regression enjoys the best of both sparse least squares regression and sparse rank regression.

3.1 Introduction

Over the past two decades, there has been a surge of literature on high dimensional statistics. We refer to Bühlmann and Van De Geer (2011) and Fan et al. (2020) for a comprehensive review of the existing work on this topic. In particular, many penalization methods have been proposed for high-dimensional regression, including ℓ_1-penalized regression (Tibshirani, 1996), the Dantzig selector (Candes and Tao, 2007), concave-penalized regression (Fan and Li, 2001), among others. These techniques are also applicable in other statistical models. The penalized least squares method is at the center of the stage in terms of theoretical and computational developments in high-dimensional regression. The theoretical setup typically assumes that the true model is a linear regression model with homoscedastic variance. As long as the error is sub-Gaussian, the penalized least squares estimator enjoys nice theoretical guarantees even if the number of covariates grows at a nearly exponential rate with sample size.

An approach for achieving a higher efficiency is the penalized Wilcoxon rank regression (or rank regression for short). Wilcoxon rank regression is well studied in the classical robust nonparametric statistics (Hettmansperger and McKean, 2010). The penalized rank regression was studied by several authors (Wang and Li, 2009) for the low dimension setting. Recently, penalized rank regression in high dimensional setting was fully investigated in Wang et al. (2020). The penalized rank regression

solves the estimator of regression coefficient through minimizing

$$\frac{1}{n(n-1)} \sum \sum_{i \neq j} |(y_i - \mathbf{x}_i^{\mathrm{T}} \beta) - (y_j - \mathbf{x}_j^{\mathrm{T}} \beta)| + p_\lambda(|\beta_j|) \tag{3.1.1}$$

over $\beta \in \mathbb{R}^p$, where $p_\lambda(\cdot)$ is some penalty function. The penalized rank regression has several advantages compared with the penalized least squares regression. First, penalized rank regression is shown to possess better efficiency than the least squares approach when error has a heavy-tailed distribution, while maintaining a good relative efficiency when error is normally distributed. Second, penalized rank regression enjoys tuning free property, which means the theoretical correct tuning parameter can be easily estimated from the dataset without any cross validation. Although tuning free property can be also obtained through other methodologies such as the square-root Lasso (Belloni, Chernozhukov and Li, 2012) and penalized quantile regression (Wang, Wu and Li, 2012), these methods do not necessarily have the first aforementioned efficiency property.

Although penalized rank regression has the aforementioned nice theoretical advantages, it can be difficult to use in practice due to computational challenges, especially when the number of covariates in the dataset is very large. It is known that high dimensional penalized regression with a smooth loss function can be efficiently computed by cyclical coordinate descent algorithm (Friedman, Hastie and Tibshirani, 2010). However, the loss function in penalized rank regression is highly non-smooth. In principle, coordinate descent may fail to deliver the right solution due to the non-smoothness of the objective function. A similar problem is quantile regression in which the check loss is nonsmooth. The computation of quantile regression is done by using interior point algorithms. One way of computing the penalized rank regression is to transform it into linear programming and then apply the interior point algorithm. However, the interior point algorithm does not scale well with high di-

mensions. Gu et al. (2018) developed an alternating directional method of multipliers for computing the high-dimensional quantile regression. Computationally speaking, sparse penalized rank regression is more challenging than penalized quantile regression. The interior point algorithm is not a suitable choice for solving high-dimensional sparse rank regression.

It is natural to ask whether the aforementioned good theoretical properties possessed by rank regression can be shared with good computational efficiency for practical applications. If one focuses on ((3.1.1)), then the only solution is to develop an efficient algorithm for solving ((3.1.1)) exactly for large p problems. Recently, Fernandes et al. (2021) proposed an interesting smoothing technique for solving quantile regression with statistical guarantees. They showed that the smoothing quantile regression can even have a smaller mean squared error than the exact quantile regression for estimating the same conditional quantile function. Their work is more interesting from a statistical perspective, because fast computation for the quantile regression has already been solved in Gu et al. (2018). Their work motivated us to develop a smooth version of sparse rank regression from the statistical perspective, as opposed to trying to solve it exactly. For easy discussion, we call the first term in ((3.1.1)) the rank regression loss, although it is not like the empirical average of a loss function in empirical risk minimization. If we could replace the rank regression loss in ((3.1.1)) with a smooth loss such that the resulting estimator still has the nice theoretical properties of sparse rank regression, then we should focus on solving the smooth problem instead of ((3.1.1)). This is what Fernandes et al. (2021) did for quantile regression. To this end, we consider the expectation of the rank regression loss with respect to the true distribution. The rank regression loss is viewed as the expectation of a random variable with respect to some empirical distribution assigning uniform discrete probability to each observed realization. If we estimate the true distribution by using a smoothed kernel density estimator, then we can take the expectation of

the same random variable with respect to the smoothed kernel density estimator. The resulting quantity is shown to be smooth, convex and has a Lipschitz continuous derivative. We name it *convoluted rank loss* because the kernel density estimator has a convolution interpretation. We then replace the rank regression loss in ((3.1.1)) with the convoluted rank loss and the resulting estimator is called sparse convoluted rank regression. By its convexity and smoothness, the sparse convoluted rank regression can be efficiently solved by using the generalized coordinate descent algorithm (Yang and Zou, 2013).

We systematically study the theoretical properties of the sparse convoluted rank regression. The goal is to show that it maintains all the essential theoretical properties of rank regression. Specifically, we first establish the rate of convergence of the ℓ_1-penalized convoluted rank regression in ultra-high dimensions without assuming a strong moment condition on the error and the ℓ_1-penalized convoluted regression is also shown to enjoy the asymptotic tuning free property. Second, we analyze the folded concave penalized convoluted rank regression and establish its strong oracle property without imposing strong moment conditions on the error. The folded concave penalized regression involves a tuning parameter. We thus further propose a high dimensional Bayesian information criterion (HBIC) and establish its consistency for the selection of the theoretically optimal tuning parameter.

The rest of this paper is organized as follows. In Section 3.2, we introduce convoluted rank regression loss and the sparse convoluted rank regression estimator. In Section 3.3, we present the theoretical justifications for the proposed estimators. We also present the HBIC criterion and its theoretical results. In Section 3.4, we derive an efficient algorithm for solving sparse convoluted rank regression for high-dimensional data. In Section 3.5, we use simulations and a real data example to compare sparse convoluted rank regression and sparse rank regression. The technical proofs are given in the supplement file.

3.2 Convoluted Rank Regression

In this section we present the main idea that leads to the convoluted rank regression loss and the sparse convoluted rank regression.

3.2.1 Notation and definitions

We begin with some necessary definitions. For an arbitrary index set $\mathbf{A} \subset \{1, \ldots, p\}$, any vector $\mathbf{c} = (c_1, \ldots, c_p)$ and any $n \times p$ matrix $\mathbf{U}$, let $\mathbf{c}_\mathbf{A} = (c_i, i \in \mathbf{A})$, and let $\mathbf{U}_\mathbf{A}$ be the submatrix with columns of $\mathbf{U}$ whose indices are in $\mathbf{A}$. The complement of an index set $\mathbf{A}$ is denoted as $\mathbf{A}^{\mathrm{c}} = \{1, \ldots, p\} \setminus \mathbf{A}$. For any finite set $\mathbf{B}$, let $|\mathbf{B}|$ be the number of elements in $\mathbf{B}$. For a vector $\mathbf{c} = (c_1, \ldots, c_p)^{\mathrm{T}}$ and $q \in [1, \infty)$, let $\|\mathbf{c}\|_q = (\sum_{j=1}^p |c_j|^q)^{\frac{1}{q}}$ be its ℓ_q norm, let $\|\mathbf{c}\|_\infty$ (or $\|\mathbf{c}\|_{\max}$) $= \max_j |c_j|$ be its ℓ_∞ norm, let $\|\mathbf{c}\|_0 = |\{j : c_j \neq 0\}|$ be its ℓ_0 norm, and let $\|\mathbf{c}\|_{\min} = \min_j |c_j|$ be its minimum absolute value. For a matrix $\mathbf{M}$, let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ be its eigenvalue with smallest absolute value and largest absolute value, respectively. This is the common notation for eigenvalues of a matrix, and $\lambda_{\min}$, $\lambda_{\max}$ should not be confused with the penalization parameter used in a penalty function. For any matrix $\mathbf{G}$, let $\|\mathbf{G}\| = \sqrt{\lambda_{\max}(\mathbf{G}^{\mathrm{T}}\mathbf{G})}$ be its spectral norm. In particular, for a vector $\mathbf{c}$, $\|\mathbf{c}\| = \|\mathbf{c}\|_2$. For $a, b \in \mathbb{R}$, let $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For a sequence $\{a_n\}$ and another nonnegative sequence $\{b_n\}$, we write $a_n = O(b_n)$ if there exists a constant $c > 0$ such that $|a_n| \leq cb_n$ for all $n \geq 1$. And we write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. Also, we use $a_n = o(b_n)$, or $a_n \ll b_n$, to represent $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$. We write $b_n \gg a_n$ if $a_n \ll b_n$. Let $(\Omega, \mathcal{G}, \mathrm{P})$ be a probability space on which all the random variables that appear in this paper are defined. Let $\mathbb{E}[\cdot]$ be the expectation with respect to the probability measure P. For a sequence of random variables $\{Z_n\}_{n\geq 1}$, we write $Z_n = O_p(1)$ if $\lim_{M\to\infty} \limsup_{n\to\infty} \mathrm{P}(|Z_n| > M) = 0$, and we write $Z_n = o_p(1)$ if $\lim_{n\to\infty} \mathrm{P}(|Z_n| > \epsilon) = 0, \forall \epsilon > 0$. For two sequences of random

variables Z_n and Z'_n, we write $Z_n = O_p(Z'_n)$ if $\frac{Z_n}{Z'_n} = O_p(1)$, and we write $Z_n = o_p(Z'_n)$ if $\frac{Z_n}{Z'_n} = o_p(1)$.

3.2.2 Canonical Convoluted Rank Regression

Suppose we have the observed data $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ where $y_i \in \mathbb{R}$ is the response value and $\mathbf{x}_i \in \mathbb{R}^p$ is the p-dimensional covariate vector for the ith subject. Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p) \in \mathbb{R}^{n\times p}$ be the design matrix, with $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ containing observations for the jth variable, $j = 1, \ldots, p$. The ith row of $\mathbf{X}$ can be written as $\mathbf{x}_i^{\mathrm{T}}$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. Let $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ be the n-dimensional response vector. For the sake of brevity, we adopt the fixed design setting in the sequel, although our methodology can also be justified under the random design setting. Assume that the data are generated from the following linear regression model $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$,

$$y_i = \mathbf{x}_i^{\mathrm{T}}\beta^* + \epsilon_i, \tag{3.2.1}$$

where $\{\epsilon_i\}_{i=1}^n$ are i.i.d. random errors, $\beta^* \in \mathbb{R}^p$ is the unknown vector to be estimated. Note that we do not assume the errors in ((3.2.1)) have mean zero. Consequently, the intercept can be absorbed into the error term.

The canonical rank regression (Jaeckel, 1972; Hettmansperger and McKean, 2010) in the fixed dimension setting proposes to estimate β^* through

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{n(n-1)} \sum\sum_{i\neq j} |(y_i - \mathbf{x}_i^{\mathrm{T}}\beta) - (y_j - \mathbf{x}_j^{\mathrm{T}}\beta)|. \tag{3.2.2}$$

Compared with the standard least squares method, the rank regression estimator of β^* can have arbitrarily high relative efficiency when error distribution is heavy-tailed, while having at least 86.4% asymptotic relative efficiency under arbitrary symmetric error distribution with finite Fisher information (Hettmansperger and McKean, 2010).

For each (i, j) pair, define $\{\zeta_{ij}\}_{i\neq j}$ with $\zeta_{ij} = (y_i - \mathbf{x}_i^{\mathrm{T}}\beta) - (y_j - \mathbf{x}_j^{\mathrm{T}}\beta)$. For the discussion in this part, we treat $(y_i, \mathbf{x}_i)_{i=1}^n$ as independent and identically distributed. Although ζ_{ij}s are not independent, they still follow an identical distribution. For any β, let $F(t, \beta)$ denote its cumulative distribution function. After taking the expectation of the objective function in ((3.2.2)) with respect to the true distribution of ζ_{ij}, the population level objective function is $\int_{-\infty}^{\infty} |t| \mathrm{d}F(t, \beta)$. Then, we can view the objective function in ((3.2.2)) as $\int_{-\infty}^{\infty} |t| \mathrm{d}\hat{F}(t, \beta)$, where $\hat{F}(t, \beta) = \frac{1}{n(n-1)} \sum\sum_{i\neq j} 1_{\{\zeta_{ij} \leq t\}}$ is the estimated cumulative distribution function for $\{\zeta_{ij}\}_{i\neq j}$. Since the estimated CDF is discontinuous, it causes the objective function in ((3.2.2)) to have the same degree of smoothness as the absolute value function. This statistical view of the objective function in rank regression suggests us to use an alternative estimator for the distribution of ζ_{ij}. If we use a smooth estimator $\tilde{F}(t, \beta)$, then $\int_{-\infty}^{\infty} |t| \mathrm{d}\tilde{F}(t, \beta)$ can be the new objective function and become smooth.

Specifically, we consider using the kernel density estimator

$$\tilde{F}(t, \beta) = \int_{-\infty}^{t} \frac{1}{n(n-1)} \sum\sum_{i\neq j} \frac{1}{h} K(\frac{v - \zeta_{ij}}{h}) \mathrm{d}v$$

with some kernel function $K : \mathbb{R} \to [0, \infty)$ satisfying $K(-t) = K(t)$, $\int_{-\infty}^{\infty} K(t) \mathrm{d}t = 1$, and $h > 0$. Replacing $\hat{F}$ with $\tilde{F}$, we obtain a new objective function

$$\begin{aligned} \int_{-\infty}^{\infty} |t| \mathrm{d}\tilde{F}(t, \beta) &= \frac{1}{n(n-1)} \sum\sum_{i\neq j} \int_{-\infty}^{\infty} \frac{1}{h} K(\frac{\zeta_{ij} - t}{h}) |t| \mathrm{d}t \\ &\triangleq \frac{1}{n(n-1)} \sum\sum_{i\neq j} L_h\big((y_i - \mathbf{x}_i^{\mathrm{T}}\beta) - (y_j - \mathbf{x}_j^{\mathrm{T}}\beta)\big), \end{aligned}$$

where $L_h(u) = \int_{-\infty}^{\infty} |u - v| \frac{1}{h} K(\frac{v}{h}) \mathrm{d}v$. It is worth noting that $L_h(\cdot)$ is a smooth convex function. The function L_h satisfies the relation $L_h = L * K_h$, where $L(u) = |u|$, $K_h(u) = \frac{1}{h} K(\frac{u}{h})$ and "$*$" stands for convolution.

Thus, in the fixed dimension setting, we propose the canonical convoluted rank regression

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{n(n-1)} \sum\sum_{i\neq j} L_h\big(y_i - \mathbf{x}_i^{\mathrm{T}}\beta) - (y_j - \mathbf{x}_j^{\mathrm{T}}\beta)\big). \tag{3.2.3}$$

It turns out that the rank regression ((3.2.2)) and the convoluted rank regression ((3.2.3)) shares interesting connection in the population sense. In fact, let $(y, \mathbf{x})$ and $(y', \mathbf{x}')$ be i.i.d. random vectors with continuous distribution in $\mathbb{R}^{p+1}$ satisfying $y = \mathbf{x}^{\mathrm{T}}\beta^* + \epsilon$ and $y' = \mathbf{x}'^{\mathrm{T}}\beta^* + \epsilon'$, where ϵ is independent from $\mathbf{x}$, and ϵ' is independent from $\mathbf{x}'$. For rank regression, it is well known that the minimizer of the population version of its loss function, i.e. $\arg\min_{\beta\in\mathbb{R}^p} \mathbb{E}[|\epsilon - \epsilon' - (\mathbf{x} - \mathbf{x}')^{\mathrm{T}}(\beta - \beta^*)|]$, is exactly the same as β^*, the true regression coefficients. This simple fact justifies that rank regression is valid in the population sense, which is necessary in order for its sample version to aim at estimating the true regression coefficients. One may naturally ask if the population version of ((3.2.3)) also has such property. Let $\beta_h^* = \arg\min_{\beta\in\mathbb{R}^p} \mathbb{E}[L_h(y - y' - (\mathbf{x} - \mathbf{x}')^{\mathrm{T}}\beta)]$. We have the following theorem, stating that smoothing via convolution does not incur any bias at all in the population sense.

Theorem 3 For any $h > 0$ and any kernel $K(\cdot)$ satisfying $\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1$ and $K(u) = K(-u), \forall u \in \mathbb{R}$, we have $\beta_h^* = \beta^*$. □

Remark 7 Note that the smoothing quantile regression (Fernandes et al., 2021) does not have the good property of zero smoothing bias as shown in Theorem 3. In fact, the proof of Theorem 3 crucially relies on the fact that the distributions of $\epsilon - \epsilon'$ and $\mathbf{x} - \mathbf{x}'$ are symmetric about zero, which can only be taken advantage of given the special form of rank regression. □

3.2.3 Sparse Convoluted Rank Regression

When p is large, we consider designing the estimator under a sparsity assumption that β^* in the data generating model has many zero components. Let $\mathbb{A} = \{j : \beta_j^* \neq 0\}$ be the support set of β^*, i.e., the set of indices of the important covariates. Let $s = |\mathbb{A}|$. Throughout this paper, we allow $p = p_n$ and $s = s_n$ to diverge with n, and we assume $s_n \geq 1$ and p_n goes to infinity as n goes to infinity. For convenience, we still use p and s to represent these quantities since no confusion is caused. In ultra-high dimensions, the dimension p is allowed to increase exponentially with the sample size n, and we assume that s is relatively of smaller order compared to n. Otherwise, no consistent estimator is possible.

To estimate β^*, we propose the sparse Convoluted Rank Regression (CRR) by solving

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_h(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \beta) + \sum_{j=1}^{p} p_\lambda(|\beta_j|).$$

Here $p_\lambda(\cdot)$ is some sparsity-inducing penalty function with a tuning parameter $\lambda > 0$, $L_h(u) = \int_{-\infty}^{\infty} |u - v| \frac{1}{h} K(\frac{v}{h}) \mathrm{d}v$, where $K : \mathbb{R} \to [0, \infty)$ is a kernel function satisfying $\int_{-\infty}^{\infty} K(u) \mathrm{d}u = 1$ and $K(u) = K(-u), \forall u$, and $h > 0$ is a constant.

Remark 8 There can be a lot of choices for the kernel function $K(\cdot)$ satisfying the conditions in our theory presented in section 3. In the numerical studies of this work, we focus on the Epanechnikov kernel $K(u) = \frac{3}{4}(1 - u^2) I(-1 \leq u \leq 1)$ for illustration purposes, where $I(\cdot)$ is the indicator function.

Intuitively, h should be small such that the sparse convoluted rank regression is very close to the sparse rank regression. As suggested by the theoretical results in Section 3, $h = O(1)$ is sufficient for our method to achieve optimal rate and oracle property. According to density estimator, the optimal rate for h is $O(n^{-1/5})$. So, we

use $h = 2.5n^{-1/5}$ as the default value in our implementation. □

3.3 Theoretical Justifications for Sparse CRR

In this section we study the theoretical properties of the ℓ_1-penalized convoluted rank regression (CRR) and the folded concave penalized CRR under the same key regularity conditions for the rank regression in Wang et al. (2020).

3.3.1 ℓ_1-penalized CRR

For a tuning parameter $\lambda_0 > 0$, we define the ℓ_1-penalized CRR estimator as

$$\tilde{\beta}^{\lambda_0} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_h(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\beta) + \lambda_0 \sum_{j=1}^{p} |\beta_j|.$$

We now state the assumptions needed throughout this paper. We make the following assumptions for the kernel function $K(\cdot)$.

Assumption 1 $K : \mathbb{R} \to [0, \infty)$ is a function satisfying the following properties: (i), $K(-t) = K(t)$, $\forall t \in \mathbb{R}$; (ii), $\exists \delta_0 > 0$ s.t. $\kappa_l := \inf_{t \in [-\delta_0, \delta_0]} K(t) > 0$; (iii), $\int_{-\infty}^{\infty} K(t)\,\mathrm{d}t = 1$; (iv), $\kappa_1 := \int_{-\infty}^{\infty} |t| K(t)\,\mathrm{d}t < \infty$. □

For the error distribution, we impose the following assumption.

Assumption 2 The errors $\{\epsilon_i\}_{i=1}^n$ are independent and identically distributed with density function $f(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}$. Besides, let $\varsigma_{ij} = \epsilon_i - \epsilon_j, 1 \leq i \neq j \leq n$. Let $g(\cdot)$ denote the probability density function of ς_{ij}, we assume $\sup_{t \in \mathbb{R}} g(t) = \mu_0 < \infty$. Meanwhile, there exist positive constants δ_1, μ_1 such that $g(t) \geq \mu_1, \forall t \in [-\delta_1, \delta_1]$. □

For any index set $\mathbf{A} \subset \{1, \ldots, p\}$, let $\mathcal{S}_{\mathbf{A}} \coloneqq \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}_{\mathbf{A}^c}\|_1 \leq 3\|\mathbf{u}_{\mathbf{A}}\|_1 \neq 0\}$. We also impose the following conditions on the design matrix.

Assumption 3 There exists a constant $M > 0$ such that $\max_{1\leq i\leq n, 1\leq j\leq p} |x_{ij}| \leq M$. Also, the covariates are centered, i.e. $\sum_{i=1}^n x_{ij} = 0, \forall j = 1, \ldots, p$. □

Assumption 4 There exists a constant $\rho > 0$ such that $\min_{\mathbf{u}\in\mathcal{S}_{\mathbb{A}}} \frac{\|\mathbf{X}\mathbf{u}\|_2^2}{n\|\mathbf{u}\|_2^2} \geq \rho$. In particular, this implies $\lambda_{\min}(\frac{\mathbf{X}_{\mathbb{A}}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}}}{n}) \geq \rho$. □

Assumption 5 is common in the fixed design case. It can be relaxed with M increasing with n at a suitable rate, without much difficulty. We keep it here for the sake of brevity. We can center the design matrix when estimating the β^* vector because centering only affects the intercept part which is a nuisance parameter in our method as well as in rank regression. Assumption 6, which is known as the restricted eigenvalue condition (RE), is needed to establish ℓ_2-type error bound for ℓ_1-penalized estimator. It is a commonly used assumption in the literature (Bühlmann and Van De Geer, 2011; Fan et al., 2020).

Theorem 4 Assume assumptions 1-6 hold, and $s = o(\sqrt{\frac{n}{\log p}})$. Let $0 < \lambda_0 = c_0\sqrt{\frac{\log p}{n}}$ with $8\sqrt{2}M < c_0 = O(1)$, and let $0 < h = O(1)$. Then the ℓ_1-penalized CRR estimator $\tilde{\beta}^{\lambda_0}$ satisfies

$$\|\tilde{\beta}^{\lambda_0} - \beta^*\|_2 \leq \frac{192M + 4c_0}{\mu_2\rho}\sqrt{\frac{s\log p}{n}}$$

with probability at least $1 - 2p^{-\left(\frac{c_0^2}{128M^2}-1\right)} - 2p^{-2}$, where $\mu_2 \coloneqq \kappa_l\mu_1(2\delta_0 \wedge \frac{\delta_1}{h})$. □

Notice that the probabilistic bound in Theorem 1 does not depend on unknown quantities, since with the design matrix at hand, M and p are both available. This

means that in principle, the λ_0 in ℓ_1-penalized CRR estimator is tuning-free. As long as c_0 is a constant which is larger than $8\sqrt{2}M$, the probabilistic lower bound in Theorem 1 converges to 1, and as a result we have $\|\tilde{\beta}^{\lambda_0} - \beta^*\|_2 = O_p(\sqrt{\frac{s\log p}{n}})$, which means the ℓ_1-penalized CRR estimator achieves the near-optimal rate.

3.3.2 Folded concave penalized CRR

It has been well established in the literature that folded concave penalized estimators can enjoy strong oracle property. We prove the same is true for convoluted rank regression. Define

$$\hat{\beta}^{\text{ora}} := \underset{\beta\in\mathbb{R}^p:\beta_{\mathbb{A}^c}=\mathbf{0}}{\arg\min} \frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j\neq i} L_h(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\beta) \tag{3.3.1}$$

as the CRR oracle estimator. It can be directly verified that $\hat{\beta}^{\text{ora}}$ exists due to the convexity of $L_h(\cdot)$, assumption 5 and assumption 6. We establish the following property for the oracle estimator.

Theorem 5 Assume assumptions 1-6 hold, $s = o(\sqrt{n})$ and $h = O(1)$. Then we have $\|\hat{\beta}^{\text{ora}} - \beta^*\|_2 = O_p\left(\sqrt{\frac{s}{n}}\right)$. □

Remark 9 In the case where $\hat{\beta}^{\text{ora}}$ is not unique, one may take any solution to ((3.3.1)), and our theory about CRR oracle estimator still holds. □

We now propose the concave penalized convoluted rank regression. It solves the following problem:

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j\neq i} L_h(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\beta) + \sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{3.3.2}$$

For the choice of $p_\lambda(\cdot)$, we adopt the folded concave penalty (Fan et al., 2014b), i.e. $p_\lambda(\cdot)$ is a function defined on $(-\infty, \infty)$ satisfying: (i), $p_\lambda(-z) = p_\lambda(z)$; (ii), $p_\lambda(z)$ is increasing and concave in $z \in [0, \infty)$, and $p_\lambda(0) = 0$; (iii), $p_\lambda(z)$ is differentiable in $z \in (0, \infty)$, and $p'_\lambda(0) := p'_\lambda(0+) \geq a_1\lambda$; (iv), $p'_\lambda(z) \geq a_1\lambda$ for $z \in (0, a_2\lambda]$; (v) $p'_\lambda(z) = 0$ for $z \in [a\lambda, \infty)$ with some pre-specified constant $a > a_2$. Here a_1 and a_2 are two fixed positive constants. Special cases of folded concave penalty are SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). The SCAD penalty has the form

$$\begin{aligned} p_\lambda(|t|) =& \lambda|t| I(0 \leq |t| < \lambda) + \frac{a\lambda|t| - (t^2 + \lambda^2)/2}{a-1} I(\lambda \leq |t| \leq a\lambda) \\ &+ \frac{(a+1)\lambda^2}{2} I(|t| > a\lambda), \text{ for some } a > 2, \end{aligned}$$

which corresponds to $a_1 = a_2 = 1$. The MCP penalty function is defined as

$$p_\lambda(|t|) = \lambda \left(|t| - \frac{t^2}{2a\lambda} \right) I(0 \leq |t| < a\lambda) + \frac{a\lambda^2}{2} I(|t| \geq a\lambda), \text{ for some } a > 1,$$

which corresponds to $a_1 = 1 - \frac{1}{a}, a_2 = 1$.

We adopt the local linear approximation (LLA) (Zou and Li, 2008) algorithm to solve ((3.3.2)). The LLA algorithm iteratively solves

$$\begin{aligned} \hat{\beta}^{(k+1)} =& \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_h(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \beta) + \sum_{j=1}^{p} p'_\lambda \left(|\beta_j^{(k)}| \right) |\beta_j|, \\ & k = 0, 1, 2, \ldots, \end{aligned} \tag{3.3.3}$$

where $\hat{\beta}^{(0)}$ is some initial estimator. We use $\hat{\beta}^\lambda$ to denote the folded concave penalized CRR estimator computed by the LLA algorithm, with tuning parameter λ. Below we establish theory for the folded concave penalized CRR estimator.

Theorem 6 Let the conditions of Theorem 1 and Theorem 5 hold. Assume that

$\sup_{t\in\mathbb{R}} K(t) = \kappa_u < \infty$. Let $a_0 = \min\{1, a_2\}$ where a_2 is the constant associated with the folded-concave penalty function. Choose the tuning parameter so that $\min_{j\in\mathbb{A}} |\beta_j^*| > (a+1)\lambda$.

(i) Suppose $s = o(\log p)$ and $h \gg \sqrt{\frac{s}{n\log p}}$. Let the tuning parameter be chosen as $\lambda = c_1\sqrt{\frac{s\log p}{n}}$ such that $\frac{192M+4c_0}{a_0\mu_2\rho} \vee \frac{32\sqrt{2}M}{a_1} < c_1 = O(1)$, where c_0 is defined in Theorem 1. Then the LLA algorithm in ((3.3.3)) initialized by $\hat{\beta}^{(0)} = \tilde{\beta}^{\lambda_0}$, with λ_0 being defined in Theorem 1, converges to $\hat{\beta}^{\text{ora}}$ in two iterations with probability converging to 1 as $n \to \infty$.

(ii) Consider the SCAD or MCP as the penalty function. Suppose $s = o(\sqrt{\log p})$ and $h \gg \frac{s}{\sqrt{n\log p}}$. Let the tuning parameter be chosen as $\lambda = c_1\sqrt{\frac{\log p}{n}}$ such that $\frac{(192M+4c_1)\sqrt{s}}{a_0\mu_2\rho} \vee \frac{32\sqrt{2}M}{a_1} \vee 8\sqrt{2}M < c_1 = O(1)$. Then the LLA algorithm in ((3.3.3)) initialized by $\hat{\beta}^{(0)} = \mathbf{0}$ converges to $\hat{\beta}^{\text{ora}}$ in three iterations with probability converging to 1 as $n \to \infty$. □

Theorem 6 shows that the folded concave penalized CRR estimator equals to the oracle estimator with overwhelming probability, which is typically referred to as strong oracle property. It means that our estimator can perform as well as if the true set of important covariates was given.

Remark 10 In Theorems 1 and 5, we only require $h = O(1)$, and in Theorem 6, $\frac{1}{\sqrt{n}} \ll h = O(1)$ is sufficient. These are weaker conditions on the smoothing bandwidth h than that is required for smoothing quantile regression (Fernandes et al., 2021) in which h should satisfy $(\frac{n}{\log n})^{-1/3} \ll h = o(1)$. Again, this is a consequence of the delicate form of rank regression which makes important first order terms vanish, as can be seen from our theoretical proofs. □

3.3.3 Consistent tuning parameter selection

For the folded concave penalization, Theorem 6 guarantees that there exists a good tuning parameter in principle. Since the tuning parameter depends on unknown quantities, a data-driven approach is needed to specify the tuning parameter in practice. Motivated by Wang et al. (2013), we propose a modified high dimensional Bayesian information criteria, defined as

$$\mathrm{HBIC}(\lambda) = \log\left(\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_h\big(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\hat{\beta}^{\lambda}\big)\right) + |M_\lambda|\frac{C_n \log p}{n},$$

where $M_\lambda \coloneqq \{j : \hat{\beta}_j^{\lambda} \neq 0\}$, and the choice of C_n is discussed in Theorem 7. The corresponding tuning parameter for the folded concave penalty is chosen by minimizing the proposed HBIC.

Theorem 7 Let $\hat{\lambda} = \arg\min_{\lambda\in\Lambda} \mathrm{HBIC}(\lambda)$, where $\Lambda = \{\lambda > 0 : |M_\lambda| \leq K_n\}$, and $K_n > s$ is allowed to diverge to infinity. Under the conditions of Theorem 6, assume that $\mathbb{E}[|\varsigma_{ij}|] < \infty$, $\phi \coloneqq \min_{|\mathbb{S}|\leq 2K_n} \lambda_{\min}(\frac{\mathbf{X}_{\mathbb{S}}^{\mathrm{T}}\mathbf{X}_{\mathbb{S}}}{n}) > 0$. If $\sqrt{\frac{C_n\sqrt{s}\log p}{n}} \vee \frac{C_n \log p\sqrt{sK_n}}{n} = o(\|\beta^*_{\mathbb{A}}\|_{\min})$, $\frac{C_n s\log p}{n} = o(1)$ and $K_n = o\big(\sqrt{\frac{n}{\log p}} \wedge \sqrt{C_n}\big)$, then we have $\mathrm{P}(M_{\hat{\lambda}} = \mathbb{A}) \to 1$ as $n \to \infty$. □

Remark 11 The condition $\min_{|\mathbb{S}|\leq 2K_n} \lambda_{\min}(\frac{\mathbf{X}_{\mathbb{S}}^{\mathrm{T}}\mathbf{X}_{\mathbb{S}}}{n}) > 0$ in Theorem 7 is known as the sparse Riesz condition and is widely used in literature on high dimensional statistics (Zhang and Huang, 2008). In our numerical studies, the sequence C_n is chosen such that $C_n \asymp \log\log n$. This is the same choice as in the HBIC for the penalized rank regression (Wang et al., 2020). □

Theorem 7 shows that with proposed HBIC, our method can exactly identify the important variables with probability approaching to 1. Unlike cross validation, the

HBIC criterion does not require sample splitting or repeated evaluation of the test error on each sub-dataset. As a result, our method requires no extra computation for tuning.

3.4 Computation

We have shown that we need to solve the folded concave penalized CRR by running the LLA iteration 2-3 times. In each LLA iteration, we need to solve a weighted ℓ_1-penalized CRR problem. In this section, we develop an efficient algorithm for computing the solution path of a weighted ℓ_1-penalized CRR.

Consider the following "weighted" ℓ_1-penalized CRR problem:

$$\arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i} L_h(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\beta) + \sum_{k=1}^{p} w_k|\beta_k|, \tag{3.4.1}$$

where each $w_k \geq 0$. In contrast to the sparse rank regression, the density convolution gives a smooth loss function L_h. To see this, recall $L_h(u) = \int_{-\infty}^{\infty} |u-v|\frac{1}{h}K(\frac{v}{h})\mathrm{d}v, u \in \mathbb{R}$, and a direct calculation gives $L_h'(u) = 2\int_{-\infty}^{u} \frac{1}{h}K(\frac{v}{h})\mathrm{d}v - 1$ and $L_h''(u) = \frac{2}{h}K(\frac{u}{h})$, $\forall u \in \mathbb{R}$. We thus establish some basic properties of $L_h(\cdot)$.

Lemma 1 Under assumption 1, for any $t_1, t_2, t \in \mathbb{R}$, we have $L_h'(-t) = -L_h'(t)$ and $|L_h(t_1) - L_h(t_2)| \leq |t_1 - t_2|$. If we use a kernel such that $\sup_{t\in\mathbb{R}} K(t) = \kappa_u < \infty$, then $|L_h'(t_1) - L_h'(t_2)| \leq \frac{2}{h}\kappa_u|t_1 - t_2|$. □

Therefore, the objective function in problem ((3.4.1)) is the summation of a convex and smooth loss function and a convex and separable penalty term. It turns out that a coordinate descent-type algorithm usually works well in this situation (Tseng, 2001).

In a coordinate-wise manner, suppose we have updated the coordinates $\beta_1, \beta_2, \ldots, \beta_{k-1}$ and we now need to update β_k. Denote by $\tilde{\boldsymbol{\beta}}$ the current solution and let $v_{ij} =$

$y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\tilde{\beta}$. The standard coordinate descent algorithm cyclically updates β_k by minimizing

$$F(\beta_k|\tilde{\beta}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_h(v_{ij} - (x_{ik} - x_{jk})(\beta_k - \tilde{\beta}_k)) + w_k|\beta_k|.$$

We observe that minimizing the above function does not have a close-form solution, so we consider a generalized coordinate descent algorithm (Yang and Zou, 2013). The idea is to perform a majorization-minimization update rather than directly minimize $F(\beta_k|\tilde{\beta})$. Specifically, we need to find a quadratic function G such that $F(\beta_k|\tilde{\beta}) = G(\beta_k|\tilde{\beta})$ and $F(\gamma|\tilde{\beta}) < G(\gamma|\tilde{\beta})$ for any $\gamma \neq \beta_k$.

From the last inequality of Lemma 1, we can obtain a quadratic majorization condition for CRR:

$$L_h(t_1) < L_h(t_2) + L_h'(t_2)(t_1 - t_2) + \frac{\kappa_u}{h}(t_1 - t_2)^2,$$

for $t_1 \neq t_2$. For each pair of $i \neq j$, by letting $t_1 = v_{ij} - (x_{ik} - x_{jk})(\beta_k - \tilde{\beta}_k)$ and $t_2 = v_{ij}$, we have the quadratic majorization function for $F(\beta_k|\tilde{\beta})$:

$$G(\beta_k|\tilde{\beta}) = \frac{\sum_{i=1}^{n} \sum_{j \neq i} L_h(v_{ij})}{n(n-1)} + a_k(\beta_k - \tilde{\beta}_k) + \frac{c_k \kappa_u}{h}(\beta_k - \tilde{\beta}_k)^2 + w_k|\beta_k|,$$

where $a_k = -\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_h'(v_{ij})(x_{ik} - x_{jk})$ and $c_k = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} (x_{ik} - x_{jk})^2$. Hence, we update β_k using the minimizer of $G\left(\beta_k|\tilde{\beta}\right)$:

$$\hat{\beta}_k = \operatorname{sgn}\left(\tilde{\beta}_k - \frac{h a_k}{2 c_k \kappa_u}\right)\left(\left|\tilde{\beta}_k - \frac{h a_k}{2 c_k \kappa_u}\right| - \frac{h w_k}{2 c_k \kappa_u}\right)_+.$$

Therefore, we solve problem ((3.4.1)) by cyclically performing the above update for each $k = 1, 2, \ldots, p$.

In our implementation, we directly compute the solution path problem ((3.4.1))

at a sequence of tuning parameters, $\lambda^{[1]}$, $\lambda^{[2]}, \ldots, \lambda^{[L]}$ instead of calling the algorithm L times for each individual parameter. We let

$$\lambda^{[1]} = \left\| \frac{1}{n(n-1)} \sum_{i \neq j} L_h' \left(y_i - y_j \right) \left(\mathbf{x}_i - \mathbf{x}_j \right) \right\|_\infty ,$$

which is the smallest penalization parameter to make all $\hat{\beta}_k = 0$. We then choose other λ-values such that they are uniformly distributed on a logarithm scale. In addition, we also employ the warm start and active set strategies to further accelerate the GCD algorithm; see details of these two strategies in Friedman, Hastie and Tibshirani (2010).

3.5 Numerical Examples

3.5.1 Simulation Study

In this section, we demonstrate the performance of the sparse convoluted rank regression in terms of estimation accuracy and variable selection using simulations. Because the most attractive property of rank regression is its efficiency argument, we focus on estimators with strong oracle properties such as the SCAD-penalized convoluted rank regression (denoted by CRR-SCAD) and SCAD-penalized rank regression (denoted by RR-SCAD). We use zero vector as the initial value in the LLA algorithm for computing CRR-SCAD, so that we do not have to compute the ℓ_1-penalized CRR in order to compute CRR-SCAD. We used the code from Wang et al. (2020) to compute RR-SCAD. In our numerical studies, we used Epanechnikov kernel as the density convolution kernel, $K(u) = \frac{3}{4}(1-u^2)I(-1 \leq u \leq 1)$, where $I(\cdot)$ is the indicator

function, and the loss function is

$$L_h(u) = \begin{cases} u, & u \geq h, \\ \frac{3u^2}{4h} - \frac{u^4}{8h^3} + \frac{3h}{8}, & -h < u < h, \\ -u, & u \leq -h. \end{cases}$$

Both the RR-SCAD and CRR-SCAD are tuned based on HBIC. For comparison, we also include the SCAD-penalized least squares (denoted by LS-SCAD) and tune it by its corresponding HBIC (Wang et al., 2013).

We consider a model $y = \mathbf{x}^{\mathrm{T}}\beta^{\star} + \epsilon$, where $\beta^{\star} = (\sqrt{3}, \sqrt{3}, \sqrt{3}, 0, 0, \ldots, 0) \in \mathbb{R}^p$, x is independently generated from $\mathrm{N}(0, \Sigma)$, and ϵ is independently generated from some certain distributions. We fix the sample size $n = 100$ and use the dimensions $p = 400$ and 3000. We consider four situations for the correlation structure of x: CS (0.2), CS (0.5), CS (0.8), and AR (0.5), where each CS (ρ) represents the compound symmetry correlation, i.e., $\Sigma_{i,j} = \rho$ if $i \neq j$ or 1 otherwise, and AR (ρ) indicates the autoregressive correlation, that is, $\Sigma = (\rho^{|i-j|})_{p\times p}$.

We compare these methods based on five criteria: ℓ_1 error ($\mathbb{E}\|\hat{\beta} - \beta^{\star}\|_1$), ℓ_2 error ($\mathbb{E}\|\hat{\beta} - \beta^{\star}\|_2$), model error, ($\mathbb{E}(\hat{\beta} - \beta^{\star})^{\mathrm{T}}\Sigma(\hat{\beta} - \beta^{\star})$), the number of false positive variables, and the number of false negative variables. All the quantities are averaged over 200 independent runs and the standard errors are provided.

Table 3.1 exhibits the simulation results when ϵ is from $\mathrm{N}(0, 1)$. In each situation, we use boldface to indicate the best performance that is evaluated based on each of the five criteria. When $p = 400$, we observe that the estimation accuracy of LS-SCAD and CRR-SCAD is similar and better than that of RR-SCAD; when $p = 3000$, the estimation accuracy of CRR-SCAD is the best. In addition, both LS-SCAD and CRR-SCAD have perfect performance in variable selection and RR-SCAD is the only method that makes mistakes. By comparing the performance of CRR-SCAD when

$p = 400$ and 3000, we see the performance of CRR-SCAD is less prone to the increase in p.

Table 3.2 summarizes the simulation results when ϵ is from a mixture normal distribution: $\epsilon \sim 0.95\mathrm{N}(0,1) + 0.05\mathrm{N}(0,100)$. From Table 3.2, we find that LS-SCAD fails to work well in this situation. For both $p = 400$ and $p = 3000$, RR-SCAD and CRR-SCAD perform similarly. Table 3.3 shows the results when $\epsilon/\sqrt{2}$ follows the t-distribution with four degrees of freedom. In all situations, CRR-SCAD performs better than the other two methods, in terms of both estimation accuracy and variable selection. When p is increased from 400 to 3000, CRR-SCAD suffers minimal impact, while RR-SCAD shows a significant loss in estimation accuracy.

3.5.2 A real data application

We illustrate our proposed method on a microarray gene expression data reported in (Scheetz et al., 2006). The dataset contains RNA expression levels of more than 31,000 gene probes from 120 twelve-week-old laboratory rats. Following Scheetz et al. (2006), we include 18,976 genes that have sufficient variation and are considered expressed in mammalian eyes. Among these genes, TRIM32 has genetic influences on a rare genetic disorder, the Bardet-Biedl syndrome (Chiang et al., 2006). Thus TRIM32 is chosen as the target variable and our goal is to identify the genes that are associated with TRIM32.

In our experiments, we randomly split the original data into a training set and a test set in the ratio 1:1. On the training set, we apply the fused Kolmogorov filter (Mai and Zou, 2015) to obtain a reduced set of 300 probes and retained the same 300 probes on the test set. We then fit SCAD-penalized least squares (SCAD), rank regression (RR-SCAD) and our convoluted rank regression (CRR-SCAD) on the training set and compute the prediction error on the test set. To illustrate the performance in higher dimensions, we repeat the same above procedure except that

Table 3.1. *Comparison of least-square regression with SCAD (LS-SCAD), rank regression with SCAD (RR-SCAD) and convoluted rank regression with SCAD (CRR-SCAD). The comparison criteria are ℓ_1 error, ℓ_2 error, model error (ME), number of false positive variables (FP) and number of false negative variables (FN). In each example, the best method evaluated based on each criterion is in boldface. All the quantities are averaged over 200 independent runs and standard errors are given in parentheses. In all the examples shown in this table, the error term in the data generating model is drawn from the standard normal distribution.*

		$p = 400$						$p = 3000$					
Σ	criterion	LS-SCAD		RR-SCAD		CRR-SCAD		LS-SCAD		RR-SCAD		CRR-SCAD	
CS (0.2)	ℓ_1	**0.31**	(0.01)	0.37	(0.01)	0.32	(0.01)	0.36	(0.01)	0.53	(0.01)	**0.33**	(0.01)
	ℓ_2	**0.18**	(0.00)	0.21	(0.01)	**0.18**	(0.00)	0.22	(0.01)	0.29	(0.01)	**0.19**	(0.00)
	ME	**0.03**	(0.00)	0.04	(0.00)	**0.03**	(0.00)	0.05	(0.00)	0.09	(0.00)	**0.04**	(0.00)
	FP	**0**	(0)	0	(0)	**0**	(0)	**0**	(0)	1	(0)	**0**	(0)
	FN	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)
CS (0.5)	ℓ_1	**0.36**	(0.01)	0.38	(0.01)	**0.36**	(0.01)	0.39	(0.01)	0.46	(0.01)	**0.37**	(0.01)
	ℓ_2	**0.21**	(0.01)	0.23	(0.01)	**0.21**	(0.01)	0.23	(0.01)	0.27	(0.01)	**0.22**	(0.01)
	ME	**0.03**	(0.00)	0.04	(0.00)	0.04	(0.00)	0.04	(0.00)	0.05	(0.00)	**0.03**	(0.00)
	FP	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)
	FN	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)
AR (0.5)	ℓ_1	**0.35**	(0.01)	0.45	(0.01)	**0.35**	(0.01)	0.39	(0.01)	0.62	(0.02)	**0.37**	(0.01)
	ℓ_2	**0.20**	(0.01)	0.23	(0.01)	0.21	(0.01)	0.23	(0.01)	0.34	(0.01)	**0.22**	(0.01)
	ME	**0.03**	(0.00)	0.05	(0.00)	**0.03**	(0.00)	**0.04**	(0.00)	0.09	(0.00)	**0.04**	(0.00)
	FP	**0**	(0)	1	(0)	**0**	(0)	**0**	(0)	0	(0)	**0**	(0)
	FN	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)

the reduced set from the fused Kolmogorov filter has 5,000 probes.

Based on 200 random partitions, we report the prediction error and run time in Table 3.4. We observe CRR-SCAD has the lowest prediction error whereas LS-SCAD has the highest error. When p grows from 300 to 5000, both RR-SCAD and CRR-SCAD become more accurate; this may be because some important variables are discarded in the screening step. In terms of speed, we see the smoothed rank loss offers some obvious benefits in the computational efficiency: CRR-SCAD is as fast as LS-SCAD and it is about two orders of magnitude faster than RR-SCAD. LS-SCAD is implemented in a standard way by using the LLA algorithm with the `glmnet` package.

Table 3.2. *Comparison of least-square regression with SCAD (LS-SCAD), rank regression with SCAD (RR-SCAD) and convoluted rank regression with SCAD (CRR-SCAD). The comparison criteria are ℓ_1 error, ℓ_2 error, model error (ME), number of false positive variables (FP) and number of false negative variables (FN). In each example, the best method evaluated based on each criterion is in boldface. All the quantities are averaged over 200 independent runs and standard errors are given in parentheses. In all the examples shown in this table, the error term in the data generating model follows a mixture normal distribution:* $\epsilon \sim 0.95\mathrm{N}(0,1) + 0.05\mathrm{N}(0,100)$.

		$p = 400$						$p = 3000$					
Σ	criterion	LS-SCAD		RR-SCAD		CRR-SCAD		LS-SCAD		RR-SCAD		CRR-SCAD	
CS (0.2)	ℓ_1	1.51	(0.08)	**0.18**	(0.01)	0.22	(0.01)	3.18	(0.14)	**0.19**	(0.01)	0.21	(0.01)
	ℓ_2	0.79	(0.04)	**0.16**	(0.01)	0.17	(0.01)	1.68	(0.07)	**0.16**	(0.01)	**0.16**	(0.01)
	ME	0.67	(0.06)	**0.03**	(0.00)	**0.03**	(0.00)	5.18	(0.33)	**0.03**	(0.00)	**0.03**	(0.00)
	FP	1	(0)	**0**	(0)	**0**	(0)	1	(0)	**0**	(0)	**0**	(0)
	FN	**0**	(0)	**0**	(0)	**0**	(0)	1	(0)	**0**	(0)	**0**	(0)
CS (0.5)	ℓ_1	1.86	(0.11)	**0.21**	(0.01)	0.24	(0.01)	3.72	(0.15)	0.25	(0.01)	**0.21**	(0.01)
	ℓ_2	0.90	(0.05)	**0.16**	(0.01)	0.18	(0.01)	1.84	(0.08)	0.18	(0.01)	**0.17**	(0.01)
	ME	0.47	(0.04)	**0.03**	(0.00)	**0.03**	(0.00)	7.82	(0.51)	**0.03**	(0.00)	**0.03**	(0.00)
	FP	2	(0)	**0**	(0)	**0**	(0)	2	(0)	**0**	(0)	**0**	(0)
	FN	**0**	(0)	**0**	(0)	**0**	(0)	1	(0)	**0**	(0)	**0**	(0)
AR (0.5)	ℓ_1	1.22	(0.05)	**0.19**	(0.01)	0.26	(0.01)	1.72	(0.07)	**0.20**	(0.01)	0.22	(0.01)
	ℓ_2	0.73	(0.03)	**0.16**	(0.01)	0.18	(0.01)	1.03	(0.04)	**0.16**	(0.01)	**0.16**	(0.01)
	ME	0.50	(0.04)	**0.03**	(0.00)	**0.03**	(0.00)	1.44	(0.10)	**0.03**	(0.00)	**0.03**	(0.00)
	FP	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)
	FN	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)

When we implemented CRR-SCAD, we made some efforts to integrate the GCD and LLA algorithms by avoiding some repeated computation, thus our CRR-SCAD is even faster than LS-SCAD when $p = 5000$. Without such implementation efforts, our CRR-SCAD would be slower than LS-SCAD.

Table 3.3. *Comparison of least-square regression with SCAD (LS-SCAD), rank regression with SCAD (RR-SCAD), and convoluted rank regression with SCAD (CRR-SCAD). The comparison criteria are ℓ_1 error, ℓ_2 error, model error (ME), number of false positive variables (FP) and number of false negative variables (FN). In each example, the best method evaluated based on each criterion is in boldface. All the quantities are averaged over 200 independent runs and standard errors are given in parentheses. In all the examples shown in this table, the error term in the data generating model $\epsilon \sim \sqrt{2}t(4)$.*

		$p = 400$						$p = 3000$					
Σ	criterion	LS-SCAD		RR-SCAD		CRR-SCAD		LS-SCAD		RR-SCAD		CRR-SCAD	
CS (0.2)	ℓ_1	1.13	(0.03)	0.79	(0.02)	**0.58**	(0.02)	3.33	(0.10)	1.69	(0.06)	**0.63**	(0.02)
	ℓ_2	0.63	(0.02)	0.43	(0.01)	**0.34**	(0.01)	1.74	(0.05)	0.82	(0.02)	**0.37**	(0.01)
	ME	0.42	(0.02)	0.19	(0.01)	**0.12**	(0.01)	4.70	(0.26)	0.64	(0.03)	**0.14**	(0.01)
	FP	1	(0)	0	(0)	**0**	(0)	2	(0)	5	(0)	**0**	(0)
	FN	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)
CS (0.5)	ℓ_1	1.33	(0.05)	0.72	(0.02)	**0.70**	(0.02)	4.01	(0.12)	1.10	(0.03)	**0.72**	(0.02)
	ℓ_2	0.69	(0.02)	0.41	(0.01)	**0.40**	(0.01)	1.95	(0.06)	0.63	(0.02)	**0.41**	(0.01)
	ME	0.34	(0.02)	0.14	(0.01)	**0.13**	(0.01)	7.53	(0.47)	0.26	(0.01)	**0.14**	(0.01)
	FP	2	(0)	**0**	(0)	**0**	(0)	4	(0)	**0**	(0)	**0**	(0)
	FN	**0**	(0)	**0**	(0)	**0**	(0)	1	(0)	**0**	(0)	**0**	(0)
AR (0.5)	ℓ_1	1.12	(0.03)	0.89	(0.03)	**0.62**	(0.02)	1.56	(0.04)	1.50	(0.04)	**0.71**	(0.02)
	ℓ_2	0.66	(0.02)	0.46	(0.01)	**0.37**	(0.01)	0.93	(0.02)	0.86	(0.03)	**0.41**	(0.01)
	ME	0.37	(0.02)	0.18	(0.01)	**0.12**	(0.01)	1.00	(0.05)	0.64	(0.03)	**0.15**	(0.01)
	FP	**0**	(0)	1	(0)	**0**	(0)	**0**	(0)	1	(0)	**0**	(0)
	FN	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)	**0**	(0)

Table 3.4. *Real data analysis. Comparison of prediction error and run time using least-square regression with SCAD (LS-SCAD), rank regression with SCAD (RR-SCAD), and convoluted rank regression with SCAD (CRR-SCAD). The data is split into a training and a test set in the ratio of 1:1 and the fused Kolmogorov filter is applied to reduced the dimension to 300 and 5000. All the quantities are averaged over 200 random partitions. The lowest prediction errors are in boldface, and standard errors are given in parentheses.*

	$p = 300$			$p = 5000$		
method	prediction error		time (sec)	prediction error		time (sec)
LS-SCAD	1.027	(0.018)	2.52	1.061	(0.017)	8.76
RR-SCAD	0.942	(0.015)	20.86	0.865	(0.012)	487.91
CRR-SCAD	**0.898**	(0.010)	1.86	**0.825**	(0.009)	7.81

Chapter 4

Density-Convoluted Support Vector Machines for High-Dimensional Classification

The support vector machine (SVM) is a popular classification method which enjoys good performance in many real applications. The SVM can be viewed as a penalized minimization problem in which the objective function is the expectation of hinge loss function with respect to the standard non-smooth empirical measure corresponding to the true underlying measure. We further extend this viewpoint and propose a smoothed SVM by substituting a kernel density estimator for the measure in the expectation calculation. The resulting method is called density convoluted support vector machine (DCSVM). We argue that the DCSVM is particularly more interesting than the standard SVM in the context of high-dimensional classification. We systematically study the rate of convergence of the elastic-net penalized DCSVM and prove it has order $O_p(\sqrt{\frac{s \log p}{n}})$ under general random design setting. We further develop novel efficient algorithm for computing elastic-net penalized DCSVM. Simulation studies and 8 benchmark datasets are used to demonstrate the superior classification performance of elastic-net DCSVM over other competitors, and it is demonstrated in these numerical studies that the computation of DCSVM can be more than 100 times faster than that of the SVM.

4.1 Introduction

Due to the advanced technology for data collection over the past decades, there has been a surge of data complexity in many research fields such as genomics, genetics and finance, among others. Consequently, it is very common for the number of predictors in the dataset to be far larger than the number of observations (Donoho et al., 2000). For example, in genomics it is crucial to build a classifier for the purpose of disease diagnosis, with thousands of candidate genes at hand but only tens of instances available for study. Such high dimensionality in data makes traditional classification methods infeasible and poses new challenges from both theoretical and computational perspectives.

One method for performing high dimensional classification is the penalized large margin classifier. The standard support vector machine (SVM), initially proposed and investigated in Boser et al. (1992) and Vapnik (1995), has an objective equal to hinge loss plus an ℓ_2 penalty. It is also referred to as ℓ_2-norm SVM. When the dimension greatly exceeds the sample size and there are many noisy features in the predictor set, it has been shown that it is more beneficial to use a sparse penalty such as the ℓ_1 norm penalty (a.k.a. the lasso) to replace the ℓ_2 norm penalty in order to perform classification and variable selection simultaneously in high dimensional setting (Zhu et al., 2003; Wang et al., 2006). Consider the ℓ_1 norm SVM for example. It can be written as

$$\min_{\beta_0,\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} L\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\beta_0)\big) + \lambda\|\boldsymbol{\beta}\|_1, \tag{4.1.1}$$

where $L(u) = (1-u)_+$ is the hinge loss. Just like in lasso regression, the ℓ_1 penalty induces sparsity in the solution and is thus capable of removing irrelevant features. More recently, Peng et al. (2016) investigated the rate of convergence of the ℓ_1-norm

SVM and an error bound of $O(\sqrt{\frac{s \log p}{n}})$ was established in their paper.

The sparse penalized SVM can be computationally intensive especially when the number of predictors is huge in the dataset, owing to the non-differentiable loss function part. It is known that penalized problem in high dimensions with a smooth loss function can be efficiently computed by cyclical coordinate descent algorithm (Friedman, Hastie and Tibshirani, 2010). Nevertheless, the SVM is based on the non-differentiable hinge loss, which means that there is no convergence guarantee if one uses cyclical coordinate descent to solve the SVM. In principle, coordinate descent may not give the right solution due to the non-differentiability of the objective function (Luo and Tseng, 1992; Tseng, 2001). A similar problem under regression context is the quantile regression, in which the check loss is not differentiable (Fan et al., 2020). The typical method of solving quantile regression is the interior point algorithm. Since ℓ_1-norm SVM can be transformed into linear programming, one may also consider interior point algorithm for solving it. However, interior point algorithm may not scale well with high dimensional input and thus is not suitable for solving SVM in high dimensions.

Recently, Fernandes et al. (2021) studied an interesting smoothing technique for solving quantile regression with statistical guarantees. Later, Tan et al. (2021) further studied the smoothing quantile regression under high dimensional settings and showed that the statistical property of quantile regression is maintained after smoothing. Motivated by their work, we develop a smooth version of SVM from statistical perspective, as opposed to trying to solve it exactly. Consider the first term in ((4.1.1))

$$\frac{1}{n}\sum_{i=1}^{n} L\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\beta_0)\big), \tag{4.1.2}$$

which is non-smooth. If we could replace it by some smooth loss such that the resulting estimator has nice theoretical properties, then we should focus on solving the

smooth problem instead of the original problem. In fact, one may view ((4.1.2)) as the expectation of the hinge loss function with respect to the empirical measure assigning $\frac{1}{n}$ probability mass to each $y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\beta_0), i = 1, \ldots, n$. The empirical measure is viewed as an estimator for the true distribution of the random variable $y(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}+\beta_0)$. Clearly, if we estimate the true distribution by using a smoothed kernel density estimator, then we can take the expectation of the hinge loss function with respect to the distribution determined by the smoothed kernel density estimator. This leads us to a new objective function

$$\frac{1}{n}\sum_{i=1}^{n} L_h\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\beta_0)\big), \tag{4.1.3}$$

which we use to replace ((4.1.2)). Here h is the bandwidth of kernel density estimator and is used to index the new classifier. The resulting estimator is named as *density convoluted support vector machine (DCSVM)*, since the kernel density estimator has a convolution interpretation. We study the following general form of penalized DCSVM in high dimensions,

$$\frac{1}{n}\sum_{i=1}^{n} L_h\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\beta_0)\big) + \lambda_0\|\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

The resulting estimator is called elastic-net DCSVM, which involves both ℓ_1-DCSVM and ℓ_2-DCSVM as special cases. By its convexity and smoothness, elastic-net DCSVM can be efficiently solved by using the generalized coordinate descent algorithm (Yang and Zou, 2013).

In this paper, we first study the theoretical properties of the elastic-net DCSVM. We show that the convergence rate of the elastic-net DCSVM is $O_p(\sqrt{\frac{s\log p}{n}})$ under the general random design setting. Furthermore, we develop novel efficient algorithm for computing elastic-net DCSVM. We use simulation studies and 8 benchmark datasets

to demonstrate that elastic-net DCSVM delivers superior classification performance over its competitors, and the computational speed of DCSVM can be two orders of magnitude faster than that of SVM.

4.2 Density-Convoluted SVM

4.2.1 Notation and definitions

We first introduce some notation that is used throughout the paper. For an arbitrary index set $\mathbf{A} \subset \{1, \ldots, p\}$, any vector $\mathbf{c} = (c_1, \ldots, c_p)$ and any $n \times p$ matrix $\mathbf{U}$, let $\mathbf{c}_\mathbf{A} = (c_i, i \in \mathbf{A})$, and let $\mathbf{U}_\mathbf{A}$ be the submatrix with columns of $\mathbf{U}$ whose indices are in $\mathbf{A}$. The complement of an index set $\mathbf{A}$ is denoted as $\mathbf{A}^{\mathrm{c}} = \{1, \ldots, p\} \setminus \mathbf{A}$. For any finite set $\mathbf{B}$, let $|\mathbf{B}|$ be the number of elements in $\mathbf{B}$. For a vector $\mathbf{c} \in \mathbb{R}^p$ and $q \in [1, \infty)$, let $\|\mathbf{c}\|_q = (\sum_{j=1}^p |c_j|^q)^{\frac{1}{q}}$ be its ℓ_q norm, let $\|\mathbf{c}\|_\infty$ (or $\|\mathbf{c}\|_{\max}$) $= \max_j |c_j|$ be its ℓ_∞ norm, and let $\|\mathbf{c}\|_{\min} = \min_j |c_j|$ be its minimum absolute value. For a matrix $\mathbf{M}$, let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ be its eigenvalue with smallest absolute value and largest absolute value, respectively. This is the common notation for eigenvalues of a matrix, and $\lambda_{\min}$, $\lambda_{\max}$ should not be confused with the penalization parameter used in a penalty function. For any matrix $\mathbf{G}$, let $\|\mathbf{G}\| = \sqrt{\lambda_{\max}(\mathbf{G}^{\mathrm{T}}\mathbf{G})}$ be its spectral norm. In particular, for a vector $\mathbf{c}$, $\|\mathbf{c}\| = \|\mathbf{c}\|_2$. For $a, b \in \mathbb{R}$, let $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For a sequence $\{a_n\}$ and another nonnegative sequence $\{b_n\}$, we write $a_n = O(b_n)$ if there exists a constant $c > 0$ such that $|a_n| \leq cb_n$ for all $n \geq 1$. Also, we use $a_n = o(b_n)$, or $a_n \ll b_n$, to represent $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$. We write $b_n \gg a_n$ if $a_n \ll b_n$. Let $(\Omega, \mathcal{G}, \mathrm{P})$ be a probability space on which all the random variables that appear in this paper are defined. Let $\mathbb{E}[\cdot]$ be the expectation corresponding to the probability measure P. Let $\psi : [0, \infty) \to [0, \infty]$ be a nondecreasing, convex function with $\psi(0) = 0$, then we denote $\|Z\|_\psi = \inf\{t > 0 : \mathbb{E}[\psi(\frac{|Z|}{t})] \leq 1\}$ as

th ψ-Orlicz norm for any random variable Z. In particular, if $p \geq 1$, let $\psi_p(x) \coloneqq \mathrm{e}^{x^p} - 1$ which is a nondecreasing convex function with $\psi_p(0) = 0$, then we denote its corresponding Orlicz norm as $\|Z\|_{\psi_p} = \inf\{t > 0 : \mathbb{E}[\mathrm{e}^{\frac{|Z|^p}{t^p}}] \leq 2\}$ where Z is any random variable. For a sequence of random variables $\{Z_n\}_{n\geq 1}$, we write $Z_n = O_p(1)$ if $\lim_{M\to\infty} \limsup_{n\to\infty} \mathrm{P}(|Z_n| > M) = 0$, and we write $Z_n = o_p(1)$ if $\lim_{n\to\infty} \mathrm{P}(|Z_n| > \epsilon) = 0, \forall \epsilon > 0$. For two sequences of random variables Z_n and Z'_n, we write $Z_n = O_p(Z'_n)$ if $\frac{Z_n}{Z'_n} = O_p(1)$, and we write $Z_n = o_p(Z'_n)$ if $\frac{Z_n}{Z'_n} = o_p(1)$.

4.2.2 Density-Convoluted SVM

Suppose the training data consists of n observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, where $y_i \in \{-1, 1\}$ is the class label and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ are predictors for the ith subject. We use $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ to denote the design matrix, where $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ contains observations for the jth variable, and use $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ to represent the response vector. We focus on the general case where the observed data $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ are i.i.d. samples from the distribution of a random vector $(y, \mathbf{x})$. Let the jth component of the random vector $\mathbf{x}$ be denoted as x_j. Meanwhile, let $\tilde{\mathbf{x}} = (1, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}}$ and $\tilde{\mathbf{x}}_i = (1, \mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}$, $i = 1, \ldots, n$. To perform the classification task, the support vector machine (SVM, Vapnik, 1995) seeks a separating hyperplane $\{\mathbf{x} : \beta_0 + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = 0\}$ where

$$\begin{aligned} \min_{\beta_0, \boldsymbol{\beta}, \xi_i} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 \\ \text{subject to} \quad & y_i\left(\beta_0 + \mathbf{x}_i^{\top}\boldsymbol{\beta}\right) \geq 1 - \xi_i, \xi_i \geq 0, \sum_{i=1}^n \xi_i \leq c. \end{aligned} \tag{4.2.1}$$

It is well known that the above problem can be equivalently formulated as a penalized empirical risk minimization problem:

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^n L\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \beta_0)\big) + \lambda_0\|\boldsymbol{\beta}\|_2^2, \tag{4.2.2}$$

where $L(u) = (1-u)_+ = \max\{1-u, 0\}$ is known as the SVM hinge loss and $\lambda_0 > 0$ is a tuning parameter that is one-to-one correspondent to the constant c in problem ((4.2.1)).

Let us consider the population version of risk appearing in ((4.2.2)), $\mathbb{E}[L\big(y(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}+\beta_0)\big)]$. If we define new random variable $U = y(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}+\beta_0)$ and let $F(u;\boldsymbol{\beta},\beta_0)$ be its cumulative distribution function (cdf), then the population risk is written as $\int_{-\infty}^{\infty} L(t)\mathrm{d}F(t;\boldsymbol{\beta},\beta_0)$. The unpenalized objective function in ((4.2.2)) can be further viewed as $\int_{-\infty}^{\infty} L(t)\mathrm{d}\hat{F}(t;\boldsymbol{\beta},\beta_0)$, where $\hat{F}(t;\boldsymbol{\beta},\beta_0) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\beta_0)\le t\}}$ is the empirical cdf based on i.i.d. realizations of U. The usage of the discontinuous empirical cdf here makes the objective in ((4.2.2)) to have the same degree of smoothness as the hinge loss $L(\cdot)$, i,e. continuous but nondifferentiable. This has motivated us to consider an alternative estimator for the cdf. If we use an estimator $\tilde{F}(\cdot\,;\boldsymbol{\beta},\beta_0)$ that is smooth enough, the $\int_{-\infty}^{\infty} L(t)\mathrm{d}\tilde{F}(t;\boldsymbol{\beta},\beta_0)$ shall lead us towards a new objective which is differentiable to certain degrees.

In particular, we consider the cdf from the kernel density estimator

$$\tilde{F}(t;\boldsymbol{\beta},\beta_0) = \int_{-\infty}^{t} \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{u - y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\beta_0)}{h}\Big)\mathrm{d}u,$$

where $K : \mathbb{R} \to [0,\infty)$ is a smooth kernel function satisfying $K(-u) = K(u), \forall u \in \mathbb{R}$, $\int_{-\infty}^{\infty} K(t)\mathrm{d}t = 1$ and $\int_{-\infty}^{\infty} |t|K(t)\mathrm{d}t < \infty$, and $h > 0$ is the bandwidth parameter to be tuned. Replacing $\hat{F}$ by $\tilde{F}$ gives the new objective function,

$$\begin{aligned}&\int_{-\infty}^{\infty} L(t)\mathrm{d}\tilde{F}(t;\boldsymbol{\beta},\beta_0)\\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} L(t)\frac{1}{h}K\Big(\frac{t - y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\beta_0)}{h}\Big)\mathrm{d}t \triangleq \frac{1}{n}\sum_{i=1}^{n} L_h\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\beta_0)\big)\end{aligned}$$

where $L_h(t) = \int_{-\infty}^{\infty}(1-u)_+\frac{1}{h}K\left(\frac{u-t}{h}\right)du$. Note that $L_h(\cdot)$ is a convex function that is at least second order differentiable. Also, it satisfies the relation $L_h = L * K_h$ where

$K_h(u) = \frac{1}{h}K(\frac{u}{h})$ and "$*$" stands for convolution.

As such, with the penalty term $\lambda_0\|\boldsymbol{\beta}\|_2^2$, we obtain

$$\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} L_h(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \beta_0)) + \lambda_0\|\boldsymbol{\beta}\|_2^2. \tag{4.2.3}$$

We treat the classifier arisen from the above problem as a new classifier and coin it the density-convoluted SVM (DCSVM).

As discussed above, DCSVM originates from a statistical view of the SVM, while it shows merit from the computational perspective as it overcomes the non-differentiability of the original SVM problem. Smoothing a non-differentiable problem through convolution can be traced back to the idea of *mollification* (Friedrichs, 1944) and has also been studied in the optimization community, for example, Bertsekas (1973) and Rubinstein (1983). The method was recently adopted to smooth the quantile regression by He et al. (2021), Fernandes et al. (2021) and Tan et al. (2021).

In this work, we focus on two most popular kernel functions, Gaussian kernel and Epanechnikov kernel in DCSVM, and we denote the corresponding convoluted loss function by $L_h^G(v)$ and $L_h^E(v)$, respectively.

For the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}}\exp\{-u^2/2\}$, one can show that

$$L_h^G(v) = (1-v)\Phi\left(\frac{1-v}{h}\right) + \frac{h}{\sqrt{2\pi}}\exp\left\{-\frac{(1-v)^2}{2h^2}\right\},$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

For the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)I(-1 \le u \le 1)$, where $I(\cdot)$ is the

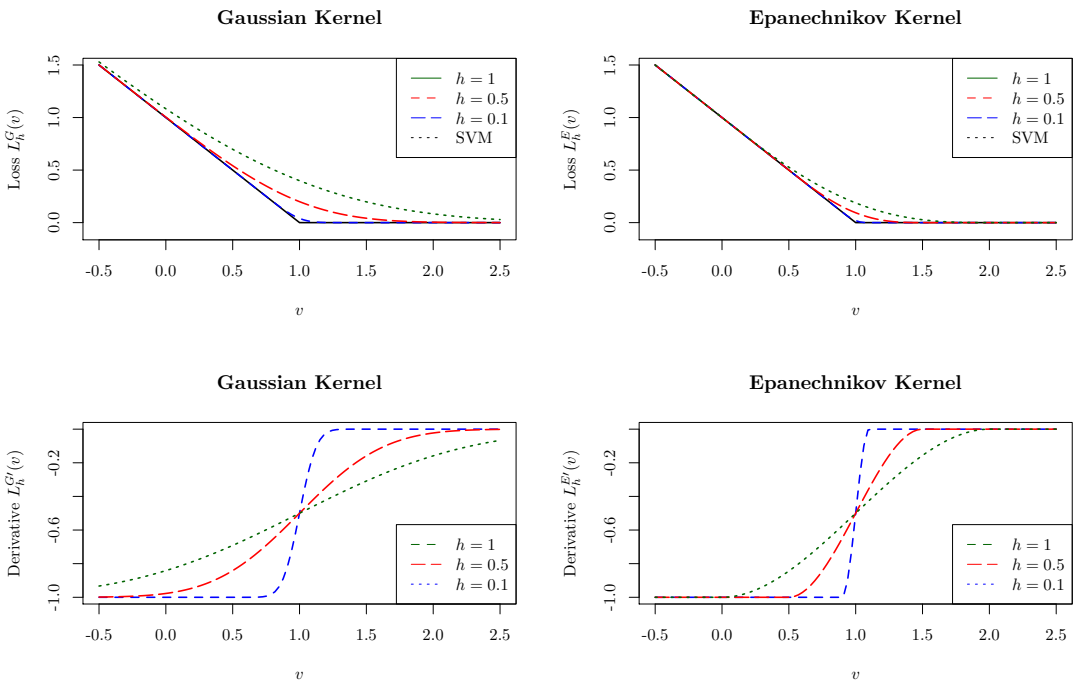

Figure 4.1. *Top row: plots of $L_h^G(v)$ and $L_h^E(v)$, the density-convoluted SVM loss functions with Gaussian kernel (left) and Epanechnikov kernels (right). Bottom row: plots of the first-order derivatives, $L_h^{G\prime}(v)$ and $L_h^{E\prime}(v)$.*

indicator function,

$$L_h^E(v) = \begin{cases} 1 - v, & v \le 1 - h, \\ \dfrac{(1 - v + h)^3(3h - (1 - v))}{16h^3}, & 1 - h < v \le 1 + h, \\ 0, & v \ge 1 + h. \end{cases}$$

The top row of Figure 4.1 depicts the DCSVM losses with Gaussian kernel and Epanechnikov kernel.

It can be shown that when $h \to 0$, $L_h(\cdot)$ converges pointwise to $L(\cdot)$ and the objective function of DCSVM reduces to that of the ordinary SVM. Thus the nonsmooth

SVM can be viewed as a marginal case of a broad family of smooth classifiers being indexed by h. In practice, the best h can be determined in a data-driven approach such as cross-validation.

4.2.3 Sparse density-convoluted SVM

Define $(\beta_0^*, \boldsymbol{\beta}^*) = \arg\min_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}\times\mathbb{R}^p} \mathbb{E}\big[L_h\big(y(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \beta_0)\big)\big]$. In high dimensions, we consider designing the estimator under a sparsity assumption that $\boldsymbol{\beta}^*$ has many zero components. Let $\mathbb{A} = \{j : \beta_j^* \neq 0, 1 \leq j \leq p\}$ be the support set of $\boldsymbol{\beta}^*$, i.e., the set of indices of the important covariates. Let $s = |\mathbb{A}|$. Throughout this paper, we allow $p = p_n$ and $s = s_n$ to diverge with n, and we assume $s_n \geq 1$ and p_n goes to infinity as n goes to infinity. For convenience, we still use p and s to represent these quantities since no confusion is caused. In ultra-high dimensions, the dimension p is allowed to increase exponentially with the sample size n. We also assume that s is relatively of smaller order compared to n, which is necessary for the existence of a consistent estimator.

To perform the classification for high-dimensional data, we present sparse DCSVM with an additional ℓ_1-penalty term

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) \coloneqq \underset{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} L_h(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \beta_0)) + \lambda_0\|\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{4.2.4}$$

The ℓ_1-penalty is used to induce sparsity in the estimator. We also consider the following version of sparse DCSVM with only an ℓ_1-penalty term:

$$(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}) \coloneqq \underset{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} L_h(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \beta_0)) + \lambda\|\boldsymbol{\beta}\|_1. \tag{4.2.5}$$

Borrowing the commonly used terminologies for different penalties in high dimensional literature, we refer to the estimator in ((4.2.4)) as elastic-net DCSVM, and refer to

the estimator in ((4.2.5)) as lasso DCSVM. Note that the lasso DCSVM is a special case of elastic-net DCSVM with $\lambda_0 = 0$.

4.3 Theoretical Studies

We now state the assumptions needed to establish our theoretical results. We first impose the following conditions on the random design.

Assumption 5 $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, $(y, \mathbf{x})$ are independent and identically distributed on $\mathbb{R} \times \mathbb{R}^p$. $\mathbf{x}$ is a zero-mean sub-exponential random vector, i.e. $\mathbb{E}[\mathbf{x}] = \mathbf{0}$, and there exists a constant $m_0 > 0$ such that

$$\sup_{\mathbf{a} \in \mathbb{R}^p : \|\mathbf{a}\|_2 \leq 1} \|\mathbf{a}^{\mathrm{T}} \mathbf{x}\|_{\psi_1} \leq m_0.$$

By definition of Orlicz norm and Markov's inequality, this further implies

$$\sup_{\mathbf{a} \in \mathbb{R}^p : \|\mathbf{a}\|_2 \leq 1} \mathrm{P}(|\mathbf{a}^{\mathrm{T}} \mathbf{x}| > t) \leq 2\mathrm{e}^{-\frac{t}{m_0}}, \forall t \geq 0. \qquad \square$$

For any index set $\mathbf{A} \subset \{1, \ldots, p\}$, consider the cone $\mathcal{S}_{\mathbf{A}} := \{(\delta, \mathbf{u}) \in \mathbb{R} \times \mathbb{R}^p : \|\mathbf{u}_{\mathbf{A}^{\mathrm{c}}}\|_1 \leq 3\|\mathbf{u}_{\mathbf{A}}\|_1 + |\delta|\}$. Such type of cone has been widely considered in literature on high dimensional statistics. Meanwhile, let $I(\beta_0, \boldsymbol{\beta}) := \mathbb{E}[L_h''\big(y(\beta_0 + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\big)\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathrm{T}}]$ be Hessian matrix of the population loss, or information matrix. We impose the following condition on the information.

Assumption 6 There exists a constant $\rho > 0$ such that

$$\min_{(\delta, \mathbf{u}) \in \mathcal{S}_{\mathbb{A}} : \delta^2 + \|\mathbf{u}\|_2^2 = O(\frac{s \log p}{n})} \lambda_{\min}\big(I(\beta_0^* + \delta, \boldsymbol{\beta}^* + \mathbf{u})\big) \geq \rho$$

for large enough n. □

Assumption 5 is a general setting in the random design, which relaxes the classical condition that the components of $\mathbf{x}$ are bounded random variables (Peng et al., 2016). Assumption 6, which is a restricted eigenvalues type of condition, is needed to establish ℓ_2-type error bound for ℓ_1-penalized type of estimator. Similar conditions have been widely adopted in the literature (Bühlmann and Van De Geer, 2011; Fan et al., 2020).

Theorem 1

Assume assumptions 5-6 hold, and $s \log p = o(n)$. Choose the tuning parameters such that $8\lambda_0 \|\beta^*\|_{\max} < \lambda$. Then there exists a large enough constant $c_0 > 0$ such that with the choice $\lambda = c_0\sqrt{\frac{\log p}{n}}$, the elastic-net DCSVM estimator $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ satisfies

$$|\hat{\beta}_0 - \beta_0^*|^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 = O_p\Big(\frac{s \log p}{n}\Big).$$

Theorem 1 shows that the sparse density convoluted SVM estimator shares the same optimal rate of convergence as the ℓ_1-SVM (Peng et al., 2016). It is worth noting that we establish Theorem 1 under sub-exponential random design, which is more general than the bounded design used in Peng et al. (2016). Meanwhile, the sparse DCSVM has better computational efficiency than penalized SVM due to the smoothness of its loss function. □

4.4 Computation

In this section, we develop an efficient algorithm for computing the solution path of DCSVM.

At the outset, we present the first-order derivative of the density-convoluted SVM

loss and show they are Lipschitz continuous in Lemma 1:

$$L_h^{G\prime}(v) = -\Phi\left(\frac{1-v}{h}\right),$$

$$L_h^{E\prime}(v) = \begin{cases} -1, & v \le 1-h, \\ -\dfrac{(1-v+h)^2(2h-(1-v))}{4h^3}, & 1-h < v \le 1+h, \\ 0, & v \ge 1+h. \end{cases}$$

Lemma 1

Let $L_h^G(v)$ and $L_h^E(v)$ be the DCSVM loss using Gaussian kernel and Epanechnikov kernel, respectively. For $v_1 < v_2$,

$$|L_h^{G\prime}(v_1) - L_h^{G\prime}(v_2)| < c_h^G |v_1 - v_2|, \tag{4.4.1}$$

$$|L_h^{E\prime}(v_1) - L_h^{E\prime}(v_2)| < c_h^E |v_1 - v_2|, \tag{4.4.2}$$

□

where the Lipschitz constants are given as $c_h^G = \frac{1}{\sqrt{2\pi}h}$ and $c_h^E = \frac{3}{4h}$.

The bottom row of Figure 4.1 depicts $L_h^{G\prime}(v)$ and $L_h^{E\prime}(v)$.

Lemma 1 gives rise to the following quadratic majorization condition for the DCSVM:

$$L_h(v_1) \le L_h(v_2) + L_h'(v_2)(v_1 - v_2) + \frac{c_h}{2}(v_1 - v_2)^2, \tag{4.4.3}$$

where L_h is exemplified by L_h^G and L_h^E and c_h is the corresponding Lipschitz constant.

Based on the Lipschitz condition, we develop a generalized coordinate descent (GCD) algorithm (Yang and Zou, 2013) to solve those sparse penalized DCSVMs. We first consider the adaptive lasso penalty. The algorithm can be easily adjusted to handle lasso and elastic net.

Without loss of generality, we assume each $\mathbf{X}_j$ has zero mean and unit length. In a coordinate-wise manner, suppose the coordinate $\beta_1, \beta_2, \ldots, \beta_{j-1}$ have been updated and we now update β_j. Denote by $\tilde{\beta}_0$ and $\tilde{\boldsymbol{\beta}}$ by the current solution and let $v_i = y_i(\tilde{\beta}_0 + \mathbf{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})$. To update β_j, instead of solving the coordinate-wise update function,

$$F(\beta_j) = \frac{1}{n}\sum_{i=1}^{n} L_h\left(v_i + y_i x_{ij}\left(\beta_j - \tilde{\beta}_j\right)\right) + \lambda w_j|\beta_j|,$$

we solve its majorization function

$$Q\left(\beta_j\right) = \frac{1}{n}\sum_{i=1}^{n} L_h\left(v_i\right) + \frac{1}{n}\sum_{i=1}^{n} L_h'\left(v_i\right) y_i x_{ij}\left(\beta_j - \tilde{\beta}_j\right) + \frac{c_h}{2}\left(\beta_j - \tilde{\beta}_j\right)^2 + \lambda w_j|\beta_j|,$$

that is obtained through the quadratic majorization condition. The minimizer of $Q\left(\beta_j\right)$ is

$$\left(\tilde{\beta}_j - \frac{1}{c_h n}\sum_{i=1}^{n} L_h'(v_i) y_i x_{ij}\right)\left(1 - \frac{\lambda w_j}{\left|c_h\tilde{\beta}_j - \frac{1}{n}\sum_{i=1}^{n} L_h'(v_i) y_i x_{ij}\right|}\right)_+.$$

Likewise, β_0 is updated to be $\tilde{\beta}_0 - \frac{1}{c_h n}\sum_{i=1}^{n} L_h'(v_i) y_i$.

In our implementation, we further apply the strong rule (Tibshirani et al., 2010), warm start, and active set strategy (Friedman, Hastie and Tibshirani, 2010) to further accelerate the algorithm.

4.5 Numerical Studies

4.5.1 Simulation

In this section, we use several simulation examples to demonstrate the performance of DCSVM.

The response variables of all the simulated data are binary and the two classes are balanced, i.e., $P(Y = 1) = P(Y = -1) = 0.5$. In each example, define the p-dimensional mean vectors $\boldsymbol{\mu}_+ = (0.7, 0.7, 0.7, 0.7, 0.7, 0, 0, \ldots, 0)$ and $\boldsymbol{\mu}_- = -\boldsymbol{\mu}_+$, where $p = 500$ or 5000 in our experiments. Each observation from the positive class is drawn from $\mathrm{N}(\boldsymbol{\mu}_+, \boldsymbol{\Sigma})$ and each observation from the negative class is drawn from $\mathrm{N}(\boldsymbol{\mu}_-, \boldsymbol{\Sigma})$. We consider three different choices of $\boldsymbol{\Sigma}$. In example 1, $\boldsymbol{\Sigma} = \mathbf{I}_{p\times p}$ so the variables are independent. In both examples 2 and 3,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{\star}_{5\times 5} & \mathbf{0}_{5\times(p-5)} \\ \mathbf{0}_{(p-5)\times 5} & \mathbf{I}_{(p-5)\times(p-5)} \end{pmatrix}$$

where $\boldsymbol{\Sigma}^{\star}_{5\times 5}$ have all diagonal elements of 1 and off-diagonal elements of ρ in example 2, and $(\boldsymbol{\Sigma}^{\star}_{5\times 5})_{i,j} = \rho^{|i-j|}$ in example 3. We use $\rho = 0.2, 0.7$, and 0.9.

We first compared elastic-net DCSVM with Gaussian kernel and Epanechnikov kernel with elastic-net SVM (Wang et al., 2006) and elastic-net logistic regression that is fitted using the R package `gcdnet` (Yang and Zou, 2013). For each example, the training size is 200 and we use five-fold cross-validation to select the best tuple of (h, λ_0, λ) where h is chosen from $0.1, 0.25, 0.5$, and 1, λ_0 is selected from $0.5 * (10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 5)$, and λ is searched along the solution path; for the SVM and logistic regression, we select λ_0 and λ in the same manner.

We record the prediction error and run time in Table 4.1. The run time include all the time spent on tuning and training the models. We observe the DCSVM with Epanechnikov kernel has slightly better performance than DCSVM with Gaussian kernel, and both of them have better prediction accuracy than the other two methods. DCSVM with Epanechnikov kernel is the fastest while the elastic-net SVM is the slowest.

All the methods exhibited in Table 4.1 use elastic-net penalty. We now study the performance when using other sparse penalities. Due to the overall best performance,

Table 4.1. *Comparison of prediction error (in percentage) and run time (in second) of elastic-net density-convoluted SVM with Gaussian and Epanechnikov kernels, elastic-net SVM, and elastic-net logistc regression. Under each simulation setting, the method with the lowest prediction error is marked by a black box. All the quantities are averaged over 50 independent runs and the standard errors of the prediction error are given in parentheses.*

		DCSVM-Gaussian			DCSVM-Epanechnikov			SVM			logistic		
p	ρ	err (%)		time	err (%)		time	err (%)		time	err (%)		time
Example 1													
500		6.83	(0.14)	267.89	**6.75**	(0.14)	29.67	9.76	(1.51)	1362.44	6.98	(0.15)	49.78
5000		**7.11**	(0.13)	771.87	7.29	(0.16)	139.07	7.90	(0.87)	28323.47	7.33	(0.17)	417.54
Example 2													
500	0.2	13.52	(0.19)	305.95	**13.48**	(0.17)	33.42	16.02	(1.26)	1687.62	13.88	(0.22)	52.44
	0.7	22.65	(0.25)	385.08	**22.50**	(0.27)	41.39	25.75	(1.21)	1585.23	22.88	(0.28)	59.99
	0.9	24.76	(0.24)	467.40	**24.57**	(0.24)	48.78	27.42	(1.16)	1510.98	24.82	(0.31)	69.52
5000	0.2	13.78	(0.18)	806.36	**13.72**	(0.21)	142.09	16.32	(1.25)	30170.44	14.12	(0.26)	420.09
	0.7	**22.66**	(0.21)	890.84	23.00	(0.24)	150.44	24.15	(0.79)	31865.01	23.03	(0.23)	435.63
	0.9	**24.70**	(0.25)	975.34	24.76	(0.24)	154.73	26.88	(1.00)	32132.55	25.03	(0.24)	450.30
Example 3													
500	0.2	10.30	(0.15)	290.41	**10.13**	(0.16)	31.53	12.04	(1.14)	1476.20	10.69	(0.24)	51.16
	0.7	19.48	(0.18)	368.74	**19.40**	(0.18)	39.71	22.90	(1.34)	1726.07	19.80	(0.25)	60.53
	0.9	**23.50**	(0.22)	435.55	23.54	(0.22)	44.92	26.55	(1.19)	1625.15	23.93	(0.28)	66.23
5000	0.2	10.51	(0.20)	793.67	**10.46**	(0.18)	141.23	13.02	(1.35)	34555.70	10.74	(0.21)	418.58
	0.7	**19.70**	(0.21)	877.54	19.89	(0.22)	146.99	22.54	(1.18)	34574.72	20.09	(0.25)	433.84
	0.9	23.85	(0.23)	944.63	**23.81**	(0.24)	152.78	26.55	(1.11)	36732.99	23.90	(0.24)	445.60

we stay with DCSVM with Epanechnikov kernel and we compare the prediction accuracy and variable selection when using lasso and elastic-net penalties. We present the results in Table 4.2. In general, we find the elastic-net has the best performance in both prediction error and variable selection.

4.5.2 Benchmark data applications

In this section, we demonstrate the performance of DCSVM using several benchmark data, which are available from UCI machine learning repository. We randomly split each data set into a training set and a test set with a 1:1 ratio. On the training set, we fit elastic-net DCSVM, elastic-net logistic regression, and elastic-net SVM, and tune each method using five-fold cross-validation. The prediction accuracy is computed

Table 4.2. *Comparison of prediction error (in percentage) and variable selection of density-convoluted SVM with Epanechnikov kernels using lasso and elastic-net (enet) penalties. Denote by C and IC the number of correctly and incorrectly selected variables, respectively. Under each simulation setting, the method with the lowest prediction error is marked by a black box. All the quantities are averaged over 50 independent runs and the standard errors of the prediction error are given in parentheses.*

		lasso-DCSVM				enet-DCSVM			
p	ρ	err (%)		C	IC	err (%)		C	IC
Example 1									
500		6.88	(0.14)	5	0	**6.77**	(0.14)	5	0
5000		7.31	(0.19)	5	0	**7.29**	(0.16)	5	0
Example 2									
500	0.2	13.89	(0.23)	5	0	**13.47**	(0.17)	5	0
	0.7	22.86	(0.20)	3	0	**22.51**	(0.27)	5	0
	0.9	24.53	(0.19)	2	0	**24.51**	(0.23)	4	0
5000	0.2	14.55	(0.25)	5	0	**13.72**	(0.21)	5	0
	0.7	23.41	(0.23)	3	0	**23.05**	(0.25)	4	0
	0.9	25.36	(0.35)	2	0	**24.76**	(0.26)	3	0
Example 3									
500	0.2	10.47	(0.22)	5	0	**10.09**	(0.15)	5	0
	0.7	19.90	(0.22)	3	0	**19.44**	(0.19)	4	0
	0.9	23.74	(0.20)	3	0	**23.49**	(0.22)	4	0
5000	0.2	10.78	(0.23)	5	0	**10.48**	(0.18)	5	0
	0.7	20.12	(0.22)	3	0	**19.89**	(0.22)	4	0
	0.9	24.34	(0.31)	2	0	**23.81**	(0.24)	3	0

based on the test set.

We present the result in Table 4.3. We observe the elastic-net DCSVM has the best performance in general.

Table 4.3. *Comparison of prediction error (in percentage) and run time (in second) of elastic-net density-convoluted SVM with Epanechnikov kernel, elastic-net SVM, and elastic-net logistc regression. For each benchmark data, the method with the lowest prediction error is marked by a black box. All the quantities are averaged over 50 independent runs and the standard errors of the prediction error are given in parentheses.*

			enet-DCSVM			enet-SVM			enet-logistic		
data	n	p	err (%)		time	err (%)		time	err (%)		time
arcene	100	9920	**32.24**	(1.46)	53.26	37.09	(1.59)	8912.87	35.82	(1.65)	219.30
breast	42	22283	**25.90**	(1.64)	51.33	30.38	(2.05)	1946.98	30.76	(2.14)	227.88
colon	62	2000	**18.13**	(1.03)	10.22	18.90	(1.55)	722.48	23.87	(1.51)	27.33
leuk	72	7128	**3.50**	(0.47)	22.98	3.89	(0.51)	1863.23	4.33	(0.61)	115.00
LSVT	126	309	16.01	(0.73)	6.25	16.20	(0.68)	74.20	**15.87**	(0.68)	9.05
malaria	71	22283	**5.37**	(0.68)	85.52	7.60	(1.21)	12046.09	6.80	(0.98)	483.20
ovarian	253	15154	**0.63**	(0.12)	189.22	4.87	(1.23)	14442.87	0.87	(0.14)	964.16
prostate	102	6033	9.25	(0.67)	29.34	**8.98**	(0.50)	2421.20	10.24	(0.61)	116.50

References

Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98**, 791–806.

Bertsekas, D. P. (1973). Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications* **12**, 218–231.

Bertsekas, D. P. (1999). *Nonlinear programming.* Athena Scientific.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37**, 1705–1732.

Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory.*

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* **35**, 2313–2351.

Carroll, R. J. and Ruppert, D. (1988). Transformation and weighting in regression. *(Vol. 30). CRC Press.*

Chen, Z., Fan, J. and Li, R. (2018). Error variance estimation in ultrahigh-dimensional additive models. *Journal of the American Statistical Association* **113**, 315–327.

Chiang, A. P., Beck, J. S. ,Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R., Nishimura, D., Braun, T. A., Kim, K.-Y., Huang, J.,

ELBEDOUR, K., CARMI, R., SLUSARSKI, D. C., CASAVANT, T. L., STONE, E. M., and SHEFFIELD, V. C. (2006). Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11). *Proceedings of the National Academy of Sciences* **103**, 6287–6292.

CLEVELAND, W. S. (1993). Visualizing data. *Hobart Press*.

COOK, R. D. and WEISBERG, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* **70**, 1–10.

COX, D. R. and SNELL, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)* **30**, 248–265.

DAYE, Z.J., CHEN, J. and LI, H. (2012). High-dimensional heteroscedastic regression with an application to eQTL data analysis. *Biometrics* **68**, 316–326.

DONOHO, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture* **1**, 32.

EFRON, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica* **1**, 93–125.

EFRON, B., HASTIE, T., JOHNSTONE, I., and TIBSHIRANI, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–499.

FAN, J., GUO, S. and HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society. Series B* **74**, 37–65.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

FAN, Y. and LI, R. (2012). Variable selection in linear mixed effects models. *Annals of statistics* **40**, 2043.

FAN, J., LI, R., ZHANG, C.-H., and ZOU, H. (2020). *Statistical foundations of data science*. Chapman & Hall, CRC.

FAN, Y. and TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society. Series B* **75**, 531–552.

FAN, J., XUE, L. and ZOU, H. (2014b). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics* **42**, 819–849.

FEIGL, P. and ZELEN, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics* 826–838.

FERNANDES, M., GUERRE, E. and HORTA, E. Smoothing Quantile Regressions. *Journal of Business and Economic Statistics* **39**, 338–357.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1.

FRIEDRICHS, K. O. (1944). The identity of weak and strong extensions of differential operators. *Transactions of the American Mathematical Society* **55**, 132–151.

GÖTZE, F., SAMBALE, H. and SINULIS, A. (2021). Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability* **26**, 1–22.

GU, Y., FAN, J., KONG, L., MA, S. and ZOU, H. (2018). ADMM for high-dimensional sparse penalized quantile regression. *Technometrics* **60**, 319–331.

GU, Y. and ZOU, H. (2016). High-dimensional generalizations of asymmetric least squares regression and their applications. *Annals of Statistics* **44**, 2661–2694.

HASTIE, T., TIBSHIRANI, R., and TIBSHIRANI, R.J. (2017). *Extended comparisons of best subset selection, forward stepwise selection, and the Lasso*. arXiv: https://arxiv.org/abs/1707.08692.

HASTIE, T., TIBSHIRANI, R., and WAINWRIGHT, M. (2015). *Statistical learning with sparsity: the LASSO and generalizations*. Boca Raton, FL: CRC Press.

HE, X., PAN, X., TAN, K. M. and ZHOU, W.-X. (2021). Smoothed quantile regression with large-scale inference. *Journal of Econometrics* .

HETTMANSPERGER, T. P. and MCKEAN, J. W. (2010). *Robust Nonparametric Statistical Methods.* CRC Press.

HOEFFDING, W. (1961). The strong law of large numbers for U-statistics. North Carolina State University. Department of Statistics.

HONG, M., WANG, X., RAZAVIYAYN, M. and LUO, Z.-Q. (2013). Iteration complexity analysis of block coordinate descent methods. *arXiv preprint arXiv:1310.6957* .

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

JAECKEL, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *The Annals of Mathematical Statistics* **58**, 1449–1458.

KADKHODAIE, M., SANJABI, M. and LUO, Z.-Q. (2014). On the linear convergence of the approximate proximal splitting method for non-smooth convex optimization. *Journal of the Operations Research Society of China* **2**, 123–141.

KOENKER, R. and BASSETT, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society* **46**, 33–50.

KOENKER, R. and BASSETT, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society* **50**, 43–61.

KOENKER, R. and ZHAO, Q. (1994). L-estimation for linear heteroscedastic models. *Journal of Nonparametric Statistics* **3**, 223–235.

LEDOUX, M. and TALAGRAND, M. (2013). Probability in Banach Spaces: isoperimetry and processes. Springer Science & Business Media.

LUO, Z.-Q. and TSENG, P. (1992). On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization* **30**, 408–425.

LUO, Z.-Q. and TSENG, P. (1993). Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research* **46**, 157–178.

MAI, Q. and ZOU, H. (2015). The fused Kolmogorov filter: A nonparametric model-free screening method. *Annals of Statistics* **43**, 1471–1497.

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Annals of Statistics* **37**, 3779–3821.

MEINSHAUSEN, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis* **52**, 374–393.

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J., and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27**, 538–557.

NEWEY, W. K. and POWELL, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society* **55**, 819–847.

PARIKH, N. and BOYD, S. (2013). Proximal algorithms. *Foundations and Trends in optimization* **1**, 123–231.

PENG, B., WANG, L. and WU, Y. (2016). An error bound for l1-norm support vector machine coefficients in ultra-high dimension. *The Journal of Machine Learning Research* **17**, 8279–8304.

RIGOLLET, P. and HÜTTER, J. C. (2015). High dimensional statistics. *Lecture notes for course 18S997.*

RUBINSTEIN, R. Y. (1983). Smoothed functionals in stochastic optimization. *Mathematics of Operations Research* **8**, 26–33.

SCHEETZ, T.E., KIM, K.-Y. A. ,SWIDERSKI, R.E., PHILP, A.R., BRAUN, T.A., KNUDTSON, K.L., DORRANCE, A.M., DIBONA, G.F., HUANG, J., CASAVANT, T.L., SHEFFIELD, V.C., and STONE, E.M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences* **103**, 14429–14434.

TAN, K. M., WANG, L. and ZHOU, W.-X. (2021). High-dimensional quantile regression: convolution smoothing and concave regularization. *arXiv preprint arXiv:2109.05640* .

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* **58**, 267–288.

TIBSHIRANI, R., BIEN, J., FRIEDMAN, J., HASTIE, T., SIMON, N., TAYLOR, J. and TIBSHIRANI, R. J. (2010). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society, Series B* **74**, 245–266.

TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications* **109**, 475–494.

VAN DER VAART, A. W. and WELLNER, J. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

VAPNIK, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* (Y. Eldar and G. Kutyniok, eds.) 210–268. Cambridge Univ. Press, Cambridge.

VERSHYNIN, R. (2018). *High-dimensional probability: An introduction with applications in data science*, vol. 47. Cambridge university press.

WANG, L., KIM, Y. and LI, R. (2013). Calibrating non-convex penalized regression in ultra-high dimension. *Annals of Statistics* **41**, 2505–2536.

WANG, L. and LI, R. (2009). Weighted Wilcoxon-type smoothly clipped absolute deviation method. *Biometrics* **65**, 564–571.

WANG, L., PENG, B., BRADIC, J., LI, R. and WU, Y. (2020). A tuning-free robust and efficient approach to high-dimensional regression (with discussion). *Journal of the American Statistical Association* **115**, 1700–1714.

WANG, L., WU, Y. and LI, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association* **107**, 214–222.

WANG, L., ZHU, J. and ZOU, H. (2006). The doubly regularized support vector machine. *Statistica Sinica* **16**, 589–616.

YANG, Y. and ZOU, H. (2013). An efficient algorithm for computing the HHSVM and its generalizations. *Journal of Computational and Graphical Statistics* **22**, 396–415.

YE, F. and ZHANG, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q Loss in ℓ_r Balls. *Journal of Machine Learning Research* **11**, 3519–3540.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38**, 894–942.

ZHANG, C.H. and HUANG, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics* **36**, 1567–1594.

ZHANG, H., JIANG, J. and LUO, Z.-Q. (2013). On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. *Journal of the Operations Research Society of China* **1**, 163–186.

ZHU, J., ROSSET, S., TIBSHIRANI, R. and HASTIE, T. J. (2003). 1-norm support vector machines. In *Advances in neural information processing systems*.

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36**, 1509–1533.

Appendix A

Proof of Chapter 2

A.1 Proofs for the main results

In Appendix A.1.1, we first present some general technical lemmas and propositions that are frequently used in the proof for our main results. We then present the proof for Theorem 1 in Appendix A.1.2 and sketch the proof of Theorem 2 in Appendix A.1.3. All the other proofs are placed in Appendix Appendix A.2–Appendix A.8.

A.1.1 General technical lemmas and propositions

Proposition 2 Let $\epsilon_i, i = 1, \ldots, n$ be i.i.d. sub-Gaussian(σ) random variables. Then, for any real numbers $a_1, \ldots, a_n$, any $t > 0$,

$$\mathrm{P}(|\sum_{i=1}^{n} a_i \epsilon_i| > t) \leq 2 \exp\Big(- \frac{t^2}{2\sigma^2 \sum_{i=1}^{n} a_i^2} \Big). \qquad \square$$

Proposition 3 Let $\epsilon_i, i = 1, \ldots, n$ be independent sub-exponential(λ) random variables, i.e. $\mathbb{E}[X] = 0$ and $\mathbb{E}[\mathrm{e}^{tX}] \leq \mathrm{e}^{\frac{t^2\lambda^2}{2}}$, $\forall |t| \leq \frac{1}{\lambda}$, for some $\lambda > 0$. Then, for any real numbers $a_1, \ldots, a_n$, any $t > 0$,

$$\mathrm{P}(|\sum_{i=1}^{n} a_i \epsilon_i| > t) \leq 2 \exp\Big[- \Big(\frac{t^2}{2\lambda^2 \sum_{i=1}^{n} a_i^2} \wedge \frac{t}{2\lambda \max_{1 \leq i \leq n} |a_i|} \Big) \Big]. \qquad \square$$

Lemma 2 Let ϵ be sub-Gaussian(σ) random variable that satisfies assumption ($\mathbf{A}_0$). Let μ be some real number satisfying $|\mu| \leq c$ for some $c > 0$. Then, $\log|\epsilon+\mu| - \mathbb{E}[\log|\epsilon+\mu|]$ is sub-exponential(η) random variables for any $\eta \geq 2\sqrt{(4\sigma^2 \mathrm{e}^{(4L+2)c}) \vee (2c\mathrm{e}^{(2L+1)c}) \vee (2 + 2C_0\mathrm{e}^{(2L+1)c})}$, with $C_0 = \sup_{x\in\mathbb{R}} f(x)$ being the maximum density of ϵ and L being the Lipschitz constant for f. □

Lemma 3 Let ϵ be a random variable that has a density f on $\mathbb{R}$ with respect to lebesgue measure. And let f satisfies $|f(x) - f(y)| \leq L|x-y|$ for some constant $L > 0$. For positive integer k define $h_k(\mu) \coloneqq \big|\mathbb{E}[\log^k|\epsilon+\mu|] - \mathbb{E}[\log^k|\epsilon|]\big|$. Then for any $\mu \in \mathbb{R}$, we have

(i) $h_1(\mu) \leq (2L+1)|\mu|$,

(ii) $h_2(\mu) \leq \mu^2 + (4L + 2\mathbb{E}\big[\big|\log|\epsilon|\big|\big])|\mu|$,

(iii) $h_4(\mu) \leq \mu^4 + 4\mathbb{E}\big[\big|\log|\epsilon|\big|\big]|\mu|^3 + 6\mathbb{E}[\log^2|\epsilon|]\mu^2 + (48L + 4\mathbb{E}\big[\big|\log^3|\epsilon|\big|\big])|\mu|$. □

Proposition 4 (Upper bounds for a generic ℓ_1 problem) A generic ℓ_1 penalized estimator is defined as

$$\tilde{\beta}^{\ell_1} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\, \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_{\text{lasso}}\|\beta\|, \tag{A.1.1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is some vector, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is some matrix and $\lambda_{\text{lasso}} > 0$. Let $\beta^* \in \mathbb{R}^p$ be some sparse vector, i.e. $|\mathbb{A}| = s < p$ where $\mathbb{A} = \{i : \beta_i^* \neq 0, 1 \leq i \leq p\}$. Assume that the matrix $\mathbf{X}$ satisfies one of the following:

RE: $\min_{\mathbf{u}\neq 0:\|\mathbf{u}_{\mathbb{A}^c}\|_{\ell_1} \leq 3\|\mathbf{u}_{\mathbb{A}}\|_{\ell_1}} \frac{\|\mathbf{X}\mathbf{u}\|_{\ell_2}^2}{n\|\mathbf{u}\|_{\ell_2}^2} \geq \alpha \in (0,\infty)$, or

GIF: $\min_{\mathbf{u}\neq 0:\|\mathbf{u}_{\mathbb{A}^c}\|_{\ell_1} \leq 3\|\mathbf{u}_{\mathbb{A}}\|_{\ell_1}} \frac{\|\mathbf{X}\mathbf{u}\|_{\ell_2}^2}{n\|\mathbf{u}_{\mathbb{A}}\|_{\ell_1}\|\mathbf{u}\|_\infty} \geq \tau \in (0,\infty)$.

Then, given $\lambda_{\text{lasso}} > \frac{2}{n}\|\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\beta^*)\|_\infty$, $\tilde{\beta}^{\ell_1}$ from ((A.1.1)) satisfies

$$\|\tilde{\beta}^{\ell_1} - \beta^*\|_2 \leq \frac{3}{\alpha}\sqrt{s}\lambda_{\text{lasso}}, \text{ if } \mathbf{RE} \text{ holds;}$$
$$\|\tilde{\beta}^{\ell_1} - \beta^*\|_\infty \leq \frac{3}{\tau}\lambda_{\text{lasso}}, \text{ if } \mathbf{GIF} \text{ holds.}$$ □

Remark 12 The proofs of the above results can be found in Appendix Appendix A.2 of the supplementary material. Proposition 2-3 give tail bounds for sub-Gaussian and sub-exponential random variables. Lemma 2 shows the the sub-exponential tail for $\log|\epsilon+\mu|$ with ϵ being sub-Gaussian, and Lemma 3 provides deviation bounds for the moments of such type of random variable. These two lemmas are newly developed and they are important to our proofs. Proposition 4 is from Hastie et al. (2015). Notice that the notation is generic, and the statement is non-stochastic. □

A.1.2 Proofs for Theorem 1

First, we define some necessary notation and quantities for the proof, as well as a useful proposition. Recall that $\ell_n(\beta) = \frac{1}{n}\sum_{i=1}^{\frac{n}{2}}(y_i - \mathbf{x}_i^{\mathrm{T}}\beta)^2$ and define $\hat{\beta}^{\mathrm{ora1}} := \arg\min_{\beta:\beta_{\mathbb{A}_1^{\mathrm{C}}}=\mathbf{0}} \ell_n(\beta)$. Similarly, we can define $\hat{\beta}^{\mathrm{ora2}} := \arg\min_{\beta:\beta_{\mathbb{A}_1^{\mathrm{C}}}=\mathbf{0}} \tilde{\ell}_n(\beta)$, where $\tilde{\ell}_n(\beta) := \frac{1}{n}\sum_{i=\frac{n}{2}+1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\beta)^2$.

Proposition 5 Let assumptions $(\mathbf{A}_0)$, $(\mathbf{C}_1)$, $(\mathbf{C}_2)$ or $(\mathbf{C}_2')$ and $(\mathbf{C}_3)$ hold. Assume that $a_0\kappa \geq 3s_1^{\frac{1}{2}}$ under $(\mathbf{C}_2)$ or $a_0\rho \geq 3$ under $(\mathbf{C}_2')$. Choose the tuning parameters so that $\|\beta^*_{\mathbb{A}_1}\|_{\min} > (a+1)(\lambda \vee \tilde{\lambda})$. Then we have

(i) $\hat{\beta}(Z^{(1)}) = \hat{\beta}^{\mathrm{ora1}}$ holds true with probability at least $1 - 2p\exp\left(-\frac{n\lambda^2}{16M\sigma^2\Omega^2}\right)$ $- 2(p-s_1)\exp\left(-\frac{a_1^2 n\lambda^2}{4\sigma^2\Omega^2 M}\right) - 2s_1\exp\left(-\frac{n\varphi(\|\beta^*_{\mathbb{A}_1}\|_{\min}-a\lambda)^2}{4\sigma^2\Omega^2}\right)$.

(ii) $\hat{\beta}(Z^{(2)}) = \hat{\beta}^{\mathrm{ora2}}$ holds true with probability at least $1 - 2p\exp\left(-\frac{n\tilde{\lambda}^2}{16M\sigma^2\Omega^2}\right)$ $- 2(p-s_1)\exp\left(-\frac{a_1^2 n\tilde{\lambda}^2}{4\sigma^2\Omega^2 M}\right) - 2s_1\exp\left(-\frac{n\varphi(\|\beta^*_{\mathbb{A}_1}\|_{\min}-a\tilde{\lambda})^2}{4\sigma^2\Omega^2}\right)$.

(iii) If we suitably choose $\lambda \asymp \tilde{\lambda} \asymp \sqrt{\frac{\log p}{n}}$, then we have $\|\hat{\beta}(Z^{(i)}) - \beta^*\|_{\ell_2} = O_p(\sqrt{\frac{s_1}{n}}), i = 1, 2$. □

The proof of Proposition 5 is relegated to Appendix Appendix A.3 of the supplementary material. Proposition 5 is used in the proof of Theorem 1.

Recall that $\ell_n^1(\gamma) = \frac{1}{n}\sum_{i=\frac{n}{2}+1}^{n}\left(\log|y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}(Z^{(1)})| - \mathbf{x}_i^{\mathrm{T}}\gamma\right)^2$ in $\mathcal{R}_2$. Similarly, we denote $\tilde{\ell}_n^1(\gamma) = \frac{1}{n}\sum_{i=1}^{\frac{n}{2}}\left(\log|y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}(Z^{(2)})| - \mathbf{x}_i^{\mathrm{T}}\gamma\right)^2$ in $\mathcal{R}_1$. We also define

$$\ell_{n\bullet}^1(\gamma) := \frac{1}{n}\sum_{i=\frac{n}{2}+1}^{n}\left(\log|y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}^{\mathrm{ora1}}| - \mathbf{x}_i^{\mathrm{T}}\gamma\right)^2,$$

$$\tilde{\ell}^1_{n\bullet}(\gamma) := \frac{1}{n}\sum_{i=1}^{\frac{n}{2}} \big(\log|y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}^{\mathrm{ora2}}| - \mathbf{x}_i^{\mathrm{T}}\gamma\big)^2.$$

Here we use a "$\bullet$" sign to indicate the corresponding quantity is hypothetical. We use this kind of notation throughout the proof part of this paper. Then, we define another two hypothetical estimators as follows:

$$\hat{\gamma}^{\mathrm{ora}}_{\bullet} := \underset{\gamma\in\mathbb{R}^p:\gamma_{\mathbb{A}_2^c}=\mathbf{0}}{\arg\min}\ \ell^1_{n\bullet}(\gamma), \quad \tilde{\gamma}^{\mathrm{ora}}_{\bullet} := \underset{\gamma\in\mathbb{R}^p:\gamma_{\mathbb{A}_2^c}=\mathbf{0}}{\arg\min}\ \tilde{\ell}^1_{n\bullet}(\gamma).$$

For (i) and (ii) of Theorem 1, due to symmetry, it suffices to prove the results for $\hat{\gamma}(Z^{(2)} \to Z^{(1)})$. We consider a general ℓ_1 penalized estimator $\tilde{\gamma}^{\mathrm{lasso}} := \arg\min_{\gamma\in\mathbb{R}^p} \tilde{\ell}^1_n(\gamma) + \tilde{\lambda}'_{\mathrm{lasso}}\sum_{j=1}^p |\gamma_j|$, where $\tilde{\lambda}'_{\mathrm{lasso}} > 0$ is some tuning parameter. Recall that with $\mathbf{0}$ as initial value, the first iteration of the LLA algorithm in $\mathcal{R}_1$ gives the above ℓ_1 penalized estimator with $p'_{\tilde{\lambda}_1}(0)$ as tuning parameter. For SCAD and MCP, $p'_{\tilde{\lambda}_1}(0) = \tilde{\lambda}_1$. For the LLA with initial value $\tilde{\gamma}^{\mathrm{lasso}}$, we have the following result.

(ii') Choose the tuning parameters such that $\|\beta^*_{\mathbb{A}_1}\|_{\min} > (a+1)(\lambda\vee\tilde{\lambda})$ and $\|\gamma^*_{\mathbb{A}_2}\|_{\min} > (a+1)(\lambda_1\vee\tilde{\lambda}_1)$. If $a_0\kappa \geq 3s_1^{\frac{1}{2}}$ and we pick $\tilde{\lambda}_1 \geq \frac{3s_2^{\frac{1}{2}}\tilde{\lambda}'_{\mathrm{lasso}}}{a_0\kappa'}$ under $(\mathbf{C}_2)$, or $a_0\rho \geq 3$ and $\tilde{\lambda}_1 \geq \frac{3\tilde{\lambda}'_{\mathrm{lasso}}}{a_0\rho'}$ under $(\mathbf{C}'_2)$, then, $\hat{\mathbb{A}}^{(2)} = \mathbb{A}_2$ holds true with probability at least

$$\begin{aligned}
&1 - 2p\exp\big(-\frac{n\tilde{\lambda}^2}{16M\sigma^2\Omega^2}\big) - 2(p-s_1)\exp\big(-\frac{a_1^2 n\tilde{\lambda}^2}{4\sigma^2\Omega^2 M}\big)\\
&- 2s_1\exp\big(-\frac{n\varphi(\|\beta^*_{\mathbb{A}_1}\|_{\min} - a\tilde{\lambda})^2}{4\sigma^2\Omega^2}\big)\\
&- 2p\exp\big(-\delta_1 n\big) - 2(p-s_2)\exp\big(-\delta'_1 n\big) - 2s_2\exp\big(-\delta''_1 n\big)\\
&- n\exp\big(-\frac{(K\wedge\frac{C_1}{(4L+2)G_1}\wedge\frac{C_2}{(4L+2)G_2}\wedge\frac{C_3}{(4L+2)G_3})^2\Psi^2\varphi}{4\sigma^2\Omega^2 s_1 M}n\big),
\end{aligned}$$

where $\delta_1 = \frac{\tilde{\lambda}'^2_{\mathrm{lasso}}}{64\eta_0^2 M}\wedge\frac{\tilde{\lambda}'_{\mathrm{lasso}}}{16\eta_0\sqrt{M}}$, $\delta'_1 = \frac{a_1^2\tilde{\lambda}_1^2}{16\eta_0^2 G_2^2}\wedge\frac{a_1\tilde{\lambda}_1}{8\eta_0 G_2}$, $\delta''_1 = \frac{\varphi'^2(\|\gamma^*_{\mathbb{A}_2}\|_{\min}-a\tilde{\lambda}_1)^2}{16\eta_0^2 s_2 M}\wedge\frac{\varphi'(\|\gamma^*_{\mathbb{A}_2}\|_{\min}-a\tilde{\lambda}_1)}{8\eta_0\sqrt{s_2 M}}$, $C_1 = \frac{\tilde{\lambda}'_{\mathrm{lasso}}}{2}$, $C_2 = a_1\tilde{\lambda}_1$, $C_3 = \|\gamma^*_{\mathbb{A}_2}\|_{\min} - a\tilde{\lambda}_1$, and G_1, G_2 and G_3 are the same as in Theorem 1 (i).

With (ii') in hand, (ii) of Theorem 1 follows by taking $\tilde{\lambda}'_{\mathrm{lasso}} = \tilde{\lambda}_1$.

The proof of (ii') makes use of Lemma 4 and Proposition 6–7 below. Lemma 4

is an newly developed concentration inequality that is frequently used to bound the relevant probabilities in the proofs of Theorem 1 and Theorem 2. Its proof is placed in Appendix Appendix A.4 of the supplementary material. The proof of Proposition 6 and the proof of Proposition 7 are placed in Appendix Appendix A.5 and Appendix Appendix A.6 of the supplementary material, respectively.

Lemma 4 Assume assumptions $(\mathbf{A}_0)$, $(\mathbf{C}_1)$, $(\mathbf{C}_3)$ hold. Let $\mathbf{a} = (a_1, a_2, \ldots, a_{\frac{n}{2}})^{\mathrm{T}}$ and satisfies $|a_i| \leq G, \forall i = 1, \ldots, \frac{n}{2}$ for some constant $G > 0$. Let $v_i := \mathrm{e}^{\mathbf{x}_i^{\mathrm{T}}\gamma^*}, i = 1, \ldots, n$. Let $\zeta_i = \frac{1}{v_i}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}(\hat{\beta}_{\mathbb{A}_1}^{\mathrm{ora2}} - \beta_{\mathbb{A}_1}^*), i = 1, \ldots, \frac{n}{2}$. For $t > 0$, $\mathcal{T}_t = \{\max_{1 \leq i \leq \frac{n}{2}} |\zeta_i| \leq t\}$. Then, we have

(i) For any $t > 0$,

$$\mathrm{P}(\mathcal{T}_t^{\mathrm{c}}) \leq n \exp\left(-\frac{t^2\Psi^2\varphi}{4\sigma^2\Omega^2 s_1 M}n\right). \tag{A.1.2}$$

(ii) Let $K > 0$ be any fixed positive constant. Then for any $C > 0$,

$$\begin{aligned}&\mathrm{P}\Big(\frac{2}{n}\Big|\mathbf{a}^{\mathrm{T}}\log|\mathbf{y}^{(1)} - \mathbf{X}_{\mathbb{A}_1}^{(1)}\hat{\beta}_{\mathbb{A}_1}^{\mathrm{ora2}}| - \mathbf{a}^{\mathrm{T}}\mathbf{X}^{(1)}\gamma^*\Big| > C, \mathcal{T}_{K\wedge\frac{C}{(4L+2)G}}\Big)\\ &\leq 2\exp\Big[-\Big(\frac{C^2}{16\eta_0^2G^2}\wedge\frac{C}{8\eta_0 G}\Big)n\Big],\end{aligned}$$

where $\eta_0 = 2\sqrt{(4\sigma^2\mathrm{e}^{(4L+2)K}) \vee (2K\mathrm{e}^{(2L+1)K}) \vee (2 + 2C_0\mathrm{e}^{(2L+1)K})}$ is a fixed positive constant. Here the applications of the functions $|\cdot|$ and $\log(\cdot)$ on any vector are componentwise. Consequently, by union bound,

$$\begin{aligned}&\mathrm{P}\Big(\frac{2}{n}\Big|\mathbf{a}^{\mathrm{T}}\log|\mathbf{y}^{(1)} - \mathbf{X}_{\mathbb{A}_1}^{(1)}\hat{\beta}_{\mathbb{A}_1}^{\mathrm{ora2}}| - \mathbf{a}^{\mathrm{T}}\mathbf{X}^{(1)}\gamma^*\Big| > C\Big)\\ &\leq 2\exp\Big[-\Big(\frac{C^2}{16\eta_0^2G^2}\wedge\frac{C}{8\eta_0 G}\Big)n\Big] + n\exp\Big(-\frac{(K\wedge\frac{C}{(4L+2)G})^2\Psi^2\varphi}{4\sigma^2\Omega^2 s_1 M}n\Big).\end{aligned}$$ □

We define $\tilde{\gamma}_{\bullet}^{\mathrm{lasso}} := \arg\min_{\gamma\in\mathbb{R}^p} \tilde{\ell}_{n\bullet}^1(\gamma) + \tilde{\lambda}'_{\mathrm{lasso}}\sum_{j=1}^p |\gamma_j|$.

Proposition 6 Under assumptions $(\mathbf{A}_0)$, $(\mathbf{C}_1)$, $(\mathbf{C}_2)$ or $(\mathbf{C}_2')$, $(\mathbf{C}_3)$, $\tilde{\gamma}_{\bullet}^{\mathrm{lasso}}$ satisfies

$$\mathrm{P}(\|\tilde{\gamma}_{\bullet}^{\mathrm{lasso}} - \gamma^*\|_{\ell_2} > 3s_2^{\frac{1}{2}}\tilde{\lambda}'_{\mathrm{lasso}}\kappa'^{-1}, \mathcal{T}_{K\wedge\frac{C_1}{(4L+2)G_1}}) \leq 2p\exp(-\delta_1 n), \text{ if } (\mathbf{C}_2) \text{ holds;}$$

$$\mathrm{P}(\|\tilde{\gamma}_{\bullet}^{\text{lasso}} - \gamma^*\|_\infty > 3\tilde{\lambda}'_{\text{lasso}}\rho'^{-1}, \mathcal{T}_{K\wedge\frac{C_1}{(4L+2)G_1}}) \le 2p\exp(-\delta_1 n), \text{ if } (\mathbf{C}'_2) \text{ holds,}$$

where $\delta_1 = \frac{\tilde{\lambda}'^2_{\text{lasso}}}{64\eta_0^2 M} \wedge \frac{\tilde{\lambda}'_{\text{lasso}}}{16\eta_0\sqrt{M}}$, $C_1 = \frac{\tilde{\lambda}'_{\text{lasso}}}{2}$ and $G_1 = \sqrt{M}$. Here $\mathcal{T}_t$ is the event that has been defined in Lemma 4. □

Proposition 7 Suppose the tuning parameters are chosen so that $\|\gamma^*_{\mathbb{A}_2}\|_{\min} > (a+1)(\lambda_1 \vee \tilde{\lambda}_1)$. Then, the LLA algorithm in $\mathcal{R}_1$ initialized by $\tilde{\gamma}^{\text{lasso}}$ converges to $\tilde{\gamma}^{\text{ora}}_{\bullet}$ after two iterations with probability at least $1 - \mathrm{P}(\hat{\beta}(Z^{(2)}) \neq \hat{\beta}^{\text{ora2}}) - p'_1 - p'_2 - p'_3 - p'_4$, where

$$\begin{aligned}
p'_1 &= \mathrm{P}(\|\tilde{\gamma}_{\bullet}^{\text{lasso}} - \gamma^*\|_\infty > a_0\tilde{\lambda}_1, \mathcal{T}_{K\wedge\frac{C_1}{(4L+2)G_1}}),\\
p'_2 &= \mathrm{P}(\|\nabla_{\mathbb{A}_2^{\mathrm{c}}}\tilde{\ell}^1_{n\bullet}(\tilde{\gamma}^{\text{ora}}_{\bullet})\|_\infty \ge a_1\tilde{\lambda}_1, \mathcal{T}_{K\wedge\frac{C_2}{(4L+2)G_2}}),\\
p'_3 &= \mathrm{P}(\min_{j\in\mathbb{A}_2}|\tilde{\gamma}^{\text{ora}}_{\bullet j}| < a\tilde{\lambda}_1, \mathcal{T}_{K\wedge\frac{C_3}{(4L+2)G_3}})\\
p'_4 &= \mathrm{P}\big(\mathcal{T}^{\mathrm{c}}_{K\wedge\frac{C_1}{(4L+2)G_1}\wedge\frac{C_2}{(4L+2)G_2}\wedge\frac{C_3}{(4L+2)G_3}}\big),
\end{aligned}$$

where C_1, G_1 have been defined in Proposition 6, C_2, G_2, C_3, G_3 are any positive constants (which will be determined later in the proof of Theorem 1), $\mathcal{T}_x$ has been defined in Lemma 4. □

Proof A.1 (Proof of Theorem 1)
We slightly abuse the notation. Let $\mathbf{y}$ and $\mathbf{X}$ be the response and design matrix in $Z^{(1)}$, i.e. $\mathbf{y} = (y_1, \ldots, y_{n/2})^{\mathrm{T}}$ and $\mathbf{X} = \mathbf{X}^{(1)}$. And we let $\mathbf{X}_j = (x_{1j}, \ldots, x_{\frac{n}{2}j})^{\mathrm{T}}$ to represent the jth column of $\mathbf{X}^{(1)}$.

Recall Proposition 7 shows that the LLA algorithm in $\mathcal{R}_1$ initialized by $\tilde{\gamma}^{\text{lasso}}$ converges to $\tilde{\gamma}^{\text{ora}}_{\bullet}$ after two iterations with probability at least $1-\mathrm{P}(\hat{\beta}(Z^{(2)}) \neq \hat{\beta}^{\text{ora2}}) - p'_1 - p'_2 - p'_3 - p'_4$. From Proposition 5 we already have

$$\begin{aligned}
&\mathrm{P}(\hat{\beta}(Z^{(2)}) \neq \hat{\beta}^{\text{ora2}}) \le 2p\exp\big(-\frac{n\tilde{\lambda}^2}{16M\sigma^2\Omega^2}\big) + 2(p-s_1)\exp\big(-\frac{a_1^2 n\tilde{\lambda}^2}{4\sigma^2\Omega^2 M}\big)\\
&+ 2s_1\exp\big(-\frac{n\varphi(\|\beta^*_{\mathbb{A}_1}\|_{\min} - a\tilde{\lambda})^2}{4\sigma^2\Omega^2}\big).
\end{aligned}$$

And from Proposition 6 we already have

$$\mathrm{P}(\|\tilde{\gamma}_{\bullet}^{\text{lasso}}-\gamma^*\|_{\ell_2}>3s_2^{\frac{1}{2}}\tilde{\lambda}'_{\text{lasso}}\kappa'^{-1},\mathcal{T}_{K\wedge\frac{C_1}{(4L+2)G_1}})\leq 2p\exp(-\delta_1 n),\ \text{if } (\mathbf{C}_2) \text{ holds};$$
$$\mathrm{P}(\|\tilde{\gamma}_{\bullet}^{\text{lasso}}-\gamma^*\|_{\infty}>3\tilde{\lambda}'_{\text{lasso}}\rho'^{-1},\mathcal{T}_{K\wedge\frac{C_1}{(4L+2)G_1}})\leq 2p\exp(-\delta_1 n),\ \text{if } (\mathbf{C}'_2) \text{ holds}.$$

Under $(\mathbf{C}_2)$, since we pick $\tilde{\lambda}_1$ such that $a_0\tilde{\lambda}_1\geq 3s_2^{\frac{1}{2}}\tilde{\lambda}'_{\text{lasso}}\kappa'^{-1}$, and $\|\tilde{\gamma}_{\bullet}^{\text{lasso}}-\gamma^*\|_{\infty}\leq\|\tilde{\gamma}_{\bullet}^{\text{lasso}}-\gamma^*\|_{\ell_2}$, we have

$$p'_1\leq\mathrm{P}(\|\tilde{\gamma}_{\bullet}^{\text{lasso}}-\gamma^*\|_{\ell_2}>3s_2^{\frac{1}{2}}\tilde{\lambda}'_{\text{lasso}}\kappa'^{-1},\mathcal{T}_{K\wedge\frac{C_1}{(4L+2)G_1}})\leq 2p\exp(-\delta_1 n).$$

Under $(\mathbf{C}'_2)$, since we pick $\tilde{\lambda}_1$ such that $a_0\tilde{\lambda}_1\geq 3\tilde{\lambda}'_{\text{lasso}}\rho'^{-1}$, we have

$$p'_1\leq\mathrm{P}(\|\tilde{\gamma}_{\bullet}^{\text{lasso}}-\gamma^*\|_{\infty}>3\tilde{\lambda}'_{\text{lasso}}\rho'^{-1},\mathcal{T}_{K\wedge\frac{C_1}{(4L+2)G_1}})\leq 2p\exp(-\delta_1 n).$$

Next, we bound p'_2 and p'_3 in Proposition 7.

We first look at $p'_2=\mathrm{P}(\|\nabla_{\mathbb{A}_2^{\mathrm{c}}}\tilde{\ell}^1_{n\bullet}(\tilde{\gamma}_{\bullet}^{\text{ora}})\|_{\infty}\geq a_1\tilde{\lambda}_1,\mathcal{T}_{K\wedge\frac{C_2}{(4L+2)G_2}})$. We denote $z_i=\log|y_i-\mathbf{x}_i^{\mathrm{T}}\hat{\beta}^{\text{ora2}}|$, $\mathbf{z}=(z_1,\ldots,z_{\frac{n}{2}})^{\mathrm{T}}$. By definition of $\tilde{\gamma}_{\bullet}^{\text{ora}}$, we know that $\tilde{\gamma}_{\bullet\mathbb{A}_2}^{\text{ora}}=(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2})^{-1}\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{z}$, and $\tilde{\gamma}_{\bullet\mathbb{A}_2^{\mathrm{c}}}^{\text{ora}}=\mathbf{0}$. So $\tilde{\ell}^1_{n\bullet}(\tilde{\gamma}_{\bullet}^{\text{ora}})=\frac{1}{n}\|\mathbf{z}-\mathbf{X}\tilde{\gamma}_{\bullet}^{\text{ora}}\|_{\ell_2}^2=\frac{1}{n}\|\mathbf{z}-\mathbf{X}_{\mathbb{A}_2}\tilde{\gamma}_{\bullet\mathbb{A}_2}^{\text{ora}}\|_{\ell_2}^2$ and $\nabla_{\mathbb{A}_2^{\mathrm{c}}}\tilde{\ell}^1_{n\bullet}(\tilde{\gamma}_{\bullet}^{\text{ora}})=-\frac{2}{n}\mathbf{X}_{\mathbb{A}_2^{\mathrm{c}}}^{\mathrm{T}}(\mathbf{z}-\mathbf{X}_{\mathbb{A}_2}\tilde{\gamma}_{\bullet\mathbb{A}_2}^{\text{ora}})$. By the expression of $\tilde{\gamma}_{\bullet\mathbb{A}_2}^{\text{ora}}$, we have $\nabla_{\mathbb{A}_2^{\mathrm{c}}}\tilde{\ell}^1_{n\bullet}(\tilde{\gamma}_{\bullet}^{\text{ora}})=-\frac{2}{n}\mathbf{X}_{\mathbb{A}_2^{\mathrm{c}}}^{\mathrm{T}}(\mathbf{z}-\mathbf{X}_{\mathbb{A}_2}(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2})^{-1}\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{z})=-\frac{2}{n}\mathbf{X}_{\mathbb{A}_2^{\mathrm{c}}}^{\mathrm{T}}(\mathbf{l}-\mathbf{H}_{\mathbb{A}_2})\mathbf{z}$, in which $\mathbf{H}_{\mathbb{A}_2}=\mathbf{X}_{\mathbb{A}_2}(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2})^{-1}\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}$. Notice that $(\mathbf{l}-\mathbf{H}_{\mathbb{A}_2})\mathbf{X}\gamma^*=(\mathbf{l}-\mathbf{H}_{\mathbb{A}_2})\mathbf{X}_{\mathbb{A}_2}\gamma^*_{\mathbb{A}_2}=\mathbf{0}$. Therefore,

$$\begin{aligned}
p'_2&=\mathrm{P}(\|-\frac{2}{n}\mathbf{X}_{\mathbb{A}_2^{\mathrm{c}}}^{\mathrm{T}}(\mathbf{l}-\mathbf{H}_{\mathbb{A}_2})(\mathbf{z}-\mathbf{X}\gamma^*)\|_{\max}\geq a_1\tilde{\lambda}_1,\mathcal{T}_{K\wedge\frac{C_2}{(4L+2)G_2}})\\
&\leq\sum_{j\in\mathbb{A}_2^{\mathrm{c}}}\mathrm{P}(\frac{2}{n}|\mathbf{X}_j^{\mathrm{T}}(\mathbf{l}-\mathbf{H}_{\mathbb{A}_2})(\mathbf{z}-\mathbf{X}\gamma^*)|\geq a_1\tilde{\lambda}_1,\mathcal{T}_{K\wedge\frac{C_2}{(4L+2)G_2}})\\
&\leq\sum_{j\in\mathbb{A}_2^{\mathrm{c}}}\mathrm{P}(\frac{2}{n}|\mathbf{w}_j^{\mathrm{T}}\mathbf{z}-\mathbf{w}_j^{\mathrm{T}}\mathbf{X}\gamma^*|\geq a_1\tilde{\lambda}_1,\mathcal{T}_{K\wedge\frac{C_2}{(4L+2)G_2}}),
\end{aligned}\tag{A.1.3}$$

where we denote $\mathbf{w}_j^{\mathrm{T}}=\mathbf{X}_j^{\mathrm{T}}(\mathbf{l}-\mathbf{H}_{\mathbb{A}_2})=(w_{1j},\ldots,w_{\frac{n}{2}j}),\forall j\in\mathbb{A}_2^{\mathrm{c}}$, and C_2, G_2 are to be determined. Let $\mathbf{H}'=\mathbf{H}_{\mathbb{A}_2}=(h'_{ij})_{\frac{n}{2}\times\frac{n}{2}}$. We claim that $|h'_{ij}|\leq\frac{2s_2M}{n\varphi'},\forall 1\leq i,j\leq\frac{n}{2}$. In fact, for any $1\leq i\leq\frac{n}{2}$, $h'_{ii}=\mathbf{e}_i^{\mathrm{T}}\mathbf{H}'\mathbf{e}_i$, where $\mathbf{e}_i$ is the unit vector with ith component being 1 and others being 0. Therefore $0\leq h'_{ii}=(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{e}_i)^{\mathrm{T}}(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2})^{-1}(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{e}_i)\leq$

$\frac{2}{n\varphi'}\|\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{e}_i\|_{\ell_2}^2 \le \frac{2s_2M}{n\varphi'}$. On the other hand, since $\mathbf{H}'$ is semi-positive definite, we have $|h'_{ij}| \le \sqrt{h'_{ii}h'_{jj}} \le \frac{2s_2M}{n\varphi'}$. So the claim is true.

Now, we have

$$|w_{ij}| = |\sum_{k=1}^{n/2} x_{kj}(1\{k=i\} - h'_{ki})| \le \sum_{k\neq i}|x_{kj}||h'_{ki}| + |x_{ij}||1-h'_{ii}|$$
$$\le (\frac{n}{2}-1)\sqrt{M}\frac{2s_2M}{n\varphi'} + \sqrt{M}(1+\frac{s_2M}{\varphi'}) \le \sqrt{M}(\frac{2s_2M}{\varphi'}+1) \triangleq G_2,$$

$\forall j \in \mathbb{A}_2^{\mathrm{c}}, \forall i = 1,\dots,n$. Therefore, we use $C = C_2 \coloneqq a_1\tilde{\lambda}_1$ and $G = G_2$ in Lemma 4, and have

$$\mathrm{P}(\frac{2}{n}|\mathbf{w}_j^{\mathrm{T}}\mathbf{z} - \mathbf{w}_j^{\mathrm{T}}\mathbf{X}\gamma^*| \ge a_1\tilde{\lambda}_1, \mathcal{T}_{K\wedge\frac{C_2}{(4L+2)G_2}}) \le 2\exp(-\delta'_1 n), \quad \forall n,$$

where $\delta'_1 = \frac{a_1^2\tilde{\lambda}_1^2}{16\eta_0^2G_2^2} \wedge \frac{a_1\tilde{\lambda}_1}{8\eta_0 G_2}$. Thus by ((A.1.3)) we have $p'_2 \le 2(p-s_2)\exp(-\delta'_1 n)$, $\forall n$.

Next, we examine $p'_3 = \mathrm{P}(\min_{j\in\mathbb{A}_2}|\tilde{\gamma}^{\mathrm{ora}}_{\bullet j}| < a\tilde{\lambda}_1, \mathcal{T}_{K\wedge\frac{C_3}{(4L+2)G_3}})$. By the choice of tuning parameters, we have $\{\min_{j\in\mathbb{A}_2}|\tilde{\gamma}^{\mathrm{ora}}_{\bullet j}| < a\tilde{\lambda}_1\} \subset \{\|\tilde{\gamma}^{\mathrm{ora}}_{\bullet\mathbb{A}_2} - \gamma^*_{\mathbb{A}_2}\|_{\max} \ge \|\gamma^*_{\mathbb{A}_2}\|_{\min} - a\tilde{\lambda}_1\}$. Let us denote $(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2})^{-1}\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}} = (\mathbf{u}_1,\dots,\mathbf{u}_{s_2})^{\mathrm{T}}$, where $\mathbf{u}_j \in \mathbb{R}^n$. Then we have $\mathbf{u}_j = \mathbf{X}_{\mathbb{A}_2}(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2})^{-1}\mathbf{e}_j$, where $\mathbf{e}_j$ is the unit vector with jth element 1 and other elements 0. Then,

$$\begin{aligned} p'_3 &\le \mathrm{P}(\|\tilde{\gamma}^{\mathrm{ora}}_{\bullet\mathbb{A}_2} - \gamma^*_{\mathbb{A}_2}\|_{\max} \ge \|\gamma^*_{\mathbb{A}_2}\|_{\min} - a\tilde{\lambda}_1, \mathcal{T}_{K\wedge\frac{C_3}{(4L+2)G_3}}) \\ &= \mathrm{P}(\|(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2})^{-1}\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}(\mathbf{z}-\mathbf{X}\gamma^*)\|_{\max} \ge \|\gamma^*_{\mathbb{A}_2}\|_{\min} - a\tilde{\lambda}_1, \mathcal{T}_{K\wedge\frac{C_3}{(4L+2)G_3}}) \\ &\le \sum_{j=1}^{s_2}\mathrm{P}(|\mathbf{u}_j^{\mathrm{T}}(\mathbf{z}-\mathbf{X}\gamma^*)| \ge \|\gamma^*_{\mathbb{A}_2}\|_{\min} - a\tilde{\lambda}_1, \mathcal{T}_{K\wedge\frac{C_3}{(4L+2)G_3}}), \end{aligned} \tag{A.1.4}$$

with C_3 and G_3 to be determined later.

Denote $\mathbf{u}_j = (u_{1j},\dots,u_{\frac{n}{2}j})^{\mathrm{T}}$, we claim that $|u_{ij}| \le \frac{2\sqrt{s_2M}}{n\varphi'}, \forall 1 \le i \le \frac{n}{2}, \forall 1 \le j \le s_2$. In fact,

$$\begin{aligned} |u_{ij}| &= |\mathbf{e}_i^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2}(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2})^{-1}\mathbf{e}_j| \le \sqrt{\mathbf{e}_i^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2}(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2})^{-1}\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{e}_i}\sqrt{\mathbf{e}_j^{\mathrm{T}}(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2})^{-1}\mathbf{e}_j} \\ &\le \sqrt{\frac{2\|\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{e}_i\|_{\ell_2}^2}{n\varphi'}}\sqrt{\frac{2}{n\varphi'}} = \frac{2\sqrt{s_2M}}{n\varphi'}. \end{aligned} \tag{A.1.5}$$

Therefore, we use $C = C_3 \coloneqq \|\gamma^*_{\mathbb{A}_2}\|_{\min} - a\tilde{\lambda}_1$ and $G = G_3 \coloneqq \frac{\sqrt{s_2 M}}{\varphi'}$ in Lemma 4, and we have

$$\mathrm{P}(|\mathbf{u}_j^{\mathrm{T}}\mathbf{z} - \mathbf{u}_j^{\mathrm{T}}(\gamma_0^*\mathbf{1}_n + \mathbf{X}\gamma^*)| \geq \|\gamma^*_{\mathbb{A}_2}\|_{\min} - a\tilde{\lambda}_1, \mathcal{T}_{K\wedge\frac{C_3}{(4L+2)G_3}}) \leq 2\exp(-\delta_1'' n), \quad \forall n,$$

where $\delta_1'' = \frac{\varphi'^2(\|\gamma^*_{\mathbb{A}_2}\|_{\min} - a\tilde{\lambda}_1)^2}{16\eta_0^2 s_2 M} \wedge \frac{\varphi'(\|\gamma^*_{\mathbb{A}_2}\|_{\min} - a\tilde{\lambda}_1)}{8\eta_0\sqrt{s_2 M}}$. Combining this result and ((A.1.4)), we have $p_3' \leq 2s_2\exp(-\delta_1'' n)$ for all n.

Finally, for p_4', by ((A.1.2)) in Lemma 4 we have

$$\begin{aligned} p_4' &= \mathrm{P}\big(\mathcal{T}^{\mathrm{c}}_{K\wedge\frac{C_1}{(4L+2)G_1}\wedge\frac{C_2}{(4L+2)G_2}\wedge\frac{C_3}{(4L+2)G_3}}\big) \\ &\leq n\exp\Big(-\frac{(K\wedge\frac{C_1}{(4L+2)G_1}\wedge\frac{C_2}{(4L+2)G_2}\wedge\frac{C_3}{(4L+2)G_3})^2\Psi^2\varphi}{4\sigma^2\Omega^2 s_1 M}n\Big), \forall n. \end{aligned}$$

Thus (ii)' is proved.

Now, we prove (iii) of Theorem 1. Due to symmetry, we only need to prove the result for $\hat{\gamma}(Z^{(2)} \to Z^{(1)})$. So we slightly abuse notation in this proof. We let $\mathbf{y}$ and $\mathbf{X}$ be the response and design matrix in $Z^{(1)}$, i.e. $\mathbf{y} = (y_1, \ldots, y_{n/2})^{\mathrm{T}}$ and $\mathbf{X} = \mathbf{X}^{(1)}$.

By (ii) of Theorem 1, it is easy to verify that if we take

$$\begin{aligned} \tilde{\lambda} &= T\Big(\sqrt{\frac{\log(p-s_1)}{n}} \vee \sqrt{\frac{\log s_1}{n}} \vee \sqrt{\frac{\log p}{n}}\Big), \\ \tilde{\lambda}_1 &= T\Big(s_2\sqrt{\frac{\log(p-s_2)}{n}} \vee \frac{s_2\log(p-s_2)}{n} \vee \sqrt{\frac{s_2\log s_2}{n}} \vee \frac{\sqrt{s_2}\log s_2}{n} \vee s_2\sqrt{\frac{s_1\log n}{n}} \\ &\qquad \vee \sqrt{\frac{s_1 s_2\log n}{n}} \vee \sqrt{\frac{\log p}{n}} \vee \frac{\log p}{n} \vee \sqrt{\frac{s_1\log n}{n}}\Big) \end{aligned}$$

with sufficiently large $T > 0$, then we can make $\mathrm{P}(\hat{\gamma}(Z^{(2)} \to Z^{(1)}) \neq \tilde{\gamma}^{\mathrm{ora}}_{\bullet}) \to 0$ as $n \to \infty$. In terms of order, this can be simplified to

$$\tilde{\lambda} \asymp \sqrt{\frac{\log p}{n}}, \ \tilde{\lambda}_1 \asymp \Big(s_2\sqrt{\frac{\log(p-s_2)}{n}} \vee \frac{s_2\log(p-s_2)}{n} \vee s_2\sqrt{\frac{s_1\log n}{n}}\Big).$$

With suitably chosen tuning parameters described above, we have $\lim_{n\to\infty}\mathrm{P}(\hat{\gamma}(Z^{(2)} \to Z^{(1)}) \neq \tilde{\gamma}^{\mathrm{ora}}_{\bullet}) = 0$.

Next, we show $\mathbb{E}[\|\tilde{\gamma}^{\mathrm{ora}}_{\bullet} - \gamma^*\|^2_{\ell_2}] = O(\frac{s}{n})$ that yields $\|\tilde{\gamma}^{\mathrm{ora}}_{\bullet} - \gamma^*\|_{\ell_2} = O_p(\sqrt{\frac{s}{n}})$, which

further implies $\|\hat{\gamma}(Z^{(2)} \to Z^{(1)}) - \gamma^*\|_{\ell_2} = O_p(\sqrt{\frac{s}{n}})$. In fact,

$$\begin{aligned}
&\mathbb{E}[\|\tilde{\gamma}^{\text{ora}}_{\bullet} - \gamma^*\|^2_{\ell_2}] = \mathbb{E}[\|(\mathbf{X}^{\mathrm{T}}_{\mathbb{A}_2}\mathbf{X}_{\mathbb{A}_2})^{-1}\mathbf{X}^{\mathrm{T}}_{\mathbb{A}_2}(\mathbf{z} - \mathbf{X}\gamma^*)\|^2_{\ell_2}] \\
&= \mathbb{E}[(\mathbf{z} - \mathbf{X}\gamma^*)^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2}(\mathbf{X}^{\mathrm{T}}_{\mathbb{A}_2}\mathbf{X}_{\mathbb{A}_2})^{-2}\mathbf{X}^{\mathrm{T}}_{\mathbb{A}_2}(\mathbf{z} - \mathbf{X}\gamma^*)] \\
&= \mathbb{E}[(\mathbf{z} - \mathbf{X}\gamma^*)^{\mathrm{T}}\mathbf{A}(\mathbf{z} - \mathbf{X}\gamma^*)], \qquad \text{(A.1.6)}
\end{aligned}$$

where we denote $\mathbf{A} \coloneqq \mathbf{X}_{\mathbb{A}_2}(\mathbf{X}^{\mathrm{T}}_{\mathbb{A}_2}\mathbf{X}_{\mathbb{A}_2})^{-2}\mathbf{X}^{\mathrm{T}}_{\mathbb{A}_2} = (a_{ij})_{1\le i,j\le \frac{n}{2}}$.

Let $v_i = \mathrm{e}^{\mathbf{x}_i^{\mathrm{T}}\gamma^*}, i = 1, \ldots, n$. We have $y_i = \mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}\beta^*_{\mathbb{A}_1} + v_i\epsilon_i, i = 1, \ldots, \frac{n}{2}$. Let $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$ be the response and the design matrix in $Z^{(2)}$, i.e. $\tilde{\mathbf{y}} = (y_{\frac{n}{2}+1}, \ldots, y_n)^{\mathrm{T}}$, and $\tilde{\mathbf{X}} = \mathbf{X}^{(2)}$. Also, let $\tilde{v}_i \coloneqq v_{i+n/2}, i = 1, \ldots, \frac{n}{2}$. Then let $\tilde{\mathbf{W}}^* = \mathbf{diag}\{\tilde{v}_1, \ldots, \tilde{v}_{n/2}\}$ and let $\tilde{\epsilon} \coloneqq (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_{n/2}) \coloneqq (\epsilon_{n/2+1}, \ldots, \epsilon_n)$. Thus we can write $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}_{\mathbb{A}_1}\beta^*_{\mathbb{A}_1} + \tilde{\mathbf{W}}^*\tilde{\epsilon}$, and we have $\hat{\beta}^{\text{ora2}}_{\mathbb{A}_1} = (\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbb{A}_1}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbb{A}_1}\tilde{\mathbf{y}}$.

Let $\zeta_i = \frac{1}{v_i}\mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}(\hat{\beta}^{\text{ora2}}_{\mathbb{A}_1} - \beta^*_{\mathbb{A}_1}) = \frac{1}{v_i}\mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}(\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbb{A}_1}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbb{A}_1}\tilde{\mathbf{W}}^*\tilde{\epsilon}, i = 1, \ldots, \frac{n}{2}$. Then, notice that

$$\begin{aligned}
&\log|y_i - \mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}\hat{\beta}^{\text{ora2}}_{\mathbb{A}_1}| - \mathbf{x}_i^{\mathrm{T}}\gamma^* = \log|v_i\epsilon_i - \mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}(\hat{\beta}^{\text{ora2}}_{\mathbb{A}_1} - \beta^*_{\mathbb{A}_1})| - \mathbf{x}_i^{\mathrm{T}}\gamma^* \\
&= \log|\epsilon_i - \frac{1}{v_i}\mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}(\hat{\beta}^{\text{ora2}}_{\mathbb{A}_1} - \beta^*_{\mathbb{A}_1})| = \log|\epsilon_i - \zeta_i|, \forall i = 1, \ldots, \frac{n}{2}.
\end{aligned}$$

Let $\eta' = (\eta'_1, \ldots, \eta'_{n/2})$ with $\eta'_i = \log|\epsilon_i - \zeta_i|$. By law of iterated expectation, we have

$$\begin{aligned}
&\mathbb{E}[(\mathbf{z} - \mathbf{X}\gamma^*)^{\mathrm{T}}\mathbf{A}(\mathbf{z} - \mathbf{X}\gamma^*)] = \mathbb{E}[\eta'^{\mathrm{T}}\mathbf{A}\eta'] = \mathbb{E}[\mathbb{E}[\eta'^{\mathrm{T}}\mathbf{A}\eta'\big|\tilde{\epsilon}]] \\
&= \mathbb{E}\big[\mathbb{E}[(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}])^{\mathrm{T}}\mathbf{A}(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}])\big|\tilde{\epsilon}]\big] + \mathbb{E}\big[\mathbb{E}[\eta'|\tilde{\epsilon}]^{\mathrm{T}}\mathbf{A}\mathbb{E}[\eta'|\tilde{\epsilon}]\big]. \qquad \text{(A.1.7)}
\end{aligned}$$

Given $\tilde{\epsilon}$, $\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}]$ has independent components with mean zero, therefore

$$\begin{aligned}
&\mathbb{E}\big[\mathbb{E}[(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}])^{\mathrm{T}}\mathbf{A}(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}])\big|\tilde{\epsilon}]\big] \\
&= \mathbb{E}\big[\mathbb{E}[\sum_{i=1}^{n/2} a_{ii}(\log|\epsilon_i - \zeta_i| - \mathbb{E}[\log|\epsilon_i - \zeta_i|\big|\tilde{\epsilon}])^2\big|\tilde{\epsilon}]\big] \\
&\le \sum_{i=1}^{n/2} a_{ii}\mathbb{E}\big[\mathbb{E}[\log^2|\epsilon_i - \zeta_i|\big|\tilde{\epsilon}]\big] \qquad \text{(A.1.8)}
\end{aligned}$$

Notice that $a_{ii} = \mathbf{e}_i^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2}(\mathbf{X}^{\mathrm{T}}_{\mathbb{A}_2}\mathbf{X}_{\mathbb{A}_2})^{-2}\mathbf{X}^{\mathrm{T}}_{\mathbb{A}_2}\mathbf{e}_i \le \frac{4}{n^2\varphi'^2}\|\mathbf{x}_{i\mathbb{A}_2}\|^2_{\ell_2} \le \frac{4s_2M}{n^2\varphi'^2}, \forall i = 1, \ldots, \frac{n}{2}$.

Combining this fact and Lemma 3 (ii), we have

$$\sum_{i=1}^{n/2} a_{ii}\mathbb{E}\big[\mathbb{E}[\log^2|\epsilon_i-\zeta_i|\big|\tilde{\epsilon}]\big] \le \frac{4s_2M}{n^2\varphi'^2}\sum_{i=1}^{n/2}\mathbb{E}\big[\mathbb{E}[\log^2|\epsilon_i-\zeta_i|\big|\tilde{\epsilon}]\big]$$
$$\le \frac{4s_2M}{n^2\varphi'^2}\sum_{i=1}^{n/2}\Big\{\mathbb{E}\left[\mathbb{E}[\log^2|\epsilon_i|]+\zeta_i^2+(4L+2\mathbb{E}\big[\big|\log|\epsilon_i|\big|\big])\cdot|\zeta_i|\right]\Big\}$$
$$\le \frac{4s_2M}{n^2\varphi'^2}\sum_{i=1}^{n/2}\Big\{\mathbb{E}[\log^2|\epsilon_1|]+\mathbb{E}[\zeta_i^2]+(4L+2\mathbb{E}\big[\big|\log|\epsilon_1|\big|\big])\sqrt{\mathbb{E}[\zeta_i^2]}\Big\}. \tag{A.1.9}$$

For the last term in ((A.1.7)), notice $\lambda_{\max}(\mathbf{A}) = \lambda_{\max}(\mathbf{X}_{\mathbb{A}_2}(\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2})^{-2}\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}) = \lambda_{\max}((\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2})^{-1}) \le \frac{2}{n\varphi'}$. Combining this fact and Lemma 3 (i), we have

$$\mathbb{E}\big[\mathbb{E}[\eta'|\tilde{\epsilon}]^{\mathrm{T}}\mathbf{A}\mathbb{E}[\eta'|\tilde{\epsilon}]\big] \le \frac{2}{n\varphi'}\sum_{i=1}^{n/2}\mathbb{E}\big[\mathbb{E}[\log|\epsilon_i-\zeta_i|\big|\tilde{\epsilon}]^2\big]$$
$$\le \frac{2}{n\varphi'}(2L+1)^2\sum_{i=1}^{n/2}\mathbb{E}[\zeta_i^2]. \tag{A.1.10}$$

It remains to bound $\mathbb{E}[\zeta_i^2]$ for any $1\le i\le\frac{n}{2}$. By the assumptions $(\mathbf{C}_1)$ and $(\mathbf{C}_3)$ we have

$$\mathbb{E}[\zeta_i^2] = \mathbb{E}[\frac{1}{v_i^2}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{W}}^*\tilde{\epsilon}\tilde{\epsilon}^{\mathrm{T}}\tilde{\mathbf{W}}^*\tilde{\mathbf{X}}_{\mathbb{A}_1}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\mathbf{x}_{i\mathbb{A}_1}]$$
$$= \mathrm{var}(\epsilon_1)\frac{1}{v_i^2}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{W}}^{*2}\tilde{\mathbf{X}}_{\mathbb{A}_1}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\mathbf{x}_{i\mathbb{A}_1}$$
$$\le \mathrm{var}(\epsilon_1)\frac{\Omega^2}{\Psi^2}\cdot\frac{2}{n\varphi}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}\mathbf{x}_{i\mathbb{A}_1} \le \mathrm{var}(\epsilon_1)\frac{\Omega^2}{\Psi^2}\frac{2s_1M}{n\varphi}. \tag{A.1.11}$$

Because ϵ_1 is sub-Gaussian, $\log|\epsilon_1|$ has sub-exponential tail by Lemma 2. So $\mathbb{E}[\log^2|\epsilon_1|] = O(1)$, and by Cauchy-Schwarz inequality, $\mathbb{E}[\big|\log|\epsilon_1|\big|] \le \sqrt{\mathbb{E}[\log^2|\epsilon_1|]}$, so $\mathbb{E}[\big|\log|\epsilon_1|\big|] = O(1)$. Collecting the results in ((A.1.6)), ((A.1.7)), ((A.1.8)), ((A.1.9)), ((A.1.10)) and ((A.1.11)), we have

$$\mathbb{E}[\|\tilde{\gamma}_\bullet^{\mathrm{ora}}-\gamma^*\|_{\ell_2}^2] \le \frac{4s_2M}{n^2\varphi'^2}\cdot\frac{n}{2}\Big\{\mathbb{E}[\log^2|\epsilon_1|]+\mathrm{var}(\epsilon_1)\frac{\Omega^2}{\Psi^2}\frac{2s_1M}{n\varphi}$$

$$+ (4L + 2\mathbb{E}\big[\big|\log|\epsilon_1|\big|\big])\sqrt{\mathrm{var}(\epsilon_1)\frac{\Omega^2}{\Psi^2}\frac{2s_1M}{n\varphi}}\Big\}$$
$$+ \frac{2}{n\varphi'}(2L+1)^2 \cdot \frac{n}{2}\Big\{\mathrm{var}(\epsilon_1)\frac{\Omega^2}{\Psi^2}\frac{2s_1M}{n\varphi}\Big\}$$
$$= O(\frac{s_2}{n}) + O(\frac{s_1}{n}) = O(\frac{s}{n}).$$

This completes the proof of (iii). □

A.1.3 Proof of Theorem 2

Proof A.2

For the sake of space, we sketch the proof of Theorem 2 here and defer details to Appendix Appendix A.8 of the supplementary material. By symmetry, we only need to prove the result for $\mathrm{HBIC}^{(3)}$, i.e. $\mathrm{P}(M^{(3)}_{\hat{\tilde{\lambda}}_1} = \mathbb{A}_2) \to 1$. So we slightly abuse notation in this proof. We use $\mathbf{y}$ and $\mathbf{X}$ to represent $\mathbf{y}^{(1)}$ and $\mathbf{X}^{(1)}$. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_{\frac{n}{2}})$ and let $\tilde{\epsilon} = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_{n/2})^{\mathrm{T}}$ with $\tilde{\epsilon}_i = \epsilon_{i+\frac{n}{2}}$, $i = 1, \ldots, \frac{n}{2}$. For any $\lambda > 0$, we use $\hat{\gamma}^\lambda$ to represent $\hat{\gamma}^\lambda(Z^{(2)} \to Z^{(1)})$, and use M_λ to represent $M^{(3)}_\lambda$, which is the support of $\hat{\gamma}^\lambda$. Also, we use HBIC to represent $\mathrm{HBIC}^{(3)}$, and we use $C_{n,p}$ to represent $C^{(3)}_{n,p}$. For any index set $A \subset \{1, \ldots, p\}$, we use $\mathbf{P}_A$ to represent the projection matrix $\mathbf{P}^{(1)}_A$ for convenience.

Let $z_i = \log|y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}^{\mathrm{ora2}}|$, $i = 1, \ldots, \frac{n}{2}$, and let $\mathbf{z} = (z_1, \ldots, z_{n/2})^{\mathrm{T}}$. For any $M \subset \{1, \ldots, p\}$, let $\mathrm{SSE}_M = \inf_{\gamma_M \in \mathbb{R}^{|M|}} \|\mathbf{z} - \mathbf{X}_\mathbf{M}\gamma_\mathbf{M}\|^2_{\ell_2}$, and let $\hat{\sigma}^2_M = \frac{2}{n}\mathrm{SSE}_M$. By definition, we have $\frac{2}{n}\|\mathbf{z} - \mathbf{X}\hat{\gamma}^\lambda\|^2_{\ell_2} \geq \mathrm{SSE}_{M_\lambda}$, $\forall \lambda > 0$. Let $\hat{\gamma}^{\mathrm{ora}}$ be the p-dimensional vector with $\hat{\gamma}^{\mathrm{ora}}_{\mathbb{A}_2} = (\mathbf{X}^{\mathrm{T}}_{\mathbb{A}_2}\mathbf{X}_{\mathbb{A}_2})^{-1}\mathbf{X}^{\mathrm{T}}_{\mathbb{A}_2}\mathbf{z}$ and $\hat{\gamma}^{\mathrm{ora}}_{\mathbb{A}_2^{\mathrm{c}}} = \mathbf{0}$.

We divide the candidate set of tuning parameters into three subsets. In particular, let $\tilde{\Lambda}_- = \{\lambda > 0 : \lambda \in \tilde{\Lambda}_1, \mathbb{A}_2 \not\subset M_\lambda\}$, $\tilde{\Lambda}_0 = \{\lambda > 0 : \lambda \in \tilde{\Lambda}_1, \mathbb{A}_2 = M_\lambda\}$, $\tilde{\Lambda}_+ = \{\lambda > 0 : \lambda \in \tilde{\Lambda}_1, \mathbb{A}_2 \subset M_\lambda \text{ and } \mathbb{A}_2 \neq M_\lambda\}$. From Theorem 1, we know there exists $\tilde{\lambda} = \tilde{\lambda}_n > 0$ such that $\mathrm{P}(\hat{\gamma}^{\tilde{\lambda}_n} = \hat{\gamma}^{\mathrm{ora}}) \to 1$ as $n \to \infty$. This implies that $\mathrm{P}(\tilde{\lambda}_n \in \tilde{\Lambda}_0) \to 1$. Therefore, it suffices to show **(i)**: $\mathrm{P}\big(\inf_{\lambda\in\tilde{\Lambda}_-} \mathrm{HBIC}(\lambda) > \mathrm{HBIC}(\tilde{\lambda}_n)\big) \to 1$ and **(ii)**: $\mathrm{P}\big(\inf_{\lambda\in\tilde{\Lambda}_+} \mathrm{HBIC}(\lambda) > \mathrm{HBIC}(\tilde{\lambda}_n)\big) \to 1$.

For (i), we can show that

$$\mathrm{P}(\inf_{\lambda\in\tilde{\Lambda}_-}[\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\tilde{\lambda}_n)] > 0)$$
$$\geq \mathrm{P}(\inf_{\lambda\in\tilde{\Lambda}_-}[\log(\hat{\sigma}^2_{M_\lambda}/\hat{\sigma}^2_{\mathbb{A}_2}) + (|M_\lambda| - s_2)\frac{C_{n,p}\log(p)}{n}] > 0) + o(1), \tag{A.1.12}$$

which follows from Theorem 1 and the property of $\hat{\sigma}^2_{M_\lambda}$. Let $\eta' = \mathbf{z} - \mathbf{X}\gamma^*$. Then we have $\hat{\sigma}^2_{\mathbb{A}_2} = \frac{2}{n}\|(\mathbf{l}_{n/2} - \mathbf{P}_{\mathbb{A}_2})\eta'\|^2_{\ell_2}$, and

$$\log\left(\frac{\widehat{\sigma}^2_{M_\lambda}}{\widehat{\sigma}^2_{\mathbb{A}_2}}\right) = \log\left(1 + \frac{n\left(\widehat{\sigma}^2_{M_\lambda} - \widehat{\sigma}^2_{\mathbb{A}_2}\right)/2}{\eta'^T\left(\mathbf{I}_{n/2} - \mathbf{P}_{\mathbb{A}_2}\right)\eta'}\right). \tag{A.1.13}$$

To evaluate $\eta'^T\left(\mathbf{I}_{n/2} - \mathbf{P}_{\mathbb{A}_2}\right)\eta'$, notice that η' involves $\hat{\beta}^{\text{ora2}}$, whose randomness comes from $\tilde{\epsilon}$. By applying conditioning argument and Lemma 3, we can establish $\eta'^T\left(\mathbf{I}_{n/2} - \mathbf{P}_{\mathbb{A}_2}\right)\eta' = O_p(n)$, which means

$$\lim_{T\to\infty}\limsup_{n\to\infty}\mathrm{P}(\eta'^T\left(\mathbf{I}_{n/2} - \mathbf{P}_{\mathbb{A}_2}\right)\eta' > Tn) = 0. \tag{A.1.14}$$

To evaluate $n(\hat{\sigma}^2_{M_\lambda} - \hat{\sigma}^2_{\mathbb{A}_2})/2$ for any $\lambda \in \tilde{\Lambda}_-$, let $\mu = \mathbf{X}_{\mathbb{A}_2}\gamma^*_{\mathbb{A}_2}$. We break the target into four terms as follows,

$$\begin{aligned}\frac{n}{2}(\hat{\sigma}^2_{M_\lambda} - \hat{\sigma}^2_{\mathbb{A}_2}) &= \mu^T(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\mu + 2\mu^T(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\eta' - \eta'^T\mathbf{P}_{M_\lambda}\eta' + \eta'^T\mathbf{P}_{\mathbb{A}_2}\eta' \\ &\triangleq I_1 + I_2 - I_3 + I_4. \end{aligned}\tag{A.1.15}$$

By condition ((2.5.2)), $I_1 = \mu^T(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\mu > nc_0'$ for sufficiently large n. Evaluating I_2, I_3 and I_4 are technically challenging. The concentration inequalities we developed to deal with these terms include Lemma 4 as well as the following lemma.

Lemma 5 Assume assumptions $(\mathbf{A}_0)$, $(\mathbf{C}_1)$, $(\mathbf{C}_3)$ hold. Let η' be $\frac{n}{2}$-dimensional vector with $\eta'_i = \log|y_i - \mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}\hat{\beta}^{\text{ora2}}_{\mathbb{A}_1}| - \mathbf{x}^{\mathrm{T}}_i\gamma^* = \log|\epsilon_i - \zeta_i|$, where $\zeta_i = \frac{1}{v_i}\mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}(\hat{\beta}^{\text{ora2}}_{\mathbb{A}_1} - \beta^*_{\mathbb{A}_1})$ and $v_i = \mathrm{e}^{\mathbf{x}_i^{\mathrm{T}}\gamma^*}, i = 1, \ldots, \frac{n}{2}$. Let $\delta > 0$ be any positive real number. Let $\mathbf{P} = (P_{ij})_{1\le i,j\le\frac{n}{2}}$ be a $\frac{n}{2}$-dimensional projection matrix (i.e. $P^2 = P$ and $P = P^{\mathrm{T}}$) of rank m, and satisfying $P_{ii} \le \frac{2}{n}G'$ for some $G' > 0$. Then we have for any $C > 0$,

$$\begin{aligned}&\mathrm{P}(\|\mathbf{P}\eta'\|_{\ell_2} > \frac{C}{(1-2\delta)_+}, \mathcal{T}_{K\wedge\frac{\sqrt{2}C}{(4L+2)\sqrt{nG'}}}) \\ &\le 2(1+\frac{1}{\delta})^m \exp\left[-\left(\frac{2C^2}{16\eta_0^2\sqrt{n}G'} \wedge \frac{\sqrt{2}C}{8\eta_0\sqrt{G'}}\right)\sqrt{n}\right],\end{aligned}$$

where $\mathcal{T}_x$ is a event that has been defined in Lemma 4 for any $x > 0$, and K is any

fixed positive number. □

The proof of Lemma 5 is deferred to Appendix Appendix A.7 of the supplementary material. With these two lemmas, we are able to show that $\sup_{\lambda\in\tilde{\Lambda}_-}|I_2| = o_p(n)$, $\sup_{\lambda\in\tilde{\Lambda}_1}|I_3| = o_p(n)$ and $I_4 = o_p(n)$ (notice that I_4 does not depend on λ). These results together with ((A.1.15)) imply that $\mathrm{P}(\frac{n}{2}(\hat{\sigma}^2_{M_\lambda} - \hat{\sigma}^2_{\mathbb{A}_2}) > \frac{c_0'}{2}n) \to 1$. Then by ((A.1.13)) and ((A.1.14)), for any $T > 0$ there exists $\epsilon_T > 0$ that depends only on T and satisfies $\epsilon_T \to 0$ as $T \to \infty$, such that

$$\liminf_{n\to\infty} \mathrm{P}\Big(\inf_{\lambda\in\tilde{\Lambda}_-} \log(\frac{\hat{\sigma}^2_{M_\lambda}}{\hat{\sigma}^2_{\mathbb{A}_2}}) > \log(1+\frac{c_0'}{2T})\Big) \geq 1-\epsilon_T.$$

Following ((A.1.12)), we have

$$\begin{aligned}
&\liminf_{n\to\infty} \mathrm{P}(\inf_{\lambda\in\tilde{\Lambda}_-}[\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\tilde{\lambda}_n)] > 0)\\
&\geq \liminf_{n\to\infty} \mathrm{P}\Big(\log(1+\frac{c_0'}{2T}) - \frac{C_{n,p}s_2\log(p)}{n} > 0\Big)\\
&\geq 1-\epsilon_T,
\end{aligned}$$

where the last step is because $\frac{C_{n,p}s_2\log(p)}{n} = o(1)$. Note the left hand side of the above inequality chain does not depend on T, so **(i)** is proved by letting $T \to \infty$.

For (ii), consider any $\lambda \in \tilde{\Lambda}_+$. Then we have $\mathbb{A}_2 \subset M_\lambda$ and $\mathbb{A}_2 \neq M_\lambda$. So by the relation $\mathbf{z} = \mathbf{X}\gamma^* + \eta'$, we have $\mathbf{z}^{\mathrm{T}}(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\mathbf{z} = \eta'^{\mathrm{T}}(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\eta'$. Therefore we have $\frac{n}{2}(\hat{\sigma}^2_{\mathbb{A}_2} - \hat{\sigma}^2_{M_\lambda}) = \eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2})\eta'$. Therefore, we have

$$0 \leq \log(\frac{\hat{\sigma}^2_{\mathbb{A}_2}}{\hat{\sigma}^2_{M_\lambda}}) = \log(1+\frac{\eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2})\eta'}{\eta'^{\mathrm{T}}(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\eta'}) \leq \frac{\eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2})\eta'}{\eta'^{\mathrm{T}}(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\eta'},$$

where the last step is due to $\log(1+x) \leq x, \forall x \geq 0$. Similar to ((A.1.12)), we can show

$$\begin{aligned}
&\mathrm{P}(\inf_{\lambda\in\tilde{\Lambda}_+}[\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\tilde{\lambda}_n)] > 0)\\
\geq&\mathrm{P}\Bigg(\inf_{\lambda\in\tilde{\Lambda}_+}\Big[\Big(|M_\lambda| - s_2\Big)\Big(\frac{C_{n,p}\log p}{n} - \frac{\eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2})\eta'/(|M_\lambda| - s_2)}{\eta'^{\mathrm{T}}(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\eta'}\Big)\Big] > 0\Bigg)\\
&+ o(1).
\end{aligned}$$

Because $|M_\lambda| - s_2 \geq 1$, it suffices to show

$$\mathrm{P}\left(\inf_{\lambda \in \tilde{\Lambda}_+} \left[\left(\frac{C_{n,p} \log p}{n} - \frac{\eta'^{\mathrm{T}} (\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}) \eta' / (|M_\lambda| - s_2)}{\eta'^{\mathrm{T}} (\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda}) \eta'} \right) \right] > 0 \right) \to 1 \qquad \text{(A.1.16)}$$

as $n \to \infty$. We first evaluate $\eta'^{\mathrm{T}} (\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda}) \eta'$. We can write it as the sum of two terms, $\eta'^{\mathrm{T}} (\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda}) \eta' = \eta'^{\mathrm{T}} (\mathbf{l}_{n/2} - \mathbf{P}_{\mathbb{A}_2}) \eta' - \eta'^{\mathrm{T}} (\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}) \eta' \triangleq I_5 - I_6$. For I_5, we have $I_5 = \eta'^{\mathrm{T}} (\mathbf{l}_{n/2} - \mathbf{P}_{\mathbb{A}_2}) \eta' = \eta'^{\mathrm{T}} \eta' - I_4$, notice that this term does not depend on λ. Since we have derived $|I_4| = o_p(n)$, we need to evaluate $\eta'^{\mathrm{T}} \eta' = (\mathbf{z} - \mathbf{X}\gamma^*)^{\mathrm{T}} (\mathbf{z} - \mathbf{X}\gamma^*)$. In particular, we show $\frac{\eta'^{\mathrm{T}} \eta'}{n/2} \xrightarrow{\mathrm{P}} \mathbb{E}[\log^2 |\epsilon_1|]$. Note that components in η' are neither independent nor identically distributed, so the law of large numbers does not apply. Instead, we show this by applying a conditioning technique, Lemma 3 and a tail integration argument to bound the deviation probability $\mathrm{P}(|\frac{\eta'^{\mathrm{T}} \eta'}{n/2} - \mathbb{E}[\log^2 |\epsilon_1|]| > t)$ and proving that it converges to zero for any $t > 0$. Therefore, with probability going to 1, we have $I_5 > \frac{n}{4} \mathbb{E}[\log^2 |\epsilon_1|]$.

For $I_6 = \eta'^{\mathrm{T}} (\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}) \eta'$, we have $0 \leq I_6 \leq \eta'^{\mathrm{T}} \mathbf{P}_{M_\lambda} \eta' = I_3$. Since we have already shown that $\sup_{\lambda \in \tilde{\Lambda}_1} |I_3| = o_p(n)$, and $\tilde{\Lambda}_+ \subset \tilde{\Lambda}_1$, we have $\sup_{\lambda \in \tilde{\Lambda}_+} |I_6| = o_p(n)$. Therefore, with probability going to 1, we have $\eta'^{\mathrm{T}} (\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda}) \eta' = I_5 - I_6 > \frac{n}{8} \mathbb{E}[\log^2 |\epsilon_1|]$ holds for all $\lambda \in \tilde{\Lambda}_+$.

It remains to evaluate the term $\eta'^{\mathrm{T}} (\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}) \eta' / (|M_\lambda| - s_2) \triangleq I_7$ in ((A.1.16)). The term $|M_\lambda| - s_2$ in the denominator makes I_7 complicated, and we need to bound the supreme of I_7 over $\lambda \in \tilde{\Lambda}_+$, which makes the problem more challenging. To bound I_7, we will first consider an event $\mathcal{T}_{\sqrt{Ts_1 \log n/n}}$ with $T > 0$ being temporarily fixed, under which the vector η' behaves well as a sub-exponential random vector. Here $\mathcal{T}_x$ refers to the event that has been defined in Lemma 4 for any $x > 0$. Then we apply the results in Götze et al. (2021) which deal with the quadratic forms of sub-exponential vector and establish the concentration of $(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}])^{\mathrm{T}} (\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}) (\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}])$ around its conditional expectation given $\tilde{\epsilon}$. Then we apply Lemma 3 to evaluate the difference between the conditional expectation and a term $\mathbb{E}[\log^2 |\epsilon_1|](|M_\lambda| - s_2)$, and show the difference is asymptotically negligible. Next, applying Lemma 3 again, we bound a term $\mathbb{E}[\eta'|\tilde{\epsilon}]^{\mathrm{T}} (\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}) \mathbb{E}[\eta'|\tilde{\epsilon}]$ and show that it is asymptotically negligible under $\mathcal{T}_{\sqrt{Ts_1 \log n/n}}$. By these results and the inequality $\eta'^{\mathrm{T}} (\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}) \eta' \leq 2(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}])^{\mathrm{T}} (\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}) (\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}]) + 2\mathbb{E}[\eta'|\tilde{\epsilon}]^{\mathrm{T}} (\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}) \mathbb{E}[\eta'|\tilde{\epsilon}]$, we then get an upper bound for $\mathrm{P}\left(\eta'^{\mathrm{T}} (\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}) \eta' - 2\mathbb{E}[\log^2 |\epsilon_1|](|M_\lambda| - s_2) \geq t, \mathcal{T}_{\sqrt{Ts_1 \log n/n}} \right)$

for any $t > 0$. By carefully evaluating a union bound and choosing an appropriate value $t = c(m \log p \vee s_1 \log n)$ with some suitable constant c, we show $\mathrm{P}\big(\sup_{\lambda \in \tilde{\Lambda}_+} I_7 \geq 2\mathbb{E}[\log^2 |\epsilon_1|] + c(\log p \vee s_1 \log n), \mathcal{T}_{\sqrt{Ts_1 \log n/n}}\big) \to 0$ as $n \to \infty$. Moreover, by choosing T large enough and applying Lemma 4, we show $\mathrm{P}(\mathcal{T}^{\mathrm{c}}_{\sqrt{Ts_1 \log n/n}}) \to 0$ as $n \to \infty$. Once again by a union bound, we concludes that $\mathrm{P}\big(\sup_{\lambda \in \tilde{\Lambda}_+} I_7 \geq c'(\log p \vee s_1 \log n)\big) \to 0$ as $n \to \infty$, for some constant $c' > 0$.

Summarizing the previous results, we get

$$\begin{aligned}
&\mathrm{P}\left(\inf_{\lambda \in \tilde{\Lambda}_+} \left[\frac{C_{n,p} \log p}{n} - \frac{\eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2})\eta'/(|M_\lambda| - s_2)}{\eta'^{\mathrm{T}}(\mathbf{I}_{n/2} - \mathbf{P}_{M_\lambda})\eta'}\right] > 0\right) \\
&\geq \mathrm{P}\left(\frac{C_{n,p} \log p}{n} - \frac{c'(\log p \vee s_1 \log n)}{\frac{n}{8}\mathbb{E}[\log^2 |\epsilon_1|]} > 0\right) + o(1) \to 1,
\end{aligned}$$

since $C_{n,p} \to \infty$ and $s_1 \log n = o(C_{n,p} \log p)$. So ((A.1.16)) is proved and the conclusion of Theorem 2 follows. □

A.2 Proofs of Proposition 2-4 and Lemma 1-2.

Proposition 2 is Corollary 1.7 in Rigollet and Hütter (2015), so we omit the proof here. Proposition 3 is standard and has similar argument such as Proposition 5.16 in Vershynin (2012). Due to the difference in constants, we give its proof here for completeness. The proof of Proposition 4 is also standard, which can be found in Hastie et al. (2015) and Fan et al. (2020). We omit its proof here.

A.2.1 Proof of Proposition 3

Proof A.3

For $t > 0$, any $0 < s \leq \frac{1}{\lambda \max_{i=1}^n |a_i|}$, we have

$$\mathrm{P}(\sum_{i=1}^n a_i\epsilon_i > t) \leq \frac{\mathbb{E}\mathrm{e}^{s\sum_{i=1}^n a_i\epsilon_i}}{\mathrm{e}^{st}} = \frac{\prod_{i=1}^n \mathbb{E}\mathrm{e}^{sa_i\epsilon_i}}{\mathrm{e}^{st}} \leq \frac{\prod_{i=1}^n \mathrm{e}^{\frac{s^2a_i^2\lambda^2}{2}}}{\mathrm{e}^{st}} \leq \mathrm{e}^{\frac{\lambda^2s^2\sum_{i=1}^n a_i^2}{2}-st}.$$

Choosing $s = \frac{t}{\lambda^2\sum_{i=1}^n a_i^2} \wedge \frac{1}{\lambda \max_{i=1}^n |a_i|}$, we get

$$\mathrm{P}(\sum_{i=1}^n a_i\epsilon_i > t) \leq \exp\left[-\left(\frac{t^2}{2\lambda^2\sum_{i=1}^n a_i^2} \wedge \frac{t}{2\lambda \max_{1\leq i\leq n}|a_i|}\right)\right].$$

Applying this inequality on $\{-\epsilon_i\}_{i=1}^n$, we have

$$\mathrm{P}(\sum_{i=1}^n a_i(-\epsilon_i) > t) \leq \exp\left[-\left(\frac{t^2}{2\lambda^2\sum_{i=1}^n a_i^2} \wedge \frac{t}{2\lambda \max_{1\leq i\leq n}|a_i|}\right)\right].$$

Combining these two inequalities, the proof is completed. □

A.2.2 Proof of Lemma 2

Proof A.4

We first claim that $\mathrm{P}(\big|\log|\epsilon_1+\mu| - \mathbb{E}[\log|\epsilon_1+\mu|]\big| > t) \leq a\mathrm{e}^{-t}$ for any $t > 0$, where $a = (4\sigma^2\mathrm{e}^{(4L+2)c}) \vee (2c\mathrm{e}^{(2L+1)c}) \vee (2 + 2C_0\mathrm{e}^{(2L+1)c})$ is a positive constant. To prove this, let $b = \mathbb{E}[\log|\epsilon_1+\mu|]$, then by assumption $(\mathbf{A}_0)$ and Lemma 3, we have $|b| \leq (2L+1)|\mu| \leq (2L+1)c$. When $t > \log|2\mu| - b$, we have $\mathrm{e}^{b+t} > 2|\mu|$, then by

Proposition 2 and assumption ($\mathbf{A}_0$),

$$\begin{aligned}&\mathrm{P}(\big|\log|\epsilon+\mu|-\mathbb{E}[\log|\epsilon+\mu|]\big|>t)=\mathrm{P}(|\epsilon+\mu|>\mathrm{e}^{t+b})+\mathrm{P}(|\epsilon+\mu|<\mathrm{e}^{b-t})\\&\leq \mathrm{P}(|\epsilon|>\frac{1}{2}\mathrm{e}^{t+b})+\int_{-\mathrm{e}^{b-t}-\mu}^{\mathrm{e}^{b-t}-\mu}\mathrm{f}(x)\,\mathrm{d}x\leq 2\mathrm{e}^{-\frac{\mathrm{e}^{2b+2t}}{8\sigma^2}}+2C_0\mathrm{e}^{b}\mathrm{e}^{-t}.\end{aligned}$$

When $t>\log 4\sigma^2-2b$, we have $\mathrm{e}^{2b+2t}\geq 4\sigma^2\mathrm{e}^t\geq 4\sigma^2\mathrm{e}^{\frac{2}{\mathrm{e}}t}$. Also, let $h(t)=\mathrm{e}^{\frac{2}{\mathrm{e}}t}-2t$, we have $h'(t)=\frac{2}{\mathrm{e}}\mathrm{e}^{\frac{2}{\mathrm{e}}t}-2$, so $h'(t)<0$ when $t<\frac{\mathrm{e}}{2}$, and $h'(t)>0$ when $t>\frac{\mathrm{e}}{2}$, which means $h(t)\geq h(\frac{\mathrm{e}}{2})=0$. So when $t>\log 4\sigma^2-2b$, $\mathrm{e}^{2b+2t}\geq 4\sigma^2 2t=8\sigma^2 t$, and $\mathrm{e}^{-\frac{\mathrm{e}^{2b+2t}}{8\sigma^2}}\leq\mathrm{e}^{-t}$. Therefore when $t>(\log 4\sigma^2+2(2L+1)c)\vee(\log 2c+(2L+1)c)$, we will simultaneously have $t>\log 4\sigma^2-2b$ and $t>\log|2\mu|-b$, so that $\mathrm{P}(\big|\log|\epsilon+\mu|-\mathbb{E}[\log|\epsilon+\mu|]\big|>t)\leq(2+2C_0\mathrm{e}^b)\mathrm{e}^{-t}\leq(2+2C_0\mathrm{e}^{(2L+1)c})\mathrm{e}^{-t}$.

When $0<t<(\log 4\sigma^2+2(2L+1)c)\vee(\log 2c+(2L+1)c)$, then we have $\mathrm{e}^{-t}\geq\frac{1}{4\sigma^2}\mathrm{e}^{-(4L+2)c}\wedge\frac{1}{2c}\mathrm{e}^{-(2L+1)c}$, and

$$\mathrm{P}(\big|\log|\epsilon+\mu|-\mathbb{E}[\log|\epsilon+\mu|]\big|>t)\leq 1\leq(4\sigma^2\mathrm{e}^{(4L+2)c})\vee(2c\mathrm{e}^{(2L+1)c})\mathrm{e}^{-t}.$$

So the claim is true. We have, for any positive integer k,

$$\begin{aligned}\mathbb{E}[\big|\log|\epsilon+\mu|-\mathbb{E}[\log|\epsilon+\mu|]\big|^k]&=\int_0^\infty\mathrm{P}(\big|\log|\epsilon+\mu|-\mathbb{E}[\log|\epsilon+\mu|]\big|>t^{\frac{1}{k}})\,\mathrm{d}t\\&\leq\int_0^\infty a\mathrm{e}^{-t^{\frac{1}{k}}}\,\mathrm{d}t\\&=\int_0^\infty a\mathrm{e}^{-u}ku^{k-1}\,\mathrm{d}u\\&=ak!.\end{aligned}$$

Then we have for any s such that $|s|\leq\frac{1}{2}$,

$$\begin{aligned}\mathbb{E}\mathrm{e}^{s(\log|\epsilon+\mu|-\mathbb{E}[\log|\epsilon+\mu|])}&\leq 1+\sum_{k=2}^\infty\frac{|s|^k\mathbb{E}[\big|\log|\epsilon+\mu|-\mathbb{E}[\log|\epsilon+\mu|]\big|^k]}{k!}\\&\leq 1+\sum_{k=2}^\infty a(|s|)^k\\&=1+as^2\frac{1}{1-|s|}\end{aligned}$$

$$\begin{aligned}
&\leq 1 + 2as^2 \\
&\leq \mathrm{e}^{2as^2} \\
&\leq \mathrm{e}^{\frac{\eta^2 s^2}{2}},
\end{aligned}$$

where the last step holds true for any $\eta \geq 2\sqrt{a}$. Since $\frac{1}{\eta} \leq \frac{1}{2\sqrt{a}} < \frac{1}{2}$, we have $\mathbb{E}\mathrm{e}^{s(\log|\epsilon+\mu|-\mathbb{E}[\log|\epsilon+\mu|])} \leq \mathrm{e}^{\frac{\eta^2 s^2}{2}}$ for any s such that $|s| \leq \frac{1}{\eta}$, for any $\eta \geq 2\sqrt{a}$. This shows that $\log|\epsilon+\mu| - \mathbb{E}[\log|\epsilon+\mu|]$ is sub-exponential(η) random variable for any $\eta \geq 2\sqrt{a}$. □

A.2.3 Proof of Lemma 3

Proof A.5

For any $x > 0$, let $g_1(x) = \log(x)1_{\{x\geq 1\}}$ and $g_2(x) = \log(x)1_{\{x<1\}}$. Notice that g_1 is differentiable almost everywhere, and the derivative has magnitude that is no greater than 1. Therefore we have $|g_1(x) - g_1(y)| \leq |x - y|$ for all $x, y > 0$.

(i): By definition, $\log(x) = g_1(x) + g_2(x)$. Consequently,

$$\begin{aligned}
&\big|\mathbb{E}[\log|\epsilon+\mu|] - \mathbb{E}[\log|\epsilon|]\big| = \Big|\mathbb{E}\big[g_1(|\epsilon+\mu|) - g_1(|\epsilon|) + g_2(|\epsilon+\mu|) - g_2(|\epsilon|)\big]\Big| \\
&\leq \Big|\mathbb{E}\big[g_1(|\epsilon+\mu|) - g_1(|\epsilon|)\big]\Big| + \Big|\mathbb{E}\big[g_2(|\epsilon+\mu|) - g_2(|\epsilon|)\big]\Big| \\
&\leq \mathbb{E}\Big[\big||\epsilon+\mu| - |\epsilon|\big|\Big] + \Big|\int_{|x|<1} \log|x| \cdot \big(f(x-\mu) - f(x)\big)\,\mathrm{d}x\Big| \\
&\leq |\mu| + \int_{|x|<1} \big|\log|x|\big| \cdot \big|f(x-\mu) - f(x)\big|\,\mathrm{d}x \\
&\leq |\mu| + L|\mu| \int_{|x|<1} \big|\log|x|\big|\,\mathrm{d}x = |\mu| + 2L|\mu| = (2L+1)|\mu|.
\end{aligned}$$

So (i) is proved.

(ii): By definition, $\log^2(x) = g_1^2(x) + g_2^2(x)$. We have

$$\begin{aligned}
&\Big|\mathbb{E}\big[\log^2|\epsilon+\mu|\big] - \mathbb{E}\big[\log^2|\epsilon|\big]\Big| \\
\leq &\Big|\mathbb{E}\big[g_1^2(|\epsilon+\mu|) - g_1^2(|\epsilon|)\big]\Big| + \Big|\mathbb{E}\big[g_2^2(|\epsilon+\mu|) - g_2^2(|\epsilon|)\big]\Big| \\
= &\Big|\mathbb{E}\Big[\big(g_1(|\epsilon+\mu|) - g_1(|\epsilon|)\big)^2 + 2\big(g_1(|\epsilon+\mu|) - g_1(|\epsilon|)\big) \cdot g_1(|\epsilon|)\Big]\Big|
\end{aligned}$$

$$
\begin{aligned}
&+\left|\mathbb{E}\Big[g_2^2(|\epsilon+\mu|)-g_2^2(|\epsilon|)\Big]\right| \\
\leq&\mathbb{E}\Big[\big(|\epsilon+\mu|-|\epsilon|\big)^2\Big]+2\mathbb{E}\Big[\big||\epsilon+\mu|-|\epsilon|\big|\cdot g_1(|\epsilon|)\Big] \\
&+\Big|\int_{|x|<1}\log^2|x|\cdot\big(f(x-\mu)-f(x)\big)\,\mathrm{d}x\Big| \\
\leq&|\mu|^2+2|\mu|\mathbb{E}\Big[g_1\big(|\epsilon|\big)\Big]+L|\mu|\int_{|x|<1}(\log|x|)^2\,\mathrm{d}x \\
\leq&|\mu|^2+\Big(2\mathbb{E}\Big[\big|\log|\epsilon|\big|\Big]+4L\Big)|\mu|.
\end{aligned}
$$

So (ii) is proved.

(iii): By definition, $\log^4(x)=g_1^4(x)+g_2^4(x)$. We have

$$
\begin{aligned}
&\Big|\mathbb{E}\big[\log^4|\epsilon+\mu|\big]-\mathbb{E}\big[\log^4|\epsilon|\big]\Big| \\
\leq&\Big|\mathbb{E}\big[g_1^4(|\epsilon+\mu|)-g_1^4(|\epsilon|)\big]\Big|+\Big|\mathbb{E}\big[g_2^4(|\epsilon+\mu|)-g_2^4(|\epsilon|)\big]\Big| \\
=&\Big|\mathbb{E}\Big[\big(g_1(|\epsilon+\mu|)-g_1(|\epsilon|)\big)^4+4\big(g_1(|\epsilon+\mu|)-g_1(|\epsilon|)\big)^3\cdot g_1(|\epsilon|) \\
&+6\big(g_1(|\epsilon+\mu|)-g_1(|\epsilon|)\big)^2\cdot\big(g_1(|\epsilon|)\big)^2+4\big(g_1(|\epsilon+\mu|)-g_1(|\epsilon|)\big)\cdot\big(g_1(|\epsilon|)\big)^3\Big]\Big| \\
&+\Big|\mathbb{E}\Big[g_2^4(|\epsilon+\mu|)-g_2^4(|\epsilon|)\Big]\Big| \\
\leq&\mathbb{E}\Big[\big(|\epsilon+\mu|-|\epsilon|\big)^4\Big]+4\mathbb{E}\Big[\big||\epsilon+\mu|-|\epsilon|\big|^3\cdot g_1(|\epsilon|)\Big] \\
&+6\mathbb{E}\Big[\big||\epsilon+\mu|-|\epsilon|\big|^2\cdot\big(g_1(|\epsilon|)\big)^2\Big]+4\mathbb{E}\Big[\big||\epsilon+\mu|-|\epsilon|\big|\cdot\big(g_1(|\epsilon|)\big)^3\Big] \\
&+\Big|\int_{|x|<1}\log^4|x|\cdot\big(f(x-\mu)-f(x)\big)\,\mathrm{d}x\Big| \\
\leq&|\mu|^4+4|\mu|^3\mathbb{E}\Big[\big|\log|\epsilon|\big|\Big]+6|\mu|^2\mathbb{E}\big[\log^2|\epsilon|\big]+4|\mu|\mathbb{E}\Big[\big|\log|\epsilon|\big|^3\Big] \\
&+L|\mu|\int_{|x|<1}(\log|x|)^4\,\mathrm{d}x \\
=&|\mu|^4+4|\mu|^3\mathbb{E}\Big[\big|\log|\epsilon|\big|\Big]+6|\mu|^2\mathbb{E}\big[\log^2|\epsilon|\big]+\Big(4\mathbb{E}\Big[\big|\log|\epsilon|\big|^3\Big]+48L\Big)|\mu|.
\end{aligned}
$$

So (iii) is proved. □

A.3 Proofs for Proposition 5

In order to prove Proposition 5, notice that by symmetry, we only need to prove the result in (i) and the result in (iii) that corresponds to $\hat{\beta}(Z^{(1)})$. The results for $\hat{\beta}(Z^{(2)})$ follows similarly. To prove (i) for $\hat{\beta}(Z^{(1)})$, we will prove a more general case where we use the ℓ_1 penalized estimator as the initial estimator of LLA. The case of using $\mathbf{0}$ as initial value follows as a special case then. Recall $\ell_n(\beta) \coloneqq \frac{1}{n}\sum_{i=1}^{\frac{n}{2}}(y_i - \mathbf{x}_i^{\mathrm{T}}\beta)^2$. The ℓ_1 penalized estimator is defined as

$$\hat{\beta}^{\text{lasso}} \coloneqq \arg\min_{\beta\in\mathbb{R}^p} \ell_n(\beta) + \lambda_{\text{lasso}} \sum_{j=1}^{p} |\beta_j|,$$

where $\lambda_{\text{lasso}} > 0$ is some tuning paramater. Recall that with $\mathbf{0}$ as initial value, the first iteration of the LLA algorithm gives the ℓ_1 penalized estimator with $p'_\lambda(0)$ as tuning parameter. For SCAD and MCP, $p'_\lambda(0) = \lambda$. Hence, the estimator given by the LLA algorithm in $\mathcal{O}_1$ with $\mathbf{0}$ as initial value is the same as that given by the LLA algorithm with $\hat{\beta}^{\text{lasso}}$ as initial value, with the specific choice $\lambda_{\text{lasso}} = \lambda$.

So instead of proving (i), we prove the following result for general λ_{lasso}

(i') If we pick $\lambda \geq \frac{3s_1^{\frac{1}{2}}\lambda_{\text{lasso}}}{a_0\kappa}$ when $(\mathbf{C}_2)$ holds and pick $\lambda \geq \frac{3\lambda_{\text{lasso}}}{a_0\rho}$ when $(\mathbf{C}'_2)$ holds, then $\hat{\beta}(Z^{(1)}) = \hat{\beta}^{\text{ora1}}$ holds with probability at least $1 - 2p\exp\left(-\frac{n\lambda_{\text{lasso}}^2}{16M\sigma^2\Omega^2}\right) - 2(p-s_1)\exp\left(-\frac{a_1^2 n\lambda^2}{4\sigma^2\Omega^2 M}\right) - 2s_1\exp\left(-\frac{n\varphi(\|\beta^*_{\mathbb{A}_1}\|_{\min} - a\lambda)^2}{4\sigma^2\Omega^2}\right)$.

We can see that (i) follows from (i') after taking $\lambda_{\text{lasso}} = \lambda$.

With $\hat{\beta}^{\text{lasso}}$ being the initial estimator, we have the following Proposition 8 and Proposition 9.

Proposition 8 Let $\hat{\beta}^{\text{lasso}}$ be the lasso estimator on $Z^{(1)}$ with tuning parameter λ_{lasso}, i.e. the initial estimator for the LLA algorithm from which we get $\hat{\beta}(Z^{(1)})$. Under assumptions $(\mathbf{A}_0)$, $(\mathbf{C}_1)$, $(\mathbf{C}_2)$ or $(\mathbf{C}'_2)$, $(\mathbf{C}_3)$, we have $\hat{\beta}^{\text{lasso}}$ satisfies

$$\|\hat{\beta}^{\text{lasso}} - \beta^*\|_{\ell_2} \leq 3s_1^{\frac{1}{2}}\lambda_{\text{lasso}}\kappa^{-1}, \text{ if } (\mathbf{C}_2) \text{ holds;}$$
$$\|\hat{\beta}^{\text{lasso}} - \beta^*\|_{\infty} \leq 3\lambda_{\text{lasso}}\rho^{-1}, \text{ if } (\mathbf{C}'_2) \text{ holds,}$$

with probability at least $1 - 2p\exp\left(-\frac{n\lambda_{\text{lasso}}^2}{16M\sigma^2\Omega^2}\right)$. □

Proposition 9 Choose the tuning parameters so that $\|\beta^*_{\mathbb{A}_1}\|_{\min} > (a+1)(\lambda \vee \tilde{\lambda})$. Let $\hat{\beta}_{\text{lasso}}$ be the same estimator as in Proposition 8. Then, the LLA algorithm in $\mathcal{O}_1$ initialized by $\hat{\beta}^0 = \hat{\beta}^{\text{lasso}}$ converges to $\hat{\beta}^{\text{oracle}}$ after two iterations with probability at least $1 - p_1 - p_2 - p_3$, where

$$
\begin{aligned}
p_1 &= \mathrm{P}(\|\hat{\beta}^{\text{lasso}} - \beta^*\|_\infty > a_0\lambda), \\
p_2 &= \mathrm{P}(\|\nabla_{\mathbb{A}_1^c} \ell_n(\hat{\beta}^{\text{oracle}})\|_\infty \geq a_1\lambda), \\
p_3 &= \mathrm{P}(\min_{j \in \mathbb{A}_1} |\hat{\beta}^{\text{oracle}}_j| \leq a\lambda).
\end{aligned}
$$

□

We prove Proposition 8 first.

A.3.1 Proof of Proposition 8

Proof A.6

For notational convenience, we slightly abuse notation in this proof. We use $\mathbf{y}$ to denote the response vector in $Z^{(1)}$, i.e. $(y_1, \ldots, y_{\frac{n}{2}})$, use $\mathbf{X}$ to represent $\mathbf{X}^{(1)}$, and use $\mathbf{X}_j = (x_{1j}, \ldots, x_{\frac{n}{2}j})^{\mathrm{T}}$ to represent the jth column of $\mathbf{X}^{(1)}$. The ith row of $\mathbf{X}^{(1)}$ is still denoted as $\mathbf{x}_i^{\mathrm{T}}, i = 1, \ldots, \frac{n}{2}$. Write $\mathbf{y} = \mathbf{X}\beta^* + \eta$, where $\eta = (\eta_1, \ldots, \eta_{\frac{n}{2}})^{\mathrm{T}}$ with $\eta_i = \mathrm{e}^{\mathbf{x}_i^{\mathrm{T}}\gamma^*}\epsilon_i$. Define the event $\mathcal{E} = \{\lambda_{\text{lasso}} \geq \frac{2}{n/2}\|\mathbf{X}^{\mathrm{T}}\eta\|_\infty\} = \{\lambda_{\text{lasso}} \geq \frac{4}{n}\|\mathbf{X}^{\mathrm{T}}\eta\|_\infty\}$. Then, by Proposition 4 and $|\mathbb{A}_1| = s_1$, we know that the event $\mathcal{E}$ implies event $\{\|\hat{\beta}^{\text{lasso}} - \beta^*\|_{\ell_2} \leq \frac{3}{\kappa}\lambda_{\text{lasso}} s_1^{\frac{1}{2}}\}$ under assumption $(\mathbf{C}_2)$, and implies $\{\|\hat{\beta}^{\text{lasso}} - \beta^*\|_\infty \leq \frac{3}{\rho}\lambda_{\text{lasso}}\}$ under assumption $(\mathbf{C}_2')$. So under assumption $(\mathbf{C}_2)$ or assumption $(\mathbf{C}_2')$, we have

$$
\begin{aligned}
\mathrm{P}(\|\hat{\beta}^{\text{lasso}} - \beta^*\|_{\ell_2} \leq \tfrac{3}{\kappa}\lambda_{\text{lasso}} s_1^{\frac{1}{2}}) &\geq \mathrm{P}(\{\lambda_{\text{lasso}} \geq \tfrac{4}{n}\|\mathbf{X}^{\mathrm{T}}\eta\|_\infty\}) \\
&= 1 - \mathrm{P}(\tfrac{4}{n}\|\mathbf{X}^{\mathrm{T}}\eta\|_\infty > \lambda_{\text{lasso}}) \\
&\text{or} \\
\mathrm{P}(\|\hat{\beta}^{\text{lasso}} - \beta^*\|_\infty \leq \tfrac{3}{\rho}\lambda_{\text{lasso}}) &\geq \mathrm{P}(\{\lambda_{\text{lasso}} \geq \tfrac{4}{n}\|\mathbf{X}^{\mathrm{T}}\eta\|_\infty\}) \\
&= 1 - \mathrm{P}(\tfrac{4}{n}\|\mathbf{X}^{\mathrm{T}}\eta\|_\infty > \lambda_{\text{lasso}}),
\end{aligned}
\tag{A.3.1}
$$

respectively. It suffices to derive an upper bound for $\mathrm{P}(\frac{4}{n}\|\mathbf{X}^{\mathrm{T}}\eta\|_\infty > \lambda_{\text{lasso}})$. In fact,

$$
\mathrm{P}(\frac{4}{n}\|\mathbf{X}^{\mathrm{T}}\eta\|_\infty > \lambda_{\text{lasso}}) \leq \sum_{j=1}^{p} \mathrm{P}(|\mathbf{X}_j^{\mathrm{T}}\eta| > \frac{n\lambda_{\text{lasso}}}{4})
$$

$$= \sum_{j=1}^{p} \mathrm{P}(|\mathbf{X}_j^{\mathrm{T}}\mathbf{W}^*\epsilon| > \frac{n\lambda_{\mathrm{lasso}}}{4}), \tag{A.3.2}$$

where $\mathbf{W}^* = \mathbf{diag}\{\mathrm{e}^{\mathbf{x}_1^{\mathrm{T}}\gamma^*}, \ldots, \mathrm{e}^{\mathbf{x}_{n/2}^{\mathrm{T}}\gamma^*}\}$. Let $\mathbf{a}^{\mathrm{T}} \equiv (a_1, \ldots, a_{n/2}) = \mathbf{X}_j^{\mathrm{T}}\mathbf{W}^*$, then $\mathbf{a}^{\mathrm{T}}\mathbf{a} = \mathbf{X}_j^{\mathrm{T}}\mathbf{W}^{*2}\mathbf{X}_j$. By Proposition 2, we have

$$\begin{aligned}
&\mathrm{P}(|\mathbf{X}_j^{\mathrm{T}}\mathbf{W}^*\epsilon| > \frac{n\lambda_{\mathrm{lasso}}}{4}) \leq 2\exp\Big(-\frac{n^2\lambda_{\mathrm{lasso}}^2}{32\sigma^2\mathbf{X}_j^{\mathrm{T}}\mathbf{W}^{*2}\mathbf{X}_j}\Big) \\
\leq &2\exp\Big(-\frac{n^2\lambda_{\mathrm{lasso}}^2}{32\sigma^2\Omega^2\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j}\Big) \leq 2\exp\Big(-\frac{n\lambda_{\mathrm{lasso}}^2}{16\sigma^2\Omega^2 M}\Big),
\end{aligned} \tag{A.3.3}$$

where we use $\mathrm{e}^{\mathbf{x}_i^{\mathrm{T}}\gamma^*} \leq \Omega, \forall i = 1, \ldots, n$ in the second inequality and $\frac{\|\mathbf{X}_j\|_{\ell_2}^2}{n/2} \leq M$ in the third inequality. Collecting the results in ((A.3.1)), ((A.3.2)) and ((A.3.3)), the proof of the proposition is completed. □

We next prove Proposition 9.

A.3.2 Proof of Proposition 9

Proof A.7

We have $\nabla^2\ell_n(\beta) = \frac{1}{n/2}\sum_{i=1}^{n/2}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$. Since $u^{\mathrm{T}}(\frac{1}{n/2}\sum_{i=1}^{n/2}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})u = \frac{1}{n/2}\sum_{i=1}^{n/2}(\mathbf{x}_i^{\mathrm{T}}u)^2 \geq 0, \forall u \in \mathbb{R}^p$, $\nabla^2\ell_n(\beta)$ is nonnegative definite. So ℓ_n is a convex function. Similarly, let $\bar{\ell}_n(\beta_{\mathbb{A}_1}) \equiv \frac{1}{n}\sum_{i=1}^{n/2}(y_i - \mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}\beta_{\mathbb{A}_1})^2 = \ell_n(\beta)$ for any β such that $\beta_{\mathbb{A}_1^{\mathrm{c}}} = 0$. Then, we have $\nabla\bar{\ell}_n(\beta_{\mathbb{A}_1}) = -\frac{1}{n/2}\sum_{i=1}^{n/2}(y_i - \mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}\beta_{\mathbb{A}_1})\mathbf{x}_{i\mathbb{A}_1}$, and also $\nabla^2\bar{\ell}_n(\beta_{\mathbb{A}_1}) = \frac{1}{n/2}\sum_{i=1}^{n/2}\mathbf{x}_{i\mathbb{A}_1}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}} = \frac{1}{n/2}\mathbf{X}_{\mathbb{A}_1}^{(1)^{\mathrm{T}}}\mathbf{X}_{\mathbb{A}_1}^{(1)}$. By assumption $(\mathbf{C}_3)$ we know $\mathbf{X}_{\mathbb{A}_1}^{(1)}$ is of full rank, so $\frac{1}{n/2}\mathbf{X}_{\mathbb{A}_1}^{(1)^{\mathrm{T}}}\mathbf{X}_{\mathbb{A}_1}^{(1)}$ is positive definite and $\bar{\ell}_n$ is strictly convex. Therefore $\hat{\beta}^{\mathrm{ora1}}$ is the unique solution to $\nabla\bar{\ell}_n(\beta_{\mathbb{A}_1}) = 0$ and $\beta_{\mathbb{A}_1^{\mathrm{c}}} = 0$. The proposition then follows from Theorem 1 and Theorem 2 in Fan et al. (2014b). □

Now we are ready to prove Proposition 5.

A.3.3 Proof of Proposition 5

Proof A.8

By symmetry we only need to prove the result in (i) and the result in (iii) for $\hat{\beta}(Z^{(1)})$. So we may abuse the notation in the same way as we did in the proof of Proposition

8. That is, we use $\mathbf{y}$ to denote the response vector in $Z^{(1)}$, i.e. $(y_1, \ldots, y_{\frac{n}{2}})$, use $\mathbf{X}$ to represent $\mathbf{X}^{(1)}$, and use $\mathbf{X}_j = (x_{1j}, \ldots, x_{\frac{n}{2}j})^{\mathrm{T}}$ to represent the jth column of $\mathbf{X}^{(1)}$. The ith row of $\mathbf{X}^{(1)}$ is still denoted as $\mathbf{x}_i^{\mathrm{T}}, i = 1, \ldots, \frac{n}{2}$.

(i): It suffices to prove (i'). Recall Proposition 9 shows that the LLA algorithm in $\mathcal{O}_1$ converges to $\hat{\beta}^{\mathrm{ora}}$ after two interations with probability at least $1-p_1-p_2-p_3$. Let $\hat{\beta}_{\mathrm{lasso}}$ be the one appeared in Proposition 8. From Proposition 8 we already have

$$\mathrm{P}(\|\hat{\beta}^{\mathrm{lasso}} - \beta^*\|_{\ell_2} > 3s_1^{\frac{1}{2}}\lambda_{\mathrm{lasso}}\kappa^{-1}) \leq 2p\exp\left(-\frac{n\lambda_{\mathrm{lasso}}^2}{16M\sigma^2\Omega^2}\right), \text{ if } (\mathbf{C}_2) \text{ holds;}$$
$$\mathrm{P}(\|\hat{\beta}^{\mathrm{lasso}} - \beta^*\|_{\infty} > 3\lambda_{\mathrm{lasso}}\rho^{-1}) \leq 2p\exp\left(-\frac{n\lambda_{\mathrm{lasso}}^2}{16M\sigma^2\Omega^2}\right), \text{ if } (\mathbf{C}_2') \text{ holds.}$$

So under $(\mathbf{C}_2)$, since we pick λ such that $a_0\lambda \geq 3s_1^{\frac{1}{2}}\lambda_{\mathrm{lasso}}\kappa^{-1}$, and $\|\hat{\beta}^{\mathrm{lasso}} - \beta^*\|_{\infty} \leq \|\hat{\beta}^{\mathrm{lasso}} - \beta^*\|_{\ell_2}$, we have

$$p_1 = \mathrm{P}(\|\hat{\beta}^{\mathrm{lasso}} - \beta^*\|_{\infty} > a_0\lambda) \leq \mathrm{P}(\|\hat{\beta}^{\mathrm{lasso}} - \beta^*\|_{\ell_2} > 3s_1^{\frac{1}{2}}\lambda_{\mathrm{lasso}}\kappa^{-1})$$
$$\leq 2p\exp\left(-\frac{n\lambda_{\mathrm{lasso}}^2}{16M\sigma^2\Omega^2}\right).$$

Similarly, under $(\mathbf{C}_2')$, since we pick λ such that $a_0\lambda \geq 3\lambda_{\mathrm{lasso}}\rho^{-1}$, we have

$$p_1 = \mathrm{P}(\|\hat{\beta}^{\mathrm{lasso}} - \beta^*\|_{\infty} > a_0\lambda) \leq 2p\exp\left(-\frac{n\lambda_{\mathrm{lasso}}^2}{16M\sigma^2\Omega^2}\right).$$

It suffices to bound p_2 and p_3 in Proposition 9.

We first look at $p_2 = \mathrm{P}(\|\nabla_{\mathbb{A}_1^{\mathrm{c}}}\ell_n(\hat{\beta}^{\mathrm{ora1}})\|_{\infty} \geq a_1\lambda)$. Recall $\hat{\beta}^{\mathrm{ora1}}_{\mathbb{A}_1} = (\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-1}\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{y}$ and $\hat{\beta}^{\mathrm{ora1}}_{\mathbb{A}_1^{\mathrm{c}}} = \mathbf{0}$. So we have $\ell_n(\hat{\beta}^{\mathrm{ora1}}) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\hat{\beta}^{\mathrm{ora1}}\|_{\ell_2}^2 = \frac{1}{n}\|\mathbf{y} - \mathbf{X}_{\mathbb{A}_1}\hat{\beta}^{\mathrm{ora1}}_{\mathbb{A}_1}\|_{\ell_2}^2$, and we also have that $\nabla_{\mathbb{A}_1^{\mathrm{c}}}\ell_n(\hat{\beta}^{\mathrm{ora1}}) = -\frac{1}{n/2}\mathbf{X}_{\mathbb{A}_1^{\mathrm{c}}}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}_{\mathbb{A}_1}\hat{\beta}^{\mathrm{ora1}}_{\mathbb{A}_1}) = -\frac{1}{n/2}\mathbf{X}_{\mathbb{A}_1^{\mathrm{c}}}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}_{\mathbb{A}_1}(\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-1}\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{y})$. Plugging in $\mathbf{y} = \mathbf{X}\beta^* + \eta$, where we denote $\eta = (\eta_1, \ldots, \eta_{\frac{n}{2}})^{\mathrm{T}}$ with $\eta_i = \mathrm{e}^{\mathbf{x}_i^{\mathrm{T}}\gamma^*}\epsilon_i$, we have $\nabla_{\mathbb{A}_1^{\mathrm{c}}}\ell_n(\hat{\beta}^{\mathrm{ora1}}) = -\frac{1}{n/2}\mathbf{X}_{\mathbb{A}_1^{\mathrm{c}}}^{\mathrm{T}}(\mathbf{l} - \mathbf{X}_{\mathbb{A}_1}(\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-1}\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}})\eta$, where $\mathbf{l} \in \mathbb{R}^{\frac{n}{2}\times\frac{n}{2}}$ is the identity matrix. Denote $\mathbf{H}_{\mathbb{A}_1} \coloneqq \mathbf{X}_{\mathbb{A}_1}(\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-1}\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}$, and denote $\mathbf{W}^* = \mathbf{diag}\{\mathrm{e}^{\mathbf{x}_1^{\mathrm{T}}\gamma^*}, \ldots, \mathrm{e}^{\mathbf{x}_{n/2}^{\mathrm{T}}\gamma^*}\}$ so that $\eta = \mathbf{W}^*\epsilon$. By Proposition 2 and assumption $(\mathbf{C}_1)$ we have

$$p_2 = \mathrm{P}(\| -\frac{1}{n/2}\mathbf{X}_{\mathbb{A}_1^{\mathrm{c}}}^{\mathrm{T}}(\mathbf{l} - \mathbf{H}_{\mathbb{A}_1})\eta\|_{\max} \geq a_1\lambda)$$

$$
\begin{aligned}
&\le \sum_{j\in\mathbb{A}_1^{\mathrm{c}}} \mathrm{P}(|\mathbf{X}_j^{\mathrm{T}}(\mathbf{l}-\mathbf{H}_{\mathbb{A}_1})\mathbf{W}^*\epsilon| \ge a_1\frac{n}{2}\lambda)\\
&\le 2\sum_{j\in\mathbb{A}_1^{\mathrm{c}}} \exp\Big(-\frac{a_1^2n^2\lambda^2}{8\sigma^2\mathbf{X}_j^{\mathrm{T}}(\mathbf{l}-\mathbf{H}_{\mathbb{A}_1})\mathbf{W}^{*2}(\mathbf{l}-\mathbf{H}_{\mathbb{A}_1})\mathbf{X}_j}\Big)\\
&\le 2\sum_{j\in\mathbb{A}_1^{\mathrm{c}}} \exp\Big(-\frac{a_1^2n^2\lambda^2}{8\sigma^2\Omega^2\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j}\Big)\\
&\le 2\sum_{j\in\mathbb{A}_1^{\mathrm{c}}} \exp\Big(-\frac{a_1^2n^2\lambda^2}{4\sigma^2\Omega^2nM}\Big)\\
&= 2(p-s_1)\exp\Big(-\frac{a_1^2n\lambda^2}{4\sigma^2\Omega^2M}\Big).
\end{aligned}
$$

Next we look at $p_3 = \mathrm{P}(\min_{j\in\mathbb{A}_1}|\hat{\beta}_j^{\mathrm{ora1}}| \le a\lambda)$. By the choice of tuning parameters, we have $\{\min_{j\in\mathbb{A}_1}|\hat{\beta}_j^{\mathrm{ora1}}| \le a\lambda\} \subset \{\|\hat{\beta}_{\mathbb{A}_1}^{\mathrm{ora1}} - \beta_{\mathbb{A}_1}^*\|_{\max} \ge \|\beta_{\mathbb{A}_1}^*\|_{\min} - a\lambda\}$. Denote $(\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-1}\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}} = (\mathbf{u}_1,\ldots,\mathbf{u}_{s_1})^{\mathrm{T}}$, where $\mathbf{u}_i \in \mathbb{R}^n$. Then $\mathbf{u}_j = \mathbf{X}_{\mathbb{A}_1}(\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-1}\mathbf{e}_j$, where $\mathbf{e}_j$ is the unit vector with jth element 1 and other elements 0. Then we have $\mathbf{u}_j^{\mathrm{T}}\mathbf{u}_j = \mathbf{e}_j^{\mathrm{T}}(\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-1}\mathbf{e}_j \le \frac{1}{\frac{n}{2}\varphi}$. Therefore by Proposition 2,

$$
\begin{aligned}
p_3 &\le \mathrm{P}(\|\hat{\beta}_{\mathbb{A}_1}^{\mathrm{ora1}} - \beta_{\mathbb{A}_1}^*\|_{\max} \ge \|\beta_{\mathbb{A}_1}^*\|_{\min} - a\lambda)\\
&= \mathrm{P}(\|(\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-1}\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\eta\|_{\max} \ge \|\beta_{\mathbb{A}_1}^*\|_{\min} - a\lambda)\\
&\le \sum_{j=1}^{s_1} \mathrm{P}(|\mathbf{u}_j^{\mathrm{T}}\mathbf{W}^*\epsilon| \ge \|\beta_{\mathbb{A}_1}^*\|_{\min} - a\lambda)\\
&\le 2\sum_{j=1}^{s_1} \exp\big(-\frac{(\|\beta_{\mathbb{A}_1}^*\|_{\min} - a\lambda)^2}{2\sigma^2\mathbf{u}_j^{\mathrm{T}}\mathbf{W}^{*2}\mathbf{u}_j}\big)\\
&\le 2s_1\exp\big(-\frac{n\varphi(\|\beta_{\mathbb{A}_1}^*\|_{\min} - a\lambda)^2}{4\sigma^2\Omega^2}\big).
\end{aligned}
$$

Thus (i') is proved and (i) follows.

(ii): Similar to (i).

(iii): From (i) and the choice of tuning parameters we already have

$$
\mathrm{P}(\hat{\beta}(Z^{(1)}) \ne \hat{\beta}^{\mathrm{ora1}}) \le 2p\mathrm{e}^{-c_1n\lambda^2} + 2p\mathrm{e}^{-c_2n\lambda^2} \tag{A.3.4}
$$

for some $c_1, c_2 > 0$ (for instance, $c_1 = \frac{1}{16M\sigma^2\Omega^2}$, $c_2 = \frac{a_1^2}{4\sigma^2\Omega^2M} \wedge \frac{\varphi}{4\sigma^2\Omega^2}$). For $T > 0$, let

$\lambda = \sqrt{T}\sqrt{\frac{\log p}{n}}$. Then we have

$$\mathrm{P}(\hat{\beta}(Z^{(1)}) \neq \hat{\beta}^{\mathrm{ora1}}) \leq 2p^{-(c_1T-1)} + 2p^{-(c_2T-1)}. \tag{A.3.5}$$

Take $T > \frac{1}{c_1} \vee \frac{1}{c_2}$, we have $\lim_{n\to\infty} \mathrm{P}(\hat{\beta}(Z^{(1)}) \neq \hat{\beta}^{\mathrm{ora1}}) = 0$.

It suffices to show $\|\hat{\beta}^{\mathrm{ora1}} - \beta^*\|_{\ell_2} = O_p(\sqrt{\frac{s_1}{n}})$ which implies $\|\hat{\beta}(Z^{(1)}) - \beta^*\|_{\ell_2} = O_p(\sqrt{\frac{s_1}{n}})$. In fact,

$$\begin{aligned}
&\mathbb{E}[\|\hat{\beta}^{\mathrm{ora1}} - \beta^*\|_{\ell_2}^2] = \mathbb{E}[\|\hat{\beta}^{\mathrm{ora1}}_{\mathbb{A}_1} - \beta^*_{\mathbb{A}_1}\|_{\ell_2}^2] \\
&= \mathbb{E}[\epsilon^{\mathrm{T}}\mathbf{W}^*\mathbf{X}_{\mathbb{A}_1}(\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-2}\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{W}^*\epsilon] \\
&= \mathrm{var}(\epsilon_1)\mathrm{tr}(\mathbf{W}^*\mathbf{X}_{\mathbb{A}_1}(\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-2}\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{W}^*) \\
&\leq \mathrm{var}(\epsilon_1)\Omega^2\mathrm{tr}(\mathbf{X}_{\mathbb{A}_1}(\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-2}\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}) \\
&= \mathrm{var}(\epsilon_1)\Omega^2\mathrm{tr}((\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-1}) \\
&\leq \mathrm{var}(\epsilon_1)\Omega^2\frac{2s_1}{n\varphi}.
\end{aligned}$$

This implies $\|\hat{\beta}^{\mathrm{ora1}} - \beta^*\|_{\ell_2} = O_p(\sqrt{\frac{s_1}{n}})$. So the proof is completed. □

A.4 Proof of Lemma 3

Proof A.9 (Proof of Lemma 4)

For notational convenience, we slightly abuse the notation in this proof. Let $\mathbf{y}$ and $\mathbf{X}$ be the response and design matrix in $Z^{(1)}$, i.e. $\mathbf{y} = (y_1, \ldots, y_{n/2})^{\mathrm{T}}$ and $\mathbf{X} = \mathbf{X}^{(1)}$. Let $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$ be the response and the design matrix in $Z^{(2)}$, i.e. $\tilde{\mathbf{y}} = (y_{\frac{n}{2}+1}, \ldots, y_n)^{\mathrm{T}}$, and $\tilde{\mathbf{X}} = \mathbf{X}^{(2)}$. Then recall that $\hat{\beta}^{\mathrm{ora1}}_{\mathbb{A}_1} = (\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_1})^{-1}\mathbf{X}_{\mathbb{A}_1}^{\mathrm{T}}\mathbf{y}$, and $\hat{\beta}^{\mathrm{ora2}}_{\mathbb{A}_1} = (\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{y}}$. Recall that $v_i = \mathrm{e}^{\mathbf{x}_i^{\mathrm{T}}\gamma^*}, i = 1, \ldots, n$. Define $\tilde{v}_i \coloneqq v_{i+n/2}, i = 1, \ldots, \frac{n}{2}$. Let $\mathbf{W}^* = \mathbf{diag}\{v_1, \ldots, v_{n/2}\}$, and denote $\tilde{\mathbf{W}}^* = \mathbf{diag}\{\tilde{v}_1, \ldots, \tilde{v}_{n/2}\}$. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_{n/2})^{\mathrm{T}}$ and $\tilde{\epsilon} \coloneqq (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_{n/2})^{\mathrm{T}} \coloneqq (\epsilon_{n/2+1}, \ldots, \epsilon_n)^{\mathrm{T}}$. By definition, we have $\zeta_i = \frac{1}{v_i}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{W}}^*\tilde{\epsilon}, i = 1, \ldots, \frac{n}{2}$.

Since $v_i = \mathrm{e}^{\mathbf{x}_i^{\mathrm{T}}\gamma^*}$, we have $y_i = \mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}\beta^*_{\mathbb{A}_1} + v_i\epsilon_i, i = 1, \ldots, \frac{n}{2}$. This can also be written as $\mathbf{y} = \mathbf{X}_{\mathbb{A}_1}\beta^*_{\mathbb{A}_1} + \mathbf{W}^*\epsilon$. Similarly we can write $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}_{\mathbb{A}_1}\beta^*_{\mathbb{A}_1} + \tilde{\mathbf{W}}^*\tilde{\epsilon}$. Then we have

$$\mathrm{P}\Big(\frac{1}{n/2}\Big|\mathbf{a}^{\mathrm{T}}\log|\mathbf{y} - \mathbf{X}_{\mathbb{A}_1}\hat{\beta}^{\mathrm{ora2}}_{\mathbb{A}_1}| - \mathbf{a}^{\mathrm{T}}\mathbf{X}\gamma^*\Big| > C\Big)$$

$$
\begin{aligned}
&= \mathrm{P}\Big(\frac{1}{n/2}\Big|\sum_{i=1}^{n/2}\big(a_i\log|y_i-\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}\hat{\beta}_{\mathbb{A}_1}^{\mathrm{ora2}}|-a_i\mathbf{x}_i^{\mathrm{T}}\gamma^*\big)\Big|>C\Big)\\
&= \mathrm{P}\Big(\frac{1}{n/2}\Big|\sum_{i=1}^{n/2}\big(a_i\log|v_i\epsilon_i-\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}(\hat{\beta}_{\mathbb{A}_1}^{\mathrm{ora2}}-\beta_{\mathbb{A}_1}^*)|-a_i\mathbf{x}_i^{\mathrm{T}}\gamma^*\big)\Big|>C\Big)\\
&= \mathrm{P}\Big(\frac{1}{n/2}\Big|\sum_{i=1}^{n/2}\big(a_i\log|\epsilon_i-\frac{1}{v_i}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}(\hat{\beta}_{\mathbb{A}_1}^{\mathrm{ora2}}-\beta_{\mathbb{A}_1}^*)|\big)\Big|>C\Big)\\
&= \mathrm{P}\Big(\frac{1}{n/2}\Big|\sum_{i=1}^{n/2}\big(a_i\log|\epsilon_i-\frac{1}{v_i}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{W}}^*\tilde{\epsilon}|\big)\Big|>C\Big). \qquad \text{(A.4.1)}
\end{aligned}
$$

Recall $\zeta_i=\frac{1}{v_i}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{W}}^*\tilde{\epsilon}$. Let $K>0$ be any fixed positive constant, and let $\tilde{C}=K\wedge\frac{C}{(4L+2)G}$. Then, by union bound, we have

$$
\begin{aligned}
&\mathrm{P}\Big(\frac{1}{n/2}\Big|\sum_{i=1}^{n/2}\big(a_i\log|\epsilon_i-\zeta_i|\big)\Big|>C\Big)\\
&\le \mathrm{P}\Big(\frac{1}{n/2}\Big|\sum_{i=1}^{n/2}\big(a_i\log|\epsilon_i-\zeta_i|\big)\Big|>C,\max_{1\le i\le\frac{n}{2}}|\zeta_i|\le\tilde{C}\Big)+\mathrm{P}\Big(\max_{1\le i\le\frac{n}{2}}|\zeta_i|>\tilde{C}\Big).
\end{aligned}
\qquad \text{(A.4.2)}
$$

For any $t>0$, let $\mathcal{T}_t=\{\max_{1\le i\le\frac{n}{2}}|\zeta_i|\le t\}$. By union bound, Proposition 2, assumptions $(\mathbf{C}_1)$ and $(\mathbf{C}_3)$, we have

$$
\begin{aligned}
&\mathrm{P}(\mathcal{T}_t^{\mathrm{c}})\le\sum_{i=1}^{n/2}\mathrm{P}(|\zeta_i|>t)\\
&\le\sum_{i=1}^{n/2}2\exp\big(-\frac{t^2}{2\sigma^2\frac{1}{v_i^2}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{W}}^{*2}\tilde{\mathbf{X}}_{\mathbb{A}_1}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\mathbf{x}_{i\mathbb{A}_1}}\big)\\
&\le\sum_{i=1}^{n/2}2\exp\big(-\frac{t^2}{2\sigma^2\frac{\Omega^2}{\Psi^2}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\mathbf{x}_{i\mathbb{A}_1}}\big)\\
&\le\sum_{i=1}^{n/2}2\exp\big(-\frac{t^2}{2\sigma^2\frac{\Omega^2}{\Psi^2}\frac{2}{n\varphi}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}\mathbf{x}_{i\mathbb{A}_1}}\big)
\end{aligned}
$$

$$\leq \sum_{i=1}^{n/2} 2\exp\left(-\frac{t^2}{2\sigma^2\frac{\Omega^2}{\Psi^2}\frac{2}{n\varphi}s_1 M}\right)$$
$$= n\exp\left(-\frac{t^2\Psi^2\varphi}{4\sigma^2\Omega^2 s_1 M}n\right). \tag{A.4.3}$$

To bound the first term on the right hand side of ((A.4.2)), notice that

$$\begin{aligned}
&\mathrm{P}\Big(\frac{1}{n/2}\Big|\sum_{i=1}^{n/2}(a_i\log|\epsilon_i-\zeta_i|)\Big|>C, \max_{1\leq i\leq\frac{n}{2}}|\zeta_i|\leq\tilde{C}\Big)\\
=&\mathbb{E}\Big[\mathrm{P}\Big(\frac{1}{n/2}\Big|\sum_{i=1}^{n/2}(a_i\log|\epsilon_i-\zeta_i|)\Big|>C, \max_{1\leq i\leq\frac{n}{2}}|\zeta_i|\leq\tilde{C}\Big|\tilde{\epsilon}\Big)\Big]\\
\leq&\mathbb{E}\Big[1_{\{\max_{1\leq i\leq\frac{n}{2}}|\zeta_i|\leq\tilde{C}\}}\cdot\Big\{\mathrm{P}\Big(\frac{1}{n/2}\Big|\sum_{i=1}^{n/2}(a_i\mathbb{E}[\log|\epsilon_i-\zeta_i||\tilde{\epsilon}])\Big|>\frac{C}{2}\Big|\tilde{\epsilon}\Big)\\
&+\mathrm{P}\Big(\frac{1}{n/2}\Big|\sum_{i=1}^{n/2}(a_i\log|\epsilon_i-\zeta_i|-a_i\mathbb{E}[\log|\epsilon_i-\zeta_i||\tilde{\epsilon}])\Big|>\frac{C}{2}\Big|\tilde{\epsilon}\Big)\Big\}\Big].
\end{aligned}\tag{A.4.4}$$

By Lemma 2, conditioning on $\tilde{\epsilon}$ where $\max_{1\leq i\leq\frac{n}{2}}|\zeta_i|\leq\tilde{C}$ holds, $\log|\epsilon_i-\zeta_i|-\mathbb{E}[\log|\epsilon_i-\zeta_i|], i=1,\ldots,\frac{n}{2}$ are independent sub-exponential(η_0) random variables, where η_0 is a fixed positive constant defined as $\eta_0=2\sqrt{(4\sigma^2\mathrm{e}^{(4L+2)K})\vee(2K\mathrm{e}^{(2L+1)K})\vee(2+2C_0\mathrm{e}^{(2L+1)K})}$. Therefore, by Proposition 3, we have

$$\begin{aligned}
&\mathbb{E}\Big[1_{\{\max_{1\leq i\leq\frac{n}{2}}|\zeta_i|\leq\tilde{C}\}}\mathrm{P}\Big(\frac{1}{n/2}\Big|\sum_{i=1}^{n/2}(a_i\log|\epsilon_i-\zeta_i|-a_i\mathbb{E}[\log|\epsilon_i-\zeta_i||\tilde{\epsilon}])\Big|>\frac{C}{2}\Big|\tilde{\epsilon}\Big)\Big]\\
&\leq 2\exp\Big[-\Big(\frac{(C/2)^2n^2}{8\eta_0^2\sum_{i=1}^{n/2}a_i^2}\wedge\frac{C/2\,n}{4\eta_0\max_{1\leq i\leq\frac{n}{2}}|a_i|}\Big)\Big]\\
&\leq 2\exp\Big[-\Big(\frac{(C/2)^2}{4\eta_0^2G^2}\wedge\frac{C/2}{4\eta_0G}\Big)n\Big]\\
&= 2\exp\Big[-\Big(\frac{C^2}{16\eta_0^2G^2}\wedge\frac{C}{8\eta_0G}\Big)n\Big].
\end{aligned}\tag{A.4.5}$$

We are left to bound the remaining term in ((A.4.4)). In fact, by Lemma 3 (i) and

$\mathbb{E}[\log|\epsilon_i|]=0$, we have that $|\mathbb{E}[\log|\epsilon_i-\zeta_i|\big|\tilde{\epsilon}]|\leq(2L+1)|\zeta_i|$. Therefore,

$$\begin{aligned}
&\mathbb{E}\Big[1_{\{\max_{1\leq i\leq\frac{n}{2}}|\zeta_i|\leq\tilde{C}\}}\mathrm{P}\Big(\frac{1}{n/2}|\sum_{i=1}^{n/2}(a_i\mathbb{E}[\log|\epsilon_i-\zeta_i|\big|\tilde{\epsilon}])|>\frac{C}{2}\Big|\tilde{\epsilon}\Big)\Big]\\
&\leq\mathbb{E}\Big[1_{\{\max_{1\leq i\leq\frac{n}{2}}|\zeta_i|\leq\tilde{C}\}}\mathrm{P}\Big(G(2L+1)\max_{1\leq i\leq\frac{n}{2}}|\zeta_i|>\frac{C}{2}\Big|\tilde{\epsilon}\Big)\Big]\\
&=0,
\end{aligned}\tag{A.4.6}$$

where the last equality holds true since we have chosen $\tilde{C}$ such that $\tilde{C}\leq\frac{C}{(4L+2)G}$. Collecting the results in ((A.4.1)), ((A.4.2)), ((A.4.3)), ((A.4.4)), ((A.4.5)), ((A.4.6)), the proof of lemma 4 is completed. □

A.5 Proof of Proposition 6

Proof A.10 (Proof of Proposition 6)
For notational convenience we slightly abuse the notation in this proof. That is, we let $\mathbf{y}$ and $\mathbf{X}$ be the response and design matrix in $Z^{(1)}$, i.e. $\mathbf{y}=(y_1,\ldots,y_{n/2})^{\mathrm{T}}$ and $\mathbf{X}=\mathbf{X}^{(1)}$. And we let $\mathbf{X}_j=(x_{1j},\ldots,x_{\frac{n}{2}j})^{\mathrm{T}}$ to represent the jth column of $\mathbf{X}^{(1)}$. Then we can write $\mathbf{y}=\mathbf{X}\beta^*+\eta$, where $\eta=(\eta_1,\ldots,\eta_{\frac{n}{2}})^{\mathrm{T}}$ with $\eta_i=\mathrm{e}^{\mathbf{x}_i^{\mathrm{T}}\gamma^*}\epsilon_i, i=1,\ldots,\frac{n}{2}$. Let $z_i=\log|y_i-\mathbf{x}_i^{\mathrm{T}}\hat{\beta}^{\mathrm{ora2}}|$, $\mathbf{z}=(z_1,\ldots,z_{\frac{n}{2}})^{\mathrm{T}}$. Let $\eta'=\mathbf{z}-\mathbf{X}\gamma^*$. Then, by the definition of $\tilde{\gamma}_\bullet^{\mathrm{lasso}}$,

$$\tilde{\gamma}_\bullet^{\mathrm{lasso}}=\arg\min_{\gamma\in\mathbb{R}^p}\frac{1}{n}\|\mathbf{z}-\mathbf{X}\gamma\|_2^2+\tilde{\lambda}'_{\mathrm{lasso}}\|\gamma\|_1.$$

Define event $\mathcal{E}=\{\tilde{\lambda}'_{\mathrm{lasso}}\geq\frac{4}{n}\|\mathbf{X}^{\mathrm{T}}\eta'\|_\infty\}$. Then, by Proposition 4 and $|\mathbb{A}_2|=s_2$, we know that under assumption $(\mathbf{C}_2)$, the event $\mathcal{E}$ implies event $\{\|\tilde{\gamma}_\bullet^{\mathrm{lasso}}-\gamma^*\|_{\ell_2}\leq\frac{3}{\kappa'}\tilde{\lambda}'_{\mathrm{lasso}}s_2^{\frac{1}{2}}\}$, and we have

$$\begin{aligned}
&\mathrm{P}(\|\tilde{\gamma}_\bullet^{\mathrm{lasso}}-\gamma^*\|_{\ell_2}>\frac{3}{\kappa'}\tilde{\lambda}'_{\mathrm{lasso}}s_2^{\frac{1}{2}},\mathcal{T}_{K\wedge\frac{C}{(4L+2)G}})\\
&\leq\mathrm{P}(\frac{4}{n}\|\mathbf{X}^{\mathrm{T}}\eta'\|_\infty>\tilde{\lambda}'_{\mathrm{lasso}},\mathcal{T}_{K\wedge\frac{C}{(4L+2)G}}).
\end{aligned}\tag{A.5.1}$$

Similarly, under $(\mathbf{C}_2')$, the event $\mathcal{E}$ implies event $\{\|\tilde{\gamma}_\bullet^{\text{lasso}} - \gamma^*\|_\infty \leq \frac{3}{\rho'}\tilde{\lambda}'_{\text{lasso}}\}$, and we have

$$\begin{aligned}
&\mathrm{P}(\|\tilde{\gamma}_\bullet^{\text{lasso}} - \gamma^*\|_{\ell_2} > \frac{3}{\rho'}\tilde{\lambda}'_{\text{lasso}}, \mathcal{T}_{K\wedge\frac{C}{(4L+2)G}}) \\
&\leq \mathrm{P}(\frac{4}{n}\|\mathbf{X}^{\mathrm{T}}\eta'\|_\infty > \tilde{\lambda}'_{\text{lasso}}, \mathcal{T}_{K\wedge\frac{C}{(4L+2)G}}).
\end{aligned} \tag{A.5.2}$$

Here, C and G are positive number to be determined later.

It suffices to derive an upper bound for $\mathrm{P}(\frac{4}{n}\|\mathbf{X}^{\mathrm{T}}\eta'\|_\infty > \tilde{\lambda}'_{\text{lasso}}, \mathcal{T}_{K\wedge\frac{C}{(4L+2)G}})$. We have

$$\begin{aligned}
&\mathrm{P}(\frac{4}{n}\|\mathbf{X}^{\mathrm{T}}\eta'\|_\infty > \tilde{\lambda}'_{\text{lasso}}, \mathcal{T}_{K\wedge\frac{C}{(4L+2)G}}) \\
&= \mathrm{P}(\frac{4}{n}\|\mathbf{X}^{\mathrm{T}}(\mathbf{z} - \mathbf{X}\gamma^*)\|_\infty > \tilde{\lambda}'_{\text{lasso}}, \mathcal{T}_{K\wedge\frac{C}{(4L+2)G}}) \\
&\leq \sum_{j=1}^{p} \mathrm{P}(\frac{4}{n}|\mathbf{X}_j^{\mathrm{T}}(\mathbf{z} - \mathbf{X}\gamma^*)| > \tilde{\lambda}'_{\text{lasso}}, \mathcal{T}_{K\wedge\frac{C}{(4L+2)G}})
\end{aligned} \tag{A.5.3}$$

Next, for any $j \in \{1, \ldots, p\}$, we give an upper bound for $\mathrm{P}(\frac{4}{n}|\mathbf{X}_j^{\mathrm{T}}(\mathbf{z} - \mathbf{X}\gamma^*)| > \tilde{\lambda}'_{\text{lasso}}, \mathcal{T}_{K\wedge\frac{C}{(4L+2)G}})$. We have

$$\begin{aligned}
&\mathrm{P}(\frac{4}{n}|\mathbf{X}_j^{\mathrm{T}}(\mathbf{z} - \mathbf{X}\gamma^*)| > \tilde{\lambda}'_{\text{lasso}}, \mathcal{T}_{K\wedge\frac{C}{(4L+2)G}}) \\
&= \mathrm{P}(\frac{2}{n}\left|\mathbf{X}_j^{\mathrm{T}}\mathbf{z} - \mathbf{X}_j^{\mathrm{T}}\mathbf{X}\gamma^*\right| > \frac{\tilde{\lambda}'_{\text{lasso}}}{2}, \mathcal{T}_{K\wedge\frac{C}{(4L+2)G}}).
\end{aligned} \tag{A.5.4}$$

Since $\mathbf{X}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^{\mathrm{T}}$ and $|x_{ij}| \leq \sqrt{M}, \forall i = 1, \ldots, n$, plugging $C = C_1 := \frac{\tilde{\lambda}'_{\text{lasso}}}{2}$ and $G = G_1 := \sqrt{M}$ into Lemma 4, we have

$$\mathrm{P}(\frac{2}{n}\left|\mathbf{X}_j^{\mathrm{T}}\mathbf{z} - \mathbf{X}_j^{\mathrm{T}}\mathbf{X}\gamma^*\right| > \frac{\tilde{\lambda}'_{\text{lasso}}}{2}, \mathcal{T}_{K\wedge\frac{C_1}{(4L+2)G_1}}) \leq 2\exp\Big[-\delta_1 n\Big], \quad \forall n. \tag{A.5.5}$$

Here, $\delta_1 = \frac{\tilde{\lambda}'^2_{\text{lasso}}}{64\eta_0^2 M} \wedge \frac{\tilde{\lambda}'_{\text{lasso}}}{16\eta_0\sqrt{M}}$. Therefore by ((A.5.1)) or ((A.5.2)), ((A.5.3)), ((A.5.4)) and ((A.5.5)), we have

$$\mathrm{P}(\|\tilde{\gamma}_\bullet^{\text{lasso}} - \gamma^*\|_{\ell_2} > \frac{3}{\kappa'}\tilde{\lambda}'_{\text{lasso}}s^{\frac{1}{2}}, \mathcal{T}_{K\wedge\frac{C_1}{(4L+2)G_1}}) \leq 2p\exp\Big[-\delta_1 n\Big], \forall n.$$

This completes the proof of Proposition 6. □

A.6 Proof of Proposition 7

Proof A.11 (Proof of Proposition 7)
For notational convenience, we abuse the notation in this proof in the same way as we did in the proof of Proposition 6. That is, we let $\mathbf{y}$ and $\mathbf{X}$ be the response and design matrix in $Z^{(1)}$, i.e. $\mathbf{y} = (y_1, \ldots, y_{n/2})^{\mathrm{T}}$ and $\mathbf{X} = \mathbf{X}^{(1)}$. And we let $\mathbf{X}_j = (x_{1j}, \ldots, x_{\frac{n}{2}j})^{\mathrm{T}}$ to represent the jth column of $\mathbf{X}^{(1)}$. We have $\nabla^2 \tilde{\ell}_n^1(\gamma) = \frac{2}{n}\sum_{i=1}^{n/2} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$. This is nonnegative definite since $\forall u \in \mathbb{R}^p$, $u^{\mathrm{T}}(\frac{2}{n}\sum_{i=1}^{n/2} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}})u = \frac{2}{n}\sum_{i=1}^{n/2}(\mathbf{x}_i^{\mathrm{T}} u)^2 \geq 0$. So $\tilde{\ell}_n^1$ is a convex function. Let $\bar{\ell}_n^1(\gamma_{\mathbb{A}_2}) \equiv \frac{1}{n}\sum_{i=1}^{n/2}(\log|y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}(Z^{(2)})| - \mathbf{x}_{i\mathbb{A}_2}^{\mathrm{T}}\gamma_{\mathbb{A}_2})^2 = \tilde{\ell}_n^1(\gamma)$ for any γ such that $\gamma_{\mathbb{A}_2^{\mathrm{c}}} = 0$. Then, we have $\nabla\bar{\ell}_n^1(\gamma_{\mathbb{A}_2}) = -\frac{2}{n}\sum_{i=1}^{n/2}(\log|y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}(Z^{(2)})| - \mathbf{x}_{i\mathbb{A}_2}^{\mathrm{T}}\gamma_{\mathbb{A}_2})\mathbf{x}_{i\mathbb{A}_2}$, and $\nabla^2\bar{\ell}_n^1(\gamma_{\mathbb{A}_2}) = \frac{2}{n}\sum_{i=1}^{n/2}\mathbf{x}_{i\mathbb{A}_2}\mathbf{x}_{i\mathbb{A}_2}^{\mathrm{T}} = \frac{2}{n}\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2}$. Since $\mathbf{X}_{\mathbb{A}_2}$ is of full rank by assumption ($\mathbf{C}_3$), $\frac{2}{n}\mathbf{X}_{\mathbb{A}_2}^{\mathrm{T}}\mathbf{X}_{\mathbb{A}_2}$ is positive definite and so $\bar{\ell}_n^1$ is strictly convex. So there is a unique solution to $\nabla\bar{\ell}_n^1(\gamma_{\mathbb{A}_2}) = 0$ and $\gamma_{\mathbb{A}_2^{\mathrm{c}}} = 0$. Following from Theorem 1 and Theorem 2 in Fan et al. (2014b), we know that the LLA algorithm in $\mathcal{R}_1$ converges to

$$\tilde{\gamma}^{\mathrm{ora}} := \underset{\gamma\in\mathbb{R}^p:\gamma_{\mathbb{A}_2^{\mathrm{c}}}=\mathbf{0}}{\arg\min}\ \tilde{\ell}_n^1(\gamma)$$

after two iterations, under the following event $\mathcal{E}_1 := \{\|\tilde{\gamma}^{\mathrm{lasso}} - \gamma^*\|_\infty \leq a_0\tilde{\lambda}_1\} \cap \{\|\nabla_{\mathbb{A}_2^{\mathrm{c}}}\tilde{\ell}_n^1(\tilde{\gamma}^{\mathrm{ora}})\|_\infty < a_1\tilde{\lambda}_1\} \cap \{\min_{j\in\mathbb{A}_2}|\tilde{\gamma}_j^{\mathrm{ora}}| \geq a\tilde{\lambda}_1\}$. Now consider the event $\mathcal{E} := \mathcal{E}_1 \cap \{\hat{\beta}(Z^{(2)}) = \hat{\beta}^{\mathrm{ora2}}\}$, then under the event $\mathcal{E}$, $\tilde{\ell}_n^1 \equiv \tilde{\ell}_{n\bullet}^1$, so the LLA algorithm in $\mathcal{R}_1$ converges to the solution to

$$\tilde{\gamma}_\bullet^{\mathrm{ora}} = \underset{\gamma\in\mathbb{R}^p:\gamma_{\mathbb{A}_2^{\mathrm{c}}}=\mathbf{0}}{\arg\min}\ \tilde{\ell}_{n\bullet}^1(\gamma),$$

after two iterations. So the LLA algorithm in $\mathcal{R}_1$ converges to $\tilde{\gamma}_\bullet^{\mathrm{ora}}$ with probability at least $\mathrm{P}(\mathcal{E})$. To lower bound $\mathrm{P}(\mathcal{E})$, we have

$$\begin{aligned}\mathrm{P}(\mathcal{E}) &= \mathrm{P}\big(\{\|\tilde{\gamma}^{\mathrm{lasso}} - \gamma^*\|_\infty \leq a_0\tilde{\lambda}_1\} \cap \{\|\nabla_{\mathbb{A}_2^{\mathrm{c}}}\tilde{\ell}_n^1(\tilde{\gamma}^{\mathrm{ora}})\|_\infty < a_1\tilde{\lambda}_1\} \\ &\qquad \cap \{\min_{j\in\mathbb{A}_2}|\tilde{\gamma}_j^{\mathrm{ora}}| \geq a\tilde{\lambda}_1\} \cap \{\hat{\beta}(Z^{(2)}) = \hat{\beta}^{\mathrm{ora2}}\}\big) \\ &= \mathrm{P}\big(\{\|\tilde{\gamma}_\bullet^{\mathrm{lasso}} - \gamma^*\|_\infty \leq a_0\tilde{\lambda}_1\} \cap \{\|\nabla_{\mathbb{A}_2^{\mathrm{c}}}\tilde{\ell}_{n\bullet}^1(\tilde{\gamma}_\bullet^{\mathrm{ora}})\|_\infty < a_1\tilde{\lambda}_1\}\end{aligned}$$

$$
\begin{aligned}
&\cap \{\min_{j\in\mathbb{A}_2}|\tilde{\gamma}^{\text{ora}}_{\bullet j}| \geq a\tilde{\lambda}_1\} \cap \{\hat{\beta}(Z^{(2)}) = \hat{\beta}^{\text{ora2}}\}\big) \\
&\geq 1 - \mathrm{P}(\hat{\beta}(Z^{(2)}) \neq \hat{\beta}^{\text{ora2}}) - \mathrm{P}\big(\{\|\tilde{\gamma}^{\text{lasso}}_{\bullet} - \gamma^*\|_\infty \leq a_0\tilde{\lambda}_1\}^{\text{c}} \cap \mathcal{T}_{K\wedge\frac{C_1}{(4L+2)G_1}}\big) \\
&\quad - \mathrm{P}\big(\{\|\nabla_{\mathbb{A}_2^{\text{c}}}\tilde{\ell}^1_{n\bullet}(\tilde{\gamma}^{\text{ora}}_{\bullet})\|_\infty < a_1\tilde{\lambda}_1\}^{\text{c}} \cap \mathcal{T}_{K\wedge\frac{C_2}{(4L+2)G_2}}\big) \\
&\quad - \mathrm{P}\big(\{\min_{j\in\mathbb{A}_2}|\tilde{\gamma}^{\text{ora}}_{\bullet j}| \geq a\tilde{\lambda}_1\}^{\text{c}} \cap \mathcal{T}_{K\wedge\frac{C_3}{(4L+2)G_3}}\big) \\
&\quad - \mathrm{P}\big(\mathcal{T}^{\text{c}}_{K\wedge\frac{C_1}{(4L+2)G_1}\wedge\frac{C_2}{(4L+2)G_2}\wedge\frac{C_3}{(4L+2)G_3}}\big),
\end{aligned}
$$

where the last inequality follows from union bound technique and the fact that $\mathcal{T}_{x_1} \cap \mathcal{T}_{x_2} = \mathcal{T}_{x_1\wedge x_2}, \forall x_1, x_2 > 0$. This completes the proof of Proposition 7. □

A.7 Proof of Lemma 4

Proof A.12 (Proof of Lemma 5)

Let $S_{m-1} = \{\mathbf{v} \in \text{span}(\mathbf{P}) : \|\mathbf{v}\|_{\ell_2} = 1\}$ be the unit sphere in the range of $\mathbf{P}$. Let N be the 2δ-packing number of S_{m-1}, i.e. the maximum N such that we can choose distinct $\mathbf{v}_1, \ldots, \mathbf{v}_N$ from S_{m-1} such that $\|\mathbf{v}_i - \mathbf{v}_j\|_{\ell_2} > 2\delta, \forall i, j$. Such choice ensures that the balls $\{\mathbf{v} \in \mathbb{R}^{n/2} : \|\mathbf{v} - \mathbf{v}_j\|_{\ell_2} \leq \delta\}$ are disjoint. Thus comparing the volume in $\mathbb{R}^m$ we have $N\delta^m \leq (1+\delta)^m$. Moreover, by definition of N, we have for any $\mathbf{v} \in S_{m-1}$, there exists a $j \in \{1, \ldots, N\}$ such that $\|\mathbf{v} - \mathbf{v}_j\|_{\ell_2} \leq 2\delta$. Since $\|\mathbf{P}\eta'\|_{\ell_2} = \sup_{\mathbf{v}\in S_{m-1}} |\mathbf{v}^{\text{T}}\mathbf{P}\eta'|$, and $\mathbf{v} = \mathbf{v}_j + (\mathbf{v} - \mathbf{v}_j)$ for any j, we have $\|\mathbf{P}\eta'\|_{\ell_2} \leq \max_{j\in\{1,\ldots,N\}} |\mathbf{v}_j^{\text{T}}\mathbf{P}\eta'| + 2\delta\|\mathbf{P}\eta'\|_{\ell_2}$. This implies $\|\mathbf{P}\eta'\|_{\ell_2} \leq \frac{1}{(1-2\delta)_+} \max_{j\in\{1,\ldots,N\}} |\mathbf{v}_j^{\text{T}}\mathbf{P}\eta'| = \frac{1}{(1-2\delta)_+} \max_{j\in\{1,\ldots,N\}} |\mathbf{v}_j^{\text{T}}\eta'|$. For any $\mathbf{v} \in S_{m-1}$, we have $\|\mathbf{v}\|_{\ell_2} = 1$ and $\mathbf{v} = \mathbf{P}\mathbf{u}$ for some $\mathbf{u} \in \mathbb{R}^{n/2}$. Therefore,

$$
\begin{aligned}
&\|\mathbf{v}\|_{\max} = \max_{i\in\{1,\ldots,n/2\}} |\mathbf{e}_i^{\text{T}}\mathbf{P}\mathbf{u}| = \max_{i\in\{1,\ldots,n/2\}} |\mathbf{e}_i^{\text{T}}\mathbf{P}\mathbf{P}\mathbf{u}| \\
&\leq \max_{i\in\{1,\ldots,n/2\}} \|\mathbf{e}_i^{\text{T}}\mathbf{P}\|_{\ell_2}\|\mathbf{P}\mathbf{u}\|_{\ell_2} = \max_{i\in\{1,\ldots,n/2\}} \sqrt{\mathbf{e}_i^{\text{T}}\mathbf{P}\mathbf{e}_i} \leq \sqrt{\frac{2}{n}}\sqrt{G'}.
\end{aligned}
$$

So applying Lemma 4 with $C = \sqrt{\frac{2}{n}}t$ and $G = \sqrt{G'}$ we get

$$
\begin{aligned}
&\mathrm{P}(|\mathbf{v}_j^{\text{T}}\eta'| > t, \mathcal{T}_{K\wedge\frac{\sqrt{2}t}{(4L+2)\sqrt{nG'}}}) = \mathrm{P}(\sqrt{\frac{2}{n}}|\mathbf{v}_j^{\text{T}}\eta'| > \sqrt{\frac{2}{n}}t, \mathcal{T}_{K\wedge\frac{\sqrt{2}t}{(4L+2)\sqrt{nG'}}}) \\
&\leq 2\exp\Big[-\Big(\frac{2t^2}{16\eta_0^2 nG'} \wedge \frac{\sqrt{2}t}{8\eta_0\sqrt{nG'}}\Big)n\Big] \leq 2\exp\Big[-\Big(\frac{2t^2}{16\eta_0^2\sqrt{n}G'} \wedge \frac{\sqrt{2}t}{8\eta_0\sqrt{G'}}\Big)\sqrt{n}\Big].
\end{aligned}
$$

Therefore, apply a union bound we have

$$\begin{aligned}
&\mathrm{P}(\|\mathbf{P}\eta'\|_{\ell_2} > \frac{t}{(1-2\delta)_+}, \mathcal{T}_{K\wedge\frac{\sqrt{2}t}{(4L+2)\sqrt{nG'}}}) \leq \mathrm{P}(\max_{j\in\{1,\ldots,N\}} |\mathbf{v}_j^{\mathrm{T}}\eta'| > t, \mathcal{T}_{K\wedge\frac{\sqrt{2}t}{(4L+2)\sqrt{nG'}}}) \\
&\leq 2N\exp\Big[-\Big(\frac{2t^2}{16\eta_0^2\sqrt{n}G'} \wedge \frac{\sqrt{2}t}{8\eta_0\sqrt{G'}}\Big)\sqrt{n}\Big] \\
&\leq 2(1+\frac{1}{\delta})^m \exp\Big[-\Big(\frac{2t^2}{16\eta_0^2\sqrt{n}G'} \wedge \frac{\sqrt{2}t}{8\eta_0\sqrt{G'}}\Big)\sqrt{n}\Big].
\end{aligned}$$

So Lemma 5 is proved after replacing t with C. □

A.8 Proof of Theorem 2

Proof A.13 (Proof of Theorem 2)
By symmetry, we only need to prove the result for HBIC$^{(3)}$, i.e. $\mathrm{P}(M^{(3)}_{\hat{\lambda}_1} = \mathbb{A}_2) \to 1$. So we slightly abuse notation in this proof. We use $\mathbf{y}$ and $\mathbf{X}$ to represent $\mathbf{y}^{(1)}$ and $\mathbf{X}^{(1)}$, and still use $\mathbf{x}_i^{\mathrm{T}}$ and $\mathbf{X}_j$ to represent the ith row and jth column of $\mathbf{X}^{(1)}$, $i = 1, \ldots, \frac{n}{2}$, $j = 1, \ldots, p$. We use $\tilde{\mathbf{y}}$ to represent $\mathbf{y}^{(2)}$, and use $\tilde{\mathbf{X}}$ to represent $\mathbf{X}^{(2)}$. Let $v_i = \mathrm{e}^{\mathbf{x}_i^{\mathrm{T}}\gamma^*}$ for $i = 1, \ldots, n$. Let $\tilde{v}_i := v_{i+n/2}, i = 1, \ldots, \frac{n}{2}$. Let $\mathbf{W}^* = \mathbf{diag}\{v_1, \ldots, v_{n/2}\}$, and let $\tilde{\mathbf{W}}^* = \mathbf{diag}\{\tilde{v}_1, \ldots, \tilde{v}_{n/2}\}$. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_{\frac{n}{2}})$ and Let $\tilde{\epsilon} = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_{n/2})^{\mathrm{T}}$ with $\tilde{\epsilon}_i = \epsilon_{i+\frac{n}{2}}$, $i = 1, \ldots, \frac{n}{2}$.

For any $\lambda > 0$, we use $\hat{\gamma}^\lambda$ to represent $\hat{\gamma}^\lambda(Z^{(2)} \to Z^{(1)})$, and use M_λ to represent $M^{(3)}_\lambda$, which is the support of $\hat{\gamma}^\lambda$. Also, we use HBIC to represent HBIC$^{(3)}$, and we use $C_{n,p}$ to represent $C^{(3)}_{n,p}$. For any index set $A \subset \{1, \ldots, p\}$, we use $\mathbf{P}_A$ to represent the projection matrix $\mathbf{P}^{(1)}_A$ in this proof, since no confusion is caused.

Meanwhile, recall that the oracle estimator that corresponds to $\hat{\beta}(Z^{(2)})$ is $\hat{\beta}^{\mathrm{ora2}}$ with $\hat{\beta}^{\mathrm{ora2}}_{\mathbb{A}_1} = (\mathbf{X}^{(2)\mathrm{T}}_{\mathbb{A}_1}\mathbf{X}^{(2)}_{\mathbb{A}_1})^{-1}\mathbf{X}^{(2)\mathrm{T}}_{\mathbb{A}_1}\mathbf{y}^{(2)}$ and $\hat{\beta}^{\mathrm{ora2}}_{\mathbb{A}_1^{\mathrm{c}}} = \mathbf{0}$. By Proposition 1, we have $\mathrm{P}(\hat{\beta}(Z^{(2)}) = \hat{\beta}^{\mathrm{ora2}}) \to 1$. Let $\hat{z}_i = \log|y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}(Z^{(2)})|$ and $z_i = \log|y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\beta}^{\mathrm{ora2}}|$, $i = 1, \ldots, \frac{n}{2}$. Consequently, $\mathrm{P}(\hat{z}_i = z_i, \forall i) \to 1$. Further, we let $\mathbf{z} = (z_1, \ldots, z_{n/2})^{\mathrm{T}}$.

Let $\tilde{\Lambda}_- = \{\lambda > 0 : \lambda \in \tilde{\Lambda}_1, \mathbb{A}_2 \not\subset M_\lambda\}$, $\tilde{\Lambda}_0 = \{\lambda > 0 : \lambda \in \tilde{\Lambda}_1, \mathbb{A}_2 = M_\lambda\}$, $\tilde{\Lambda}_+ = \{\lambda > 0 : \lambda \in \tilde{\Lambda}_1, \mathbb{A}_2 \subset M_\lambda \text{ and } \mathbb{A}_2 \neq M_\lambda\}$. For any $M \subset \{1, \ldots, p\}$, let $\mathrm{SSE}_M = \inf_{\gamma_M \in \mathbb{R}^{|M|}} \|\mathbf{z} - \mathbf{X}_{\mathbf{M}}\gamma_{\mathbf{M}}\|^2_{\ell_2}$, and let $\hat{\sigma}^2_M = \frac{2}{n}\mathrm{SSE}_M$. By definition, we have $\frac{2}{n}\|\mathbf{z} - \mathbf{X}\hat{\gamma}^\lambda\|^2_{\ell_2} \geq \mathrm{SSE}_{M_\lambda}$, $\forall\lambda > 0$. Let $\hat{\gamma}^{\mathrm{ora}}$ be the p-dimensional vector which satisfies $\hat{\gamma}^{\mathrm{ora}}_{\mathbb{A}_2} = (\mathbf{X}^{\mathrm{T}}_{\mathbb{A}_2}\mathbf{X}_{\mathbb{A}_2})^{-1}\mathbf{X}^{\mathrm{T}}_{\mathbb{A}_2}\mathbf{z}$ and $\hat{\gamma}^{\mathrm{ora}}_{\mathbb{A}_2^{\mathrm{c}}} = \mathbf{0}$. By Theorem 1, there exists $\tilde{\lambda} = \tilde{\lambda}_n > 0$

such that $\mathrm{P}(\hat{\gamma}^{\tilde{\lambda}_n} = \hat{\gamma}^{\mathrm{ora}}) \to 1$ as $n \to \infty$. This implies that $\mathrm{P}(\tilde{\lambda}_n \in \tilde{\Lambda}_0) \to 1$. Therefore, it suffices to show **(i)**: $\mathrm{P}\big(\inf_{\lambda\in\tilde{\Lambda}_-} \mathrm{HBIC}(\lambda) > \mathrm{HBIC}(\tilde{\lambda}_n)\big) \to 1$ and **(ii)**: $\mathrm{P}\big(\inf_{\lambda\in\tilde{\Lambda}_+} \mathrm{HBIC}(\lambda) > \mathrm{HBIC}(\tilde{\lambda}_n)\big) \to 1$.

For (i), we have

$$\begin{aligned}
&\mathrm{P}(\inf_{\lambda\in\tilde{\Lambda}_-}[\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\tilde{\lambda}_n)] > 0) \\
&= \mathrm{P}(\inf_{\lambda\in\tilde{\Lambda}_-}[\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\tilde{\lambda}_n)] > 0, \hat{\gamma}^{\tilde{\lambda}_n} = \hat{\gamma}^{\mathrm{ora}}) \\
&\quad + \mathrm{P}(\inf_{\lambda\in\tilde{\Lambda}_-}[\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\tilde{\lambda}_n)] > 0, \hat{\gamma}^{\tilde{\lambda}_n} \neq \hat{\gamma}^{\mathrm{ora}}) \\
&\geq \mathrm{P}(\inf_{\lambda\in\tilde{\Lambda}_-}[\log(\hat{\sigma}^2_{M_\lambda}/\hat{\sigma}^2_{\mathbb{A}_2}) + (|M_\lambda| - s_2)\frac{C_{n,p}\log(p)}{n}] > 0) + o(1). \qquad \text{(A.8.1)}
\end{aligned}$$

Here, the last inequality follows from $\mathrm{P}(\hat{\gamma}^{\tilde{\lambda}_n} = \hat{\gamma}^{\mathrm{ora}}) \to 1$, $\mathrm{P}(\hat{z}_i = z_i, \forall i) \to 1$, and $\frac{2}{n}\|\mathbf{z} - \mathbf{X}\hat{\gamma}^{\lambda}\|^2_{\ell_2} \geq \mathrm{SSE}_{M_\lambda}$. Write $\mathbf{z} = \mathbf{X}\gamma^* + \eta'$, where $\eta' = (\eta'_1, \ldots, \eta'_{\frac{n}{2}})^{\mathrm{T}}$. Then we have $\hat{\sigma}^2_{\mathbb{A}_2} = \frac{2}{n}\|(\mathbf{l}_{n/2} - \mathbf{P}_{\mathbb{A}_2})\mathbf{z}\|^2_{\ell_2} = \frac{2}{n}\|(\mathbf{l}_{n/2} - \mathbf{P}_{\mathbb{A}_2})\eta'\|^2_{\ell_2}$. So we have

$$\log\left(\frac{\widehat{\sigma}^2_{M_\lambda}}{\widehat{\sigma}^2_{\mathbb{A}_2}}\right) = \log\left(1 + \frac{n\left(\widehat{\sigma}^2_{M_\lambda} - \widehat{\sigma}^2_{\mathbb{A}_2}\right)/2}{\eta'^T\left(\mathbf{I}_{n/2} - \mathbf{P}_{\mathbb{A}_2}\right)\eta'}\right). \qquad \text{(A.8.2)}$$

To evaluate $\eta'^T\left(\mathbf{I}_{n/2} - \mathbf{P}_{\mathbb{A}_2}\right)\eta'$, first we have $\eta'^T(\mathbf{l}_{n/2} - \mathbf{P}_{\mathbb{A}_2})\eta' \leq \eta'^T\eta'$. Then notice that

$$\begin{aligned}
\eta'_i &= \log|y_i - \mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}\hat{\beta}^{\mathrm{ora2}}_{\mathbb{A}_1}| - \mathbf{x}^{\mathrm{T}}_i\gamma^* = \log|v_i\epsilon_i - \mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}(\hat{\beta}^{\mathrm{ora2}}_{\mathbb{A}_1} - \beta^*_{\mathbb{A}_1})| - \mathbf{x}^{\mathrm{T}}_i\gamma^* \\
&= \log|\epsilon_i - \frac{1}{v_i}\mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}(\hat{\beta}^{\mathrm{ora2}}_{\mathbb{A}_1} - \beta^*_{\mathbb{A}_1})| \triangleq \log|\epsilon_i - \zeta_i|, \forall i = 1, \ldots, \frac{n}{2},
\end{aligned}$$

where we denote $\zeta_i = \frac{1}{v_i}\mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}(\hat{\beta}^{\mathrm{ora2}}_{\mathbb{A}_1} - \beta^*_{\mathbb{A}_1})$. By definition of $\hat{\beta}^{\mathrm{ora2}}_{\mathbb{A}_1}$, we have $\zeta_i = \frac{1}{v_i}\mathbf{x}^{\mathrm{T}}_{i\mathbb{A}_1}(\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbb{A}_1}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbb{A}_1}\tilde{\mathbf{W}}^*\tilde{\epsilon}$, $i = 1, \ldots, \frac{n}{2}$. So the randomness of ζ_i comes from $\tilde{\epsilon}$. By law of iterated expectation and Lemma 3 (ii) we have

$$\begin{aligned}
\mathbb{E}[\eta'^{\mathrm{T}}\eta'] &= \sum_{i=1}^{n/2}\mathbb{E}[\log^2|\epsilon_i - \zeta_i|] = \sum_{i=1}^{n/2}\mathbb{E}\big[\mathbb{E}[\log^2|\epsilon_i - \zeta_i|\big|\tilde{\epsilon}]\big] \\
&\leq \sum_{i=1}^{n/2}\Big\{\mathbb{E}\big[\mathbb{E}[\log^2|\epsilon_i|] + \zeta_i^2 + (4L + 2\mathbb{E}\big[\big|\log|\epsilon_i|\big|\big]) \cdot |\zeta_i|\big]\Big\}
\end{aligned}$$

$$\leq \sum_{i=1}^{n/2} \Big\{ \mathbb{E}[\log^2 |\epsilon_1|] + \mathbb{E}[\zeta_i^2] + (4L + 2\mathbb{E}\big[\big|\log|\epsilon_1|\big|\big])\sqrt{\mathbb{E}[\zeta_i^2]} \Big\}.$$

In the proof of Theorem 1 (iii), we have established that $\mathbb{E}[\zeta_i^2] \leq \mathrm{var}(\epsilon_1)\frac{\Omega^2}{\Psi^2}\frac{2s_1 M}{n\varphi}$. Plugging it in we get $\mathbb{E}[\eta'^{\mathrm{T}}\eta'] = O(n)$. This implies that $\eta'^T \left(\mathbf{I}_{n/2} - \mathbf{P}_{\mathbb{A}_2}\right)\eta' = O_p(n)$.

Next, for any $\lambda \in \tilde{\Lambda}_-$, we evaluate $n(\hat{\sigma}^2_{M_\lambda} - \hat{\sigma}^2_{\mathbb{A}_2})/2$. Let $\mu = \mathbf{X}_{\mathbb{A}_2}\gamma^*_{\mathbb{A}_2}$, we have

$$\begin{aligned}\frac{n}{2}(\hat{\sigma}^2_{M_\lambda} - \hat{\sigma}^2_{\mathbb{A}_2}) &= \mu^T(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\mu + 2\mu^T(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\eta' - \eta'^T\mathbf{P}_{M_\lambda}\eta' + \eta'^T\mathbf{P}_{\mathbb{A}_2}\eta' \\ &\triangleq I_1 + I_2 - I_3 + I_4. \end{aligned} \tag{A.8.3}$$

By condition ((2.5.2)), $I_1 = \mu^T(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\mu > nc_0'$ for sufficiently large n. To evaluate I_2, we first write $I_2 = 2\mathbf{a}^{\mathrm{T}}_{M_\lambda}\eta'$, where $\mathbf{a}^{\mathrm{T}}_{M_\lambda} = \mu^T(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})$. We first show that $\|\mathbf{a}_{M_\lambda}\|_\infty \leq k_0\sqrt{K_n}$ for some constant $k_0 > 0$. In fact, $\|\mathbf{a}_{M_\lambda}\|_\infty \leq \|\mu\|_\infty + \|\mathbf{P}_{M_\lambda}\mu\|_\infty$. By assumption $(\mathbf{C}_1)$, we know there exists some constant $\kappa_0 > 0$ such that $|\mathbf{x}_{i\mathbb{A}_2}\gamma^*_{\mathbb{A}_2}| \leq \kappa_0, \forall i$. This gives $\|\mu\|_\infty \leq \kappa_0$. To bound $\|\mathbf{P}_{M_\lambda}\mu\|_\infty$, let $\mathbf{P}_{M_\lambda} = (p_{ij})_{1\leq i,j\leq \frac{n}{2}}$. By the condition given in the theorem, we have $p_{ii} = \mathbf{e}_i^{\mathrm{T}}\mathbf{X}_{M_\lambda}(\mathbf{X}^{\mathrm{T}}_{M_\lambda}\mathbf{X}_{M_\lambda})^{-1}\mathbf{X}^{\mathrm{T}}_{M_\lambda}\mathbf{e}_i \leq \frac{2}{n\phi}\|\mathbf{x}_{i,M_\lambda}\|^2_{\ell_2} \leq \frac{2K_nM}{n\phi}, \forall i = 1, \ldots, \frac{n}{2}$. Then we have for any $1 \leq i \leq \frac{n}{2}$, $|\mu^{\mathrm{T}}\mathbf{P}_{M_\lambda}\mathbf{e}_i| \leq \sqrt{\mu^{\mathrm{T}}\mathbf{P}_{M_\lambda}\mu}\sqrt{\mathbf{e}_i^{\mathrm{T}}\mathbf{P}_{M_\lambda}\mathbf{e}_i} \leq \|\mu\|_2\sqrt{p_{ii}} \leq \sqrt{\frac{n}{2}}\kappa_0\sqrt{\frac{2K_nM}{n\phi}} \triangleq \kappa_1\sqrt{K_n}$, with κ_1 being some positive constant. Thus we have $\|\mathbf{a}_{M_\lambda}\|_\infty \leq \kappa_0 + \kappa_1\sqrt{K_n} \leq \kappa_2\sqrt{K_n}$ for some constant $\kappa_2 > 0$.

Thus for any λ, any $t > 0$, plugging $C = t$ and $G = 1$ into Lemma 4 we get

$$\begin{aligned}&\mathrm{P}\Big(\frac{2}{n\kappa_2\sqrt{K_n}}|\mathbf{a}^{\mathrm{T}}_{M_\lambda}\eta'| > t, \mathcal{T}_{K\wedge\frac{t}{4L+2}}\Big) \leq 2\exp\Big[-\Big(\frac{t^2}{16\eta_0^2} \wedge \frac{t}{8\eta_0}\Big)n\Big], \\ &\mathrm{P}\big(\mathcal{T}^{\mathrm{c}}_{K\wedge\frac{t}{4L+2}}\big) \leq n\exp\big(-\frac{(K \wedge \frac{t}{4L+2})^2\Psi^2\varphi}{4\sigma^2\Omega^2 s_1 M}n\big),\end{aligned}$$

where K can be any positive constant, η_0 and L are fixed positive constants that have been defined in Lemma 4, and $\mathcal{T}_x$ is some event that has also been defined in Lemma 4 for any positive x.

Notice that the number of possible outcomes for M_λ is at most $\sum_{i=0}^{K_n}\binom{p}{i} \leq$

$\sum_{i=0}^{K_n} p^i = \frac{p^{K_n+1}-1}{p-1} \le 2p^{K_n}$. So applying a union bound, we have for any $t \ge 0$,

$$\mathrm{P}(\sup_{\lambda\in\tilde{\Lambda}_-} \frac{2}{n\kappa_2\sqrt{K_n}}|\mathbf{a}_{M_\lambda}^{\mathrm{T}}\eta'| > t, \mathcal{T}_{K\wedge\frac{t}{4L+2}}) \le 4p^{K_n}\exp\Big[-\Big(\frac{t^2}{16\eta_0^2}\wedge\frac{t}{8\eta_0}\Big)n\Big].$$

Therefore, again by union bound, we have

$$\begin{aligned}&\mathrm{P}(\sup_{\lambda\in\tilde{\Lambda}_-} \frac{2}{n\kappa_2\sqrt{K_n}}|\mathbf{a}_{M_\lambda}^{\mathrm{T}}\eta'| > t)\\ \le& 4p^{K_n}\exp\Big[-\Big(\frac{t^2}{16\eta_0^2}\wedge\frac{t}{8\eta_0}\Big)n\Big] + n\exp\big(-\frac{(K\wedge\frac{t}{4L+2})^2\Psi^2\varphi}{4\sigma^2\Omega^2 s_1 M}n\big).\end{aligned}$$

To make the right hand side of the above inequality goes to zero, it suffices to take $t = T(\sqrt{\frac{K_n\log p}{n}}\vee\frac{K_n\log p}{n}\vee\sqrt{\frac{s_1\log n}{n}})$ with sufficiently large $T>0$. This concludes that $\sup_{\lambda\in\tilde{\Lambda}_-}\big|\frac{2\mathbf{a}_{M_\lambda}^{\mathrm{T}}\eta'}{n\sqrt{K_n}}\big| = O_p(\sqrt{\frac{K_n\log p}{n}}\vee\frac{K_n\log p}{n}\vee\sqrt{\frac{s_1\log n}{n}}) = O_p(\sqrt{\frac{K_n\log p}{n}})$, by the condition on K_n. So $\sup_{\lambda\in\tilde{\Lambda}_-}|2\mathbf{a}_{M_\lambda}^{\mathrm{T}}\eta'| = O_p(n\sqrt{\frac{K_n^2\log p}{n}}) = o_p(n)$ since $\frac{K_n^2\log p}{n} = o(1)$. So with probability going to 1, we have $I_1 + I_2 > \frac{c_0'}{2}n$ holds for all $\lambda\in\tilde{\Lambda}_-$.

To evaluate $I_3 = \eta'^T\mathbf{P}_{M_\lambda}\eta'$, notice that $\eta'\mathbf{P}_{M_\lambda}\eta' = \|\mathbf{P}_{M_\lambda}\eta'\|_{\ell_2}^2$. Recall $\mathbf{P}_{M_\lambda} = (p_{ij})_{1\le i,j\le\frac{n}{2}}$ and we have shown $p_{ii}\le\frac{2K_nM}{n\phi}, \forall i = 1,\dots,\frac{n}{2}$. Let m be the rank of M_λ, so we have $m = \mathrm{tr}(M_\lambda) \le |M_\lambda| \le K_n$. Therefore, applying Lemma 5 with $C = t$, $G' = \frac{K_nM}{\phi}$, and $\delta = \frac{1}{4}$, we have

$$\begin{aligned}&\mathrm{P}(\sqrt{I_3} > 2t, \mathcal{T}_{K\wedge\frac{\sqrt{2\phi}t}{(4L+2)\sqrt{nK_nM}}}) = \mathrm{P}(\|\mathbf{P}_{M_\lambda}\eta'\|_{\ell_2} > 2t, \mathcal{T}_{K\wedge\frac{\sqrt{2\phi}t}{(4L+2)\sqrt{nK_nM}}})\\ &\le 2\cdot 5^m\exp\Big[-\Big(\frac{2\phi t^2}{16\eta_0^2\sqrt{n}K_nM}\wedge\frac{\sqrt{2\phi}t}{8\eta_0\sqrt{K_nM}}\Big)\sqrt{n}\Big]\\ &\le 2\cdot 5^{K_n}\exp\Big[-\Big(\frac{2\phi t^2}{16\eta_0^2\sqrt{n}K_nM}\wedge\frac{\sqrt{2\phi}t}{8\eta_0\sqrt{K_nM}}\Big)\sqrt{n}\Big].\end{aligned}$$

Next, similar as before, we apply a union bound. Notice that the above inequality does not depend on the relation between $\mathbb{A}_2$ and M_λ. As long as $\lambda\in\tilde{\Lambda}_1$, we have $|M_\lambda|\le K_n$, so that the number of possible outcomes for M_λ is at most $2p^{K_n}$. Therefore we have

$$\mathrm{P}(\sup_{\lambda\in\tilde{\Lambda}_1}\sqrt{I_3} > 2t, \mathcal{T}_{K\wedge\frac{\sqrt{2\phi}t}{(4L+2)\sqrt{nK_nM}}})$$

$$\leq 4 \cdot (5p)^{K_n} \exp\Big[- \Big(\frac{2\phi t^2}{16\eta_0^2 \sqrt{n} K_n M} \wedge \frac{\sqrt{2\phi} t}{8\eta_0 \sqrt{K_n M}} \Big) \sqrt{n} \Big].$$

Also, by Lemma 4 we know

$$\mathrm{P}(\mathcal{T}^{\mathrm{c}}_{K \wedge \frac{\sqrt{2\phi} t}{(4L+2)\sqrt{n K_n M}}}) \leq n \exp\big(-\frac{(K \wedge \frac{\sqrt{2\phi} t}{(4L+2)\sqrt{n K_n M}})^2 \Psi^2 \varphi}{4\sigma^2 \Omega^2 s_1 M} n\big),$$

therefore by union bound again we have

$$\begin{aligned} \mathrm{P}(\sup_{\lambda \in \tilde{\Lambda}_1} \sqrt{I_3} > 2t) \leq & 4 \cdot (5p)^{K_n} \exp\Big[- \Big(\frac{2\phi t^2}{16\eta_0^2 \sqrt{n} K_n M} \wedge \frac{\sqrt{2\phi} t}{8\eta_0 \sqrt{K_n M}} \Big) \sqrt{n} \Big] \\ & + n \exp\big(-\frac{(K \wedge \frac{\sqrt{2\phi} t}{(4L+2)\sqrt{n K_n M}})^2 \Psi^2 \varphi}{4\sigma^2 \Omega^2 s_1 M} n\big). \end{aligned}$$

To make the right hand side of the above inequality go to zero, it suffices to take $t = T(K_n \sqrt{\log p} \vee \frac{K_n \sqrt{K_n} \log p}{\sqrt{n}} \vee \sqrt{K_n s_1 \log n})$ with sufficiently large $T > 0$ (notice that $\log 5p$ is of same order as $\log p$ since $p \to \infty$). Since we require $K_n^2 \log p = o(n)$, and we have the relations $s_1 < K_n$, $n < p$, this implies $\sup_{\lambda \in \tilde{\Lambda}_1} \sqrt{I_3} = O_p(K_n \sqrt{\log p})$, and therefore $\sup_{\lambda \in \tilde{\Lambda}_1} I_3 = O_p(K_n^2 \log p) = o_p(n)$. In particular, $\sup_{\lambda \in \tilde{\Lambda}_-} |I_3| = o_p(n)$. Thus with probability going to 1, we have $I_1 + I_2 - I_3 > \frac{c_0'}{4} n$ holds for all $\lambda \in \tilde{\Lambda}_-$.

For $I_4 = \eta'^T \mathbf{P}_{\mathbb{A}_2} \eta'$, notice this term does not depend on λ. The procedure of bounding I_4 is similar to I_3, but we need to replace $\mathbf{P}_{M_\lambda}$ with $\mathbf{P}_{\mathbb{A}_2}$. We now have $\mathbf{e}_i^{\mathrm{T}} \mathbf{P}_{\mathbb{A}_2} \mathbf{e}_i \leq \frac{2 s_2 M}{n\phi}$, and the rank of $\mathbf{P}_{\mathbb{A}_2}$ is s_2 (by assumption $(\mathbf{C}_3)$). So similar as before, we apply Lemma 5 with $C = t$, $G' = \frac{s_2 M}{\phi}$, $\delta = \frac{1}{4}$, and then apply a union bound, we obtain

$$\begin{aligned} \mathrm{P}(\sqrt{I_4} > 2t) \leq & 2 \cdot 5^{s_2} \exp\Big[- \Big(\frac{2\phi t^2}{16\eta_0^2 \sqrt{n} s_2 M} \wedge \frac{\sqrt{2\phi} t}{8\eta_0 \sqrt{s_2 M}} \Big) \sqrt{n} \Big] \\ & + n \exp\big(-\frac{(K \wedge \frac{\sqrt{2\phi} t}{(4L+2)\sqrt{n s_2 M}})^2 \Psi^2 \varphi}{4\sigma^2 \Omega^2 s_1 M} n\big). \end{aligned}$$

By similar technique as before, it follows that $\sqrt{I_4} = O_p(\sqrt{s_2 s_1 \log n}) = o_p(\sqrt{n})$ since $\max\{s_1, s_2\} < K_n$, $\log n < \log p$ and $K_n^2 \log p = o(n)$. So $I_4 = o_p(n)$. Thus with probability going to 1, we have $I_1 + I_2 - I_3 + I_4 > \frac{c_0'}{8} n$ holds for all $\lambda \in \tilde{\Lambda}_-$.

Now, following ((A.8.3)), we have $\mathrm{P}(\frac{n}{2}(\hat{\sigma}^2_{M_\lambda} - \hat{\sigma}^2_{\mathbb{A}_2}) > \frac{c_0'}{8} n) \to 1$. Recall that we

have ((A.8.2)). By the fact that we have derived $\eta'^T \left(\mathbf{I}_{n/2} - \mathbf{P}_{\mathbb{A}_2}\right) \eta' = O_p(n)$, we have

$$\lim_{T\to\infty} \limsup_{n\to\infty} \mathrm{P}(\eta'^T \left(\mathbf{I}_{n/2} - \mathbf{P}_{\mathbb{A}_2}\right) \eta' > Tn) = 0.$$

Therefore, there exists $\epsilon_T > 0$ that depends only on T and satisfies $\epsilon_T \to 0$ as $T \to \infty$, such that

$$\liminf_{n\to\infty} \mathrm{P}\big(\inf_{\lambda\in\tilde{\Lambda}_-} \log(\frac{\hat{\sigma}^2_{M_\lambda}}{\hat{\sigma}^2_{\mathbb{A}_2}}) > \log(1 + \frac{c_0'}{8T})\big) \geq 1 - \epsilon_T.$$

Following ((A.8.1)), we have

$$\begin{aligned}
&\liminf_{n\to\infty} \mathrm{P}\big(\inf_{\lambda\in\tilde{\Lambda}_-} [\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\tilde{\lambda}_n)] > 0\big) \\
&\geq \liminf_{n\to\infty} \mathrm{P}\Big(\log(1 + \frac{c_0'}{8T}) - \frac{C_{n,p} s_2 \log(p)}{n} > 0\Big) \\
&\geq 1 - \epsilon_T,
\end{aligned}$$

where the last step is because $\frac{C_{n,p} s_2 \log(p)}{n} = o(1)$. The left hand side of the above inequality chain does not depend on T, so let $T \to \infty$ we get

$$\liminf_{n\to\infty} \mathrm{P}\big(\inf_{\lambda\in\tilde{\Lambda}_-} [\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\tilde{\lambda}_n)] > 0\big) \geq 1,$$

which implies $\mathrm{P}(\inf_{\lambda\in\tilde{\Lambda}_-} [\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\tilde{\lambda}_n)] > 0) \to 1$ as $n \to \infty$.

For (ii), consider any $\lambda \in \tilde{\Lambda}_+$. Then we have $\mathbb{A}_2 \subset M_\lambda$ and $\mathbb{A}_2 \neq M_\lambda$. So by the relation $\mathbf{z} = \mathbf{X}\gamma^* + \eta'$, we have $\mathbf{z}^{\mathrm{T}}(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\mathbf{z} = \eta'^{\mathrm{T}}(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\eta'$. Therefore we have $\frac{n}{2}(\hat{\sigma}^2_{\mathbb{A}_2} - \hat{\sigma}^2_{M_\lambda}) = \eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2})\eta'$. Therefore, we have

$$0 \leq \log(\frac{\hat{\sigma}^2_{\mathbb{A}_2}}{\hat{\sigma}^2_{M_\lambda}}) = \log(1 + \frac{\eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2})\eta'}{\eta'^{\mathrm{T}}(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\eta'}) \leq \frac{\eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2})\eta'}{\eta'^{\mathrm{T}}(\mathbf{l}_{n/2} - \mathbf{P}_{M_\lambda})\eta'},$$

where the last step is due to $\log(1 + x) \leq x, \forall x \geq 0$. Similar to ((A.8.1)), we have

$$\begin{aligned}
&\mathrm{P}\big(\inf_{\lambda\in\tilde{\Lambda}_+} [\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\tilde{\lambda}_n)] > 0\big) \\
\geq &\mathrm{P}\big(\inf_{\lambda\in\tilde{\Lambda}_+} [-\log(\frac{\hat{\sigma}^2_{\mathbb{A}_2}}{\hat{\sigma}^2_{M_\lambda}}) + (|M_\lambda| - s_2)\frac{C_{n,p}\log p}{n}] > 0\big) + o(1)
\end{aligned}$$

$$
\begin{aligned}
&\geq \mathrm{P}\Big(\inf_{\lambda\in\tilde{\Lambda}_+}\big[-\frac{\eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda}-\mathbf{P}_{\mathbb{A}_2})\eta'}{\eta'^{\mathrm{T}}(\mathbf{l}_{n/2}-\mathbf{P}_{M_\lambda})\eta'}+(|M_\lambda|-s_2)\frac{C_{n,p}\log p}{n}\big]>0\Big)+o(1)\\
&\geq \mathrm{P}\left(\inf_{\lambda\in\tilde{\Lambda}_+}\Big[\Big(|M_\lambda|-s_2\Big)\Big(\frac{C_{n,p}\log p}{n}-\frac{\eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda}-\mathbf{P}_{\mathbb{A}_2})\eta'/(|M_\lambda|-s_2)}{\eta'^{\mathrm{T}}(\mathbf{l}_{n/2}-\mathbf{P}_{M_\lambda})\eta'}\Big)\Big]>0\right)\\
&\quad+o(1).
\end{aligned}
$$

Because $|M_\lambda|-s_2\geq 1$, it suffices to show that

$$
\mathrm{P}\left(\inf_{\lambda\in\tilde{\Lambda}_+}\Big[\Big(\frac{C_{n,p}\log p}{n}-\frac{\eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda}-\mathbf{P}_{\mathbb{A}_2})\eta'/(|M_\lambda|-s_2)}{\eta'^{\mathrm{T}}(\mathbf{l}_{n/2}-\mathbf{P}_{M_\lambda})\eta'}\Big)\Big]>0\right)\to 1 \tag{A.8.4}
$$

as $n\to\infty$. We first evaluate $\eta'^{\mathrm{T}}(\mathbf{l}_{n/2}-\mathbf{P}_{M_\lambda})\eta'$. We can write it as the sum of two terms, $\eta'^{\mathrm{T}}(\mathbf{l}_{n/2}-\mathbf{P}_{M_\lambda})\eta'=\eta'^{\mathrm{T}}(\mathbf{l}_{n/2}-\mathbf{P}_{\mathbb{A}_2})\eta'-\eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda}-\mathbf{P}_{\mathbb{A}_2})\eta'\triangleq I_5-I_6$. For I_5, we have $I_5=\eta'^{\mathrm{T}}(\mathbf{l}_{n/2}-\mathbf{P}_{\mathbb{A}_2})\eta'=\eta'^{\mathrm{T}}\eta'-I_4$, notice that this term does not depend on λ. Since we have derived $|I_4|=o_p(n)$, we need to evaluate $\eta'^{\mathrm{T}}\eta'=\sum_{i=1}^{n/2}\log^2|\epsilon_i-\zeta_i|$. We claim that $\frac{\eta'^{\mathrm{T}}\eta'}{n/2}\xrightarrow{\mathrm{P}}\mathbb{E}[\log^2|\epsilon_1|]$. To prove this, first we have for any $t>0$,

$$
\begin{aligned}
&\mathrm{P}(|\frac{\eta'^{\mathrm{T}}\eta'}{n/2}-\mathbb{E}[\log^2|\epsilon_1|]|>t)=\mathbb{E}\big[\mathrm{P}\big(|\frac{\eta'^{\mathrm{T}}\eta'}{n/2}-\mathbb{E}[\log^2|\epsilon_1|]|>t\big|\tilde{\epsilon}\big)\big]\\
&\leq \mathbb{E}\big[\mathrm{P}\big(|\frac{\sum_{i=1}^{n/2}(\log^2|\epsilon_i-\zeta_i|-\mathbb{E}\big[\log^2|\epsilon_i-\zeta_i|\big|\tilde{\epsilon}\big])}{n/2}|>\frac{t}{2}\big|\tilde{\epsilon}\big)\big]\\
&+\mathbb{E}\big[\mathrm{P}\big(|\frac{\sum_{i=1}^{n/2}(\mathbb{E}\big[\log^2|\epsilon_i-\zeta_i|\big|\tilde{\epsilon}\big]-\mathbb{E}[\log^2|\epsilon_i|])}{n/2}|>\frac{t}{2}\big|\tilde{\epsilon}\big)\big]\\
&\triangleq P_1+P_2.
\end{aligned}
\tag{A.8.5}
$$

For P_1, notice that conditioning on $\tilde{\epsilon}$, $\log^2|\epsilon_i-\zeta_i|-\mathbb{E}\big[\log^2|\epsilon_i-\zeta_i|\big|\tilde{\epsilon}\big]$ are independent with mean zero. So by Lemma 3 (iii), we have

$$
\begin{aligned}
P_1&\leq \mathbb{E}\Big[\frac{\sum_{i=1}^{n/2}\mathrm{var}\big(\log^2|\epsilon_i-\zeta_i|\big|\tilde{\epsilon}\big)}{n^2t^2/4}\Big]\leq \mathbb{E}\Big[\frac{\sum_{i=1}^{n/2}\mathbb{E}\big[\log^4|\epsilon_i-\zeta_i|\big|\tilde{\epsilon}\big]}{n^2t^2/4}\Big]\\
&\leq \frac{\mathbb{E}[\log^4|\epsilon_1|]}{nt^2/4}+\frac{1}{n^2t^2/4}\cdot\mathbb{E}\Big[\sum_{i=1}^{n/2}\zeta_i^4+4\mathbb{E}\big[\big|\log|\epsilon_1|\big|\big]|\zeta_i|^3\\
&\quad+6\mathbb{E}[\log^2|\epsilon_1|]\zeta_i^2+(48L+4\mathbb{E}\big[\big|\log^3|\epsilon_1|\big|\big])|\zeta_i|\Big]
\end{aligned}
$$

$$\leq \frac{4}{nt^2}\Big(\mathbb{E}[\log^4|\epsilon_1|] + \max_i \mathbb{E}[\zeta_i^4] + 4\mathbb{E}\big[\big|\log|\epsilon_1|\big|\big]\cdot\max_i \mathbb{E}[|\zeta_i|^3] \\ + 6\mathbb{E}[\log^2|\epsilon_1|]\cdot\max_i \mathbb{E}[\zeta_i^2] + (48L + 4\mathbb{E}\big[\big|\log^3|\epsilon_1|\big|\big])\cdot\sqrt{\max_i \mathbb{E}[\zeta_i^2]}\Big). \quad \text{(A.8.6)}$$

Here all the maximum is taken over $i \in \{1,\ldots,\frac{n}{2}\}$. Recall from the proof of Theorem 1 (iii) that $\mathbb{E}[\zeta_i^2] \leq \mathrm{var}(\epsilon_1)\frac{\Omega^2}{\Psi^2}\frac{2s_1M}{n\varphi} = o(1)$. By Cauchy-Schwarz inequality we have $\max_i \mathbb{E}[|\zeta_i|^3] \leq \sqrt{\max_i \mathbb{E}[\zeta_i^4]}\sqrt{\max_i \mathbb{E}[\zeta_i^2]}$. For any i, we give an upper bound for $\mathbb{E}[\zeta_i^4]$. Recall that $\zeta_i = \frac{1}{v_i}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{W}}^*\tilde{\epsilon}$. By Proposition 2, assumptions $(\mathbf{C}_1)$ and $(\mathbf{C}_3)$ we have

$$\mathrm{P}(|\zeta_i| > t) \leq 2\exp\{-\frac{t^2}{2\sigma^2\frac{1}{v_i^2}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{W}}^{*2}\tilde{\mathbf{X}}_{\mathbb{A}_1}(\tilde{\mathbf{X}}_{\mathbb{A}_1}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathbb{A}_1})^{-1}\mathbf{x}_{i\mathbb{A}_1}}\} \\ \leq 2\exp\{-\frac{t^2}{2\sigma^2\frac{\Omega^2}{\Psi^2}\frac{2}{n\varphi}\mathbf{x}_{i\mathbb{A}_1}^{\mathrm{T}}\mathbf{x}_{i\mathbb{A}_1}}\} \leq 2\exp\{-\frac{t^2}{2\sigma^2\frac{\Omega^2}{\Psi^2}\frac{2s_1M}{n\varphi}}\}.$$

Therefore,

$$\mathbb{E}[|\zeta_i|^4] = 4\int_0^\infty t^3\mathrm{P}(|\zeta_i| > t)\,\mathrm{d}t \leq 8\int_0^\infty t^3\mathrm{e}^{-\frac{t^2}{2\sigma^2\frac{\Omega^2}{\Psi^2}\frac{2s_1M}{n\varphi}}}\,\mathrm{d}t \\ = 4\int_{-\infty}^\infty |t|^3\mathrm{e}^{-\frac{t^2}{2\sigma^2\frac{\Omega^2}{\Psi^2}\frac{2s_1M}{n\varphi}}}\,\mathrm{d}t \\ = \sqrt{2\pi\sigma^2\frac{\Omega^2}{\Psi^2}\frac{2s_1M}{n\varphi}}\cdot 4\mathbb{E}[|X|^3] \text{ where } X \sim N(0,\sigma^2\frac{\Omega^2}{\Psi^2}\frac{2s_1M}{n\varphi}) \\ = \sqrt{2\pi\sigma^2\frac{\Omega^2}{\Psi^2}\frac{2s_1M}{n\varphi}}\cdot 4\big(\sigma\frac{\Omega}{\Psi}\sqrt{\frac{2s_1M}{n\varphi}}\big)^3\cdot\mathbb{E}[|Z|^3] \text{ where } Z \sim N(0,1) \\ = O(\frac{s_1^2}{n^2}) = o(1),$$

and notice that this bound does not depend on i. Therefore, we have already shown that $\max_i \mathbb{E}[\zeta_i^4] = o(1)$, $\max_i \mathbb{E}[|\zeta_i|^3] = o(1)$, and $\max_i \mathbb{E}[\zeta_i^2] = o(1)$. Plugging these into ((A.8.6)) we have $P_1 = O(\frac{1}{n})$, so $P_1 \to 0$ as $n \to \infty$.

Now we look at P_2. We have

$$P_2 \leq \mathbb{E}\Big[\mathrm{P}\Big(\frac{2}{n}\sum_{i=1}^{n/2}\big|\mathbb{E}\big[\log^2|\epsilon_i - \zeta_i|\big|\tilde{\epsilon}\big] - \mathbb{E}[\log^2|\epsilon_i|]\big| > \frac{t}{2}\Big|\tilde{\epsilon}\Big)\Big]$$

$$
\begin{aligned}
&\leq \mathrm{P}\Big(\frac{2}{n}\sum_{i=1}^{n/2}\big\{\zeta_i^2 + (4L + 2\mathbb{E}\big[\big|\log|\epsilon_1|\big|\big]) \cdot |\zeta_i|\big\} > \frac{t}{2}\Big) \\
&\leq \mathrm{P}(\frac{2}{n}\sum_{i=1}^{n/2}\zeta_i^2 > \frac{t}{4}) + \mathrm{P}(\frac{2}{n}\sum_{i=1}^{n/2}|\zeta_i| > \frac{t}{4(4L + 2\mathbb{E}\big[\big|\log|\epsilon_1|\big|\big])}) \\
&\triangleq P_{21} + P_{22},
\end{aligned}
$$

where the first step is by triangle inequality, the second step is by Lemma 3 (ii), and the third step is by union bound. Now we have

$$
\begin{aligned}
P_{21} &\leq \frac{\mathbb{E}[\frac{2}{n}\sum_{i=1}^{n/2}\zeta_i^2]}{t/4} \leq \frac{\max_i \mathbb{E}[\zeta_i^2]}{t/4} \overset{n\to\infty}{\to} 0, \\
P_{22} &\leq \frac{\mathbb{E}[\frac{2}{n}\sum_{i=1}^{n/2}|\zeta_i|]}{t/(16L + 8\mathbb{E}\big[\big|\log|\epsilon_1|\big|\big])} \leq \frac{\max_i \mathbb{E}[|\zeta_i|]}{t/(16L + 8\mathbb{E}\big[\big|\log|\epsilon_1|\big|\big])} \\
&\leq \frac{\sqrt{\max_i \mathbb{E}[\zeta_i^2]}}{t/(16L + 8\mathbb{E}\big[\big|\log|\epsilon_1|\big|\big])} \overset{n\to\infty}{\to} 0,
\end{aligned}
$$

since we have shown $\max_i \mathbb{E}[\zeta_i^2] = o(1)$. Therefore, $P_2 \to 0$ as $n \to \infty$. So by ((A.8.5)) and arbitrariness of $t > 0$, we get $\frac{\eta'^{\mathrm{T}}\eta'}{n/2} \overset{\mathrm{P}}{\to} \mathbb{E}[\log^2|\epsilon_1|]$. Therefore, with probability going to 1, we have $I_5 > \frac{n}{4}\mathbb{E}[\log^2|\epsilon_1|]$.

For $I_6 = \eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2})\eta'$, we have $0 \leq I_6 \leq \eta'^{\mathrm{T}}\mathbf{P}_{M_\lambda}\eta' = I_3$. Since we have already shown that $\sup_{\lambda\in\tilde{\Lambda}_1}|I_3| = o_p(n)$, and $\tilde{\Lambda}_+ \subset \tilde{\Lambda}_1$, we have $\sup_{\lambda\in\tilde{\Lambda}_+}|I_6| = o_p(n)$. Therefore, with probability going to 1, we have $\eta'^{\mathrm{T}}(\mathbf{I}_{n/2} - \mathbf{P}_{M_\lambda})\eta' = I_5 - I_6 > \frac{n}{8}\mathbb{E}[\log^2|\epsilon_1|]$ holds for all $\lambda \in \tilde{\Lambda}_+$.

Next, it remains to evaluate the term $\eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2})\eta'/(|M_\lambda| - s_2) \triangleq I_7$ that appeared in ((A.8.4)). By the condition given in the theorem, and the fact that $\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}$ is a projection matrix, we have $\mathrm{tr}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}) = \mathrm{rank}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}) = |M_\lambda| - s_2$. For ease of notation let $B = \mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2}$. Then we have $\mathrm{tr}(B) \leq K_n - s_2$ since $|M_\lambda| \leq K_n$. Since B is positive definite for $\lambda \in \tilde{\Lambda}_+$, we have $\eta'^{\mathrm{T}}B\eta' \leq 2(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}])^{\mathrm{T}}B(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}]) + 2(\mathbb{E}[\eta'|\tilde{\epsilon}])^{\mathrm{T}}B(\mathbb{E}[\eta'|\tilde{\epsilon}])$. Let $K' = \sqrt{\frac{Ts_1\log n}{n}}$, where $T > 0$ can be any positive constant. We take T as fixed at present. Recall $\mathcal{T}_{K'} = \{\max_{1\leq i\leq\frac{n}{2}}|\zeta_i| \leq K'\}$. For a fixed $\lambda \in \tilde{\Lambda}_+$ which satisfies $\mathrm{tr}(B) = m \in \{1, \ldots, K_n - s_2\}$, we have for

any $t \geq 0$,

$$\begin{aligned}
&\mathrm{P}(\eta'^{\mathrm{T}} B\eta' - 2\mathbb{E}[\log^2|\epsilon_1|]m \geq t, \mathcal{T}_{K'}) \\
&= \mathbb{E}\big[\mathrm{P}\big(\eta'^{\mathrm{T}} B\eta' - 2\mathbb{E}[\log^2|\epsilon_1|]m \geq t, \mathcal{T}_{K'}\big|\tilde{\epsilon}\big)\big] \\
&\leq \mathbb{E}\big[\mathrm{P}\big(2(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}])^{\mathrm{T}} B(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}]) - 2\mathbb{E}[\log^2|\epsilon_1|]m \geq \frac{t}{2}, \mathcal{T}_{K'}\big|\tilde{\epsilon}\big)\big] \\
&+ \mathbb{E}\big[\mathrm{P}\big(2\mathbb{E}[\eta'|\tilde{\epsilon}]^{\mathrm{T}} B\mathbb{E}[\eta'|\tilde{\epsilon}] \geq \frac{t}{2}, \mathcal{T}_{K'}\big|\tilde{\epsilon}\big)\big] \\
&\triangleq I_8 + I_9. \qquad \text{(A.8.7)}
\end{aligned}$$

We bound I_8 first. Let $\sigma_i^2 = \mathrm{var}(\log|\epsilon_i - \zeta_i|\big|\tilde{\epsilon})$, so we have $\sigma_i^2 \leq \mathbb{E}[\log^2|\epsilon_i - \zeta_i|\big|\tilde{\epsilon}]$. Further, by Lemma 3 we know $\sigma_i^2 \leq \mathbb{E}[\log^2|\epsilon_1|] + \zeta_i^2 + (4L + 2\mathbb{E}\big[\big|\log|\epsilon_1|\big|\big])|\zeta_i|$. Under $\mathcal{T}_{K'}$ and by the choice of K', there exists some fixed positive constant c_1 such that for sufficiently large n,

$$\sigma_i^2 - \mathbb{E}[\log^2|\epsilon_1|] \leq K'^2 + (4L + 2\mathbb{E}\big[\big|\log|\epsilon_1|\big|\big])K' \leq c_1\sqrt{\frac{Ts_1\log n}{n}}, \qquad \text{(A.8.8)}$$

since our conditions in the theorem imply that $\frac{s_1\log n}{n} = o(1)$.

Let b_{ij} be the (i,j)th component of the matrix B. Also, let $\mathbf{b}_i^{\mathrm{T}} = (b_{i1}, \ldots, b_{i\frac{n}{2}})$ be the ith row of B. We have

$$\begin{aligned}
I_8 &\leq \mathbb{E}\big[\mathrm{P}\big(2(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}])^{\mathrm{T}} B(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}]) - 2\sum_{i=1}^{n/2}\sigma_i^2 b_{ii} \geq \frac{t}{4}, \mathcal{T}_{K'}\big|\tilde{\epsilon}\big)\big] \\
&+ \mathbb{E}\big[\mathrm{P}\big(2\sum_{i=1}^{n/2}\sigma_i^2 b_{ii} - 2\mathbb{E}[\log^2|\epsilon_1|]m \geq \frac{t}{4}, \mathcal{T}_{K'}\big|\tilde{\epsilon}\big)\big] \\
&\triangleq I_{81} + I_{82}.
\end{aligned}$$

For I_{81}, we have

$$I_{81} = \mathbb{E}\big[1_{\mathcal{T}_{K'}}\mathrm{P}\big((\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}])^{\mathrm{T}} B(\eta' - \mathbb{E}[\eta'|\tilde{\epsilon}]) - \sum_{i=1}^{n/2}\sigma_{ii}^2 b_{ii} \geq \frac{t}{8}\big|\tilde{\epsilon}\big)\big]. \qquad \text{(A.8.9)}$$

Since $K' = o(1)$, by Lemma 2, there exists some fixed positive constant c_2, such that under the event $\mathcal{T}_{K'}$ and conditioning on $\tilde{\epsilon}$, $\log|\epsilon_i - \zeta_i| - \mathbb{E}[\log|\epsilon_i - \zeta_i|\big|\tilde{\epsilon}], i = 1, \ldots, \frac{n}{2}$

are independent sub-exponential(c_2) random variable. Equivalently, this means that there exists some fixed positive constant c_2 such that, conditioning on $\tilde{\epsilon}$ and under $\mathcal{T}_{K'}$, $\|\log|\epsilon_i-\zeta_i|-\mathbb{E}[\log|\epsilon_i-\zeta_i||\tilde{\epsilon}]\|_{\psi_1}\le c_2, \forall i=1,\dots,\frac{n}{2}$. Here $\|X\|_{\psi_1}:=\inf\{t>0:\mathbb{E}[\exp(|X|/t)]\le 2\}$ refers to the Orlicz norm of a random variable X.

Next, applying Proposition 1.5 of Götze et al. (2021) with $q=2$, we have

$$
\begin{aligned}
&1_{\mathcal{T}_{K'}}\mathrm{P}\big((\eta'-\mathbb{E}[\eta'|\tilde{\epsilon}])^{\mathrm{T}}B(\eta'-\mathbb{E}[\eta'|\tilde{\epsilon}])-\sum_{i=1}^{n/2}\sigma_{ii}^2 b_{ii}\ge\frac{t}{8}\big|\tilde{\epsilon}\big)\\
\le& 2\exp\Bigg(-\frac{1}{c_3}\min\Bigg\{\frac{t^2}{64c_2^4\|B\|_F^2},\frac{t}{8c_2^2\|B\|},\\
&\qquad\Big(\frac{t}{8c_2^2\max_{i=1,\dots,n/2}\|\mathbf{b}_i\|_{\ell_2}}\Big)^{\frac{2}{3}},\Big(\frac{t}{8c_2^2\|B\|_\infty}\Big)^{\frac{1}{2}}\Bigg\}\Bigg),
\end{aligned}
\tag{A.8.10}
$$

where $\|B\|_F=\sqrt{\mathrm{tr}(B^{\mathrm{T}}B)}$ is the Frobenius norm and $\|B\|_\infty=\max_{i,j}|b_{ij}|$, and $c_3>0$ is some absolute constant. Since B is a projection matrix with rank m, we have $\|B\|_F=\sqrt{\mathrm{tr}(B)}=\sqrt{m}$, $\|B\|=\lambda_{\max}(B)=1$. Also, by the condition given in the theorem, we have $b_{ii}=\mathbf{e}_i^{\mathrm{T}}(\mathbf{P}_{M_\lambda}-\mathbf{P}_{\mathbb{A}_2})\mathbf{e}_i\le\mathbf{e}_i^{\mathrm{T}}\mathbf{P}_{M_\lambda}\mathbf{e}_i=\mathbf{e}_i^{\mathrm{T}}\mathbf{X}_{M_\lambda}(\mathbf{X}_{M_\lambda}^{\mathrm{T}}\mathbf{X}_{M_\lambda})^{-1}\mathbf{X}_{M_\lambda}^{\mathrm{T}}\mathbf{e}_i\le\frac{2}{n\phi}\|\mathbf{x}_{i,M_\lambda}\|_{\ell_2}^2\le\frac{2|M_\lambda|M}{n\phi},\forall i=1,\dots,\frac{n}{2}$. Since $|M_\lambda|=m+s_2$, we have $b_{ii}\le\frac{2(m+s_2)M}{n\phi}$. Since B is positive definite, we have $|b_{ij}|\le\sqrt{b_{ii}b_{jj}}\le\frac{2(m+s_2)M}{n\phi},\forall i,j$. So $\|B\|_\infty\le\frac{2(m+s_2)M}{n\phi}$. For any $i\in\{1,\dots,\frac{n}{2}\}$, we have $\|\mathbf{b}_i\|_{\ell_2}=\|\mathbf{e}_i^{\mathrm{T}}B\|_{\ell_2}=\sqrt{\mathbf{e}_i^{\mathrm{T}}BB\mathbf{e}_i}=\sqrt{\mathbf{e}_i^{\mathrm{T}}B\mathbf{e}_i}=\sqrt{b_{ii}}\le\sqrt{\frac{2(m+s_2)M}{n\phi}}$. Combining these bound with ((A.8.9)) and ((A.8.10)), we have

$$
\begin{aligned}
I_{81}\le 2\exp\Bigg(-\frac{1}{c_3}\min\Bigg\{&\frac{t^2}{64c_2^4m},\frac{t}{8c_2^2},\\
&\Big(\frac{t\sqrt{n\phi}}{8c_2^2\sqrt{2(m+s_2)M}}\Big)^{\frac{2}{3}},\Big(\frac{tn\phi}{16c_2^2(m+s_2)M}\Big)^{\frac{1}{2}}\Bigg\}\Bigg).
\end{aligned}
$$

Now let us take $t=c_4(m\log p\vee s_1\log n)$, where c_4 can be any fixed positive constant. Then $t\ge c_4 m\log p$ and we have for sufficiently large n,

$$
\begin{aligned}
I_{81}\le 2\exp\Bigg(-\frac{1}{c_3}\min\Bigg\{&\frac{c_4^2m^2\log^2 p}{64c_2^4m},\frac{c_4m\log p}{8c_2^2},\\
&\Big(\frac{c_4m\log p\sqrt{n\phi}}{8c_2^2\sqrt{2(m+s_2)M}}\Big)^{\frac{2}{3}},\Big(\frac{c_4m\log p\,n\phi}{16c_2^2(m+s_2)M}\Big)^{\frac{1}{2}}\Bigg\}\Bigg)
\end{aligned}
$$

$$\leq 2\exp\left\{-\frac{1}{c_3}\frac{c_4 m\log p}{8c_2^2}\right\} \tag{A.8.11}$$

holds true for all $\lambda\in\tilde{\Lambda}_+$ that satisfies $\mathrm{tr}(B)=m$. Here the last inequality is because we have $m+s_2\leq K_n$ and $K_n^2\log p=o(n)$.

For I_{82}, we have

$$I_{82}=\mathbb{E}\big[1_{\mathcal{T}_{K'}}\mathrm{P}\big(\sum_{i=1}^{n/2}\sigma_i^2 b_{ii}-\mathbb{E}[\log^2|\epsilon_1|]m\geq\frac{t}{8}\big|\tilde{\epsilon}\big)\big].$$

By ((A.8.8)), we know $\sigma_i^2\leq\mathbb{E}[\log^2|\epsilon_1|]+c_1\sqrt{\frac{Ts_1\log n}{n}}$, so $\sum_{i=1}^{n/2}\sigma_i^2 b_{ii}\leq\big(\mathbb{E}[\log^2|\epsilon_1|]+c_1\sqrt{\frac{Ts_1\log n}{n}}\big)\cdot\sum_{i=1}^{n/2}b_{ii}=\big(\mathbb{E}[\log^2|\epsilon_1|]+c_1\sqrt{\frac{Ts_1\log n}{n}}\big)m$. So we have $\sum_{i=1}^{n/2}\sigma_i^2 b_{ii}-\mathbb{E}[\log^2|\epsilon_1|]m\leq c_1 m\sqrt{\frac{Ts_1\log n}{n}}$. Since $\frac{s_1\log n}{n}=o(1)$, $c_1 m\sqrt{\frac{Ts_1\log n}{n}}<c_4 m\log p\ /\ 8\leq t/8$ for sufficiently large n. Therefore, for sufficiently large n, $I_{82}=0$ for any $\lambda\in\tilde{\Lambda}_+$ that satisfies $\mathrm{tr}(B)=m$.

We bound I_9 now. Recall that

$$I_9=\mathbb{E}\big[1_{\mathcal{T}_{K'}}\mathrm{P}\big(\mathbb{E}[\eta'|\tilde{\epsilon}]^{\mathrm{T}}B\mathbb{E}[\eta'|\tilde{\epsilon}]\geq\frac{t}{4}\big|\tilde{\epsilon}\big)\big]. \tag{A.8.12}$$

By Lemma 3, under the event $\mathcal{T}_{K'}$, we have $\mathbb{E}[\log|\epsilon_i-\zeta_i|\big|\tilde{\epsilon}]\leq(2L+1)|\zeta_i|\leq(2L+1)K'$. So under $\mathcal{T}_{K'}$, $\|\mathbb{E}[\eta'|\tilde{\epsilon}]\|_\infty\leq(2L+1)K'$, and we have $\|B\mathbb{E}[\eta'|\tilde{\epsilon}]\|_{\ell_2}=\sup_{\|\mathbf{v}\|_{\ell_2}=1}|\mathbf{v}^{\mathrm{T}}B\mathbb{E}[\eta'|\tilde{\epsilon}]|\leq\sup_{\|\mathbf{v}\|_{\ell_2}=1}\|B\mathbf{v}\|_{\ell_1}(2L+1)K'$. Notice $\sup_{\|\mathbf{v}\|_{\ell_2}=1}\|B\mathbf{v}\|_{\ell_1}\leq\sup_{\|\mathbf{v}\|_{\ell_2}=1}\sqrt{\frac{n}{2}}\|B\mathbf{v}\|_{\ell_2}\leq\sqrt{\frac{n}{2}}\lambda_{\max}(B)=\sqrt{\frac{n}{2}}$, we know $\|B\mathbb{E}[\eta'|\tilde{\epsilon}]\|_{\ell_2}\leq\sqrt{\frac{n}{2}}(2L+1)K'$. So under $\mathcal{T}_{K'}$, $\mathbb{E}[\eta'|\tilde{\epsilon}]^{\mathrm{T}}B\mathbb{E}[\eta'|\tilde{\epsilon}]=\|B\mathbb{E}[\eta'|\tilde{\epsilon}]\|_{\ell_2}^2\leq(2L+1)^2\cdot\frac{n}{2}\cdot K'^2=(2L+1)^2\cdot\frac{n}{2}\cdot\frac{Ts_1\log n}{n}\triangleq c_6 Ts_1\log n$ where $c_6=\frac{(2L+1)^2}{2}$ is a fixed positive constant. Therefore, under $\mathcal{T}_{K'}$, $\mathbb{E}[\eta'|\tilde{\epsilon}]^{\mathrm{T}}B\mathbb{E}[\eta'|\tilde{\epsilon}]\leq c_6 Ts_1\log n<\frac{c_4 s_1\log n}{4}\leq\frac{t}{4}$ if we choose $c_4>4c_6T$. Under such choice, following ((A.8.12)) we have $I_9=0$. This holds for any $\lambda\in\tilde{\Lambda}_+$.

Now, combining ((A.8.7)) with the bound ((A.8.11)) and $I_{82}=I_9=0$ we have for sufficiently large n,

$$\mathrm{P}(\eta'^{\mathrm{T}}B\eta'-2\mathbb{E}[\log^2|\epsilon_1|]m\geq c_4(m\log p\vee s_1\log n),\mathcal{T}_{K'})\leq 2\exp\left\{-\frac{1}{c_3}\frac{c_4 m\log p}{8c_2^2}\right\}$$

holds for any $\lambda\in\Lambda_+$ such that $\mathrm{tr}(B)=m$. Recall that $I_7=\frac{\eta'^{\mathrm{T}}B\eta'}{\mathrm{tr}(B)}$, and $m=\mathrm{tr}(B)\geq$

1, so the above implies

$$\mathrm{P}(I_7 - 2\mathbb{E}[\log^2|\epsilon_1|] \geq c_4(\log p \vee s_1 \log n), \mathcal{T}_{K'}) \leq 2\exp\left\{-\frac{1}{c_3}\frac{c_4 m \log p}{8c_2^2}\right\}.$$

Now applying union bound we have

$$\begin{aligned}
&\mathrm{P}\Big(\sup_{\lambda\in\tilde{\Lambda}_+} I_7 \geq 2\mathbb{E}[\log^2|\epsilon_1|] + c_4(\log p \vee s_1 \log n), \mathcal{T}_{K'}\Big) \\
&\leq \sum_{m=1}^{p-s_2}\binom{p-s_2}{m} 2\exp\left\{-\frac{1}{c_3}\frac{c_4 m\log p}{8c_2^2}\right\}.
\end{aligned}$$

Take c_4 large enough so that $\frac{c_4}{8c_3c_2^2} \geq 2$, then we have

$$\begin{aligned}
&\mathrm{P}\Big(\sup_{\lambda\in\tilde{\Lambda}_+} I_7 \geq 2\mathbb{E}[\log^2|\epsilon_1|] + c_4(\log p \vee s_1 \log n), \mathcal{T}_{K'}\Big) \\
&\leq \sum_{m=1}^{p-s_2}\binom{p-s_2}{m} 2\exp\{-2m\log p\} = 2\sum_{m=1}^{p-s_2}\binom{p-s_2}{m}(\frac{1}{p^2})^m \\
&= 2\Big((1+\frac{1}{p^2})^{p-s_2} - 1\Big) \to 0.
\end{aligned}$$

Again, by Lemma 4 and our choice $K' = \sqrt{\frac{Ts_1\log n}{n}}$, we have

$$\mathrm{P}(\mathcal{T}_{K'}^{\mathrm{c}}) \leq n\exp(-\frac{K'^2\Psi^2\varphi}{4\sigma^2\Omega^2 s_1 M}n) = \exp\Big(-(Tc_7 - 1)\log n\Big)$$

where $c_7 \coloneqq \frac{\Psi^2\varphi}{4\sigma^2\Omega^2 M}$ is a fixed positive constant. Let us finally choose $T > \frac{1}{c_7}$ so that $\mathrm{P}(\mathcal{T}_{K'}^{\mathrm{c}}) \to 0$. Then by union bound we have

$$\begin{aligned}
&\mathrm{P}\Big(\sup_{\lambda\in\tilde{\Lambda}_+} I_7 \geq 2\mathbb{E}[\log^2|\epsilon_1|] + c_4(\log p \vee s_1 \log n)\Big) \\
&\leq \mathrm{P}\Big(\sup_{\lambda\in\tilde{\Lambda}_+} I_7 \geq 2\mathbb{E}[\log^2|\epsilon_1|] + c_4(\log p \vee s_1 \log n), \mathcal{T}_{K'}\Big) + \mathrm{P}(\mathcal{T}_{K'}^{\mathrm{c}}) \\
&\to 0.
\end{aligned}$$

Notice that there exists some fixed positive constant c_8 such that $2\mathbb{E}[\log^2|\epsilon_1|] + c_4(\log p \vee s_1 \log n) \leq c_8(\log p \vee s_1 \log n)$ for sufficiently large n. Therefore we can

conclude that for sufficiently large n, with probability going to 1, $\sup_{\lambda\in\tilde{\Lambda}_+}|I_7| \le \mathbb{E}[\log^2|\epsilon_1|] + c_4(\log p \vee s_1 \log n) \le c_8(\log p \vee s_1 \log n)$. Therefore we have

$$\begin{aligned}
&\mathrm{P}\left(\inf_{\lambda\in\tilde{\Lambda}_+}\left[\frac{C_{n,p}\log p}{n} - \frac{\eta'^{\mathrm{T}}(\mathbf{P}_{M_\lambda} - \mathbf{P}_{\mathbb{A}_2})\eta'/(|M_\lambda| - s_2)}{\eta'^{\mathrm{T}}(\mathbf{I}_{n/2} - \mathbf{P}_{M_\lambda})\eta'}\right] > 0\right) \\
&\ge \mathrm{P}\left(\frac{C_{n,p}\log p}{n} - \frac{c_8(\log p \vee s_1 \log n)}{\frac{n}{8}\mathbb{E}[\log^2|\epsilon_1|]} > 0\right) + o(1) \to 1
\end{aligned}$$

since $C_{n,p} \to \infty$ and $s_1 \log n = o(C_{n,p}\log p)$. So ((A.8.4)) is proved and the whole proof for the theorem is completed. □

Appendix B

Proof of Chapter 3

In this appendix we present technical proofs for all theoretical results. We will frequently need the following expressions for the loss function L_h and its derivatives in our proofs. Recall $L_h(u) = \int_{-\infty}^{\infty} |u-v| \frac{1}{h} K(\frac{v}{h}) \mathrm{d}v, u \in \mathbb{R}$. A direct calculation gives

$$\begin{aligned} L_h(u) &= u \int_{-u}^{u} \frac{1}{h} K(\frac{v}{h}) \mathrm{d}v - 2 \int_{-\infty}^{u} \frac{v}{h} K(\frac{v}{h}) \mathrm{d}v, \\ L_h'(u) &= 2 \int_{-\infty}^{u} \frac{1}{h} K(\frac{v}{h}) \mathrm{d}v - 1 = 2 \int_{0}^{u} \frac{1}{h} K(\frac{v}{h}) \mathrm{d}v, \\ L_h''(u) &= \frac{2}{h} K(\frac{u}{h}), \ \forall u \in \mathbb{R}. \end{aligned} \tag{B.0.1}$$

Meanwhile, it can be directly checked that the following identity holds

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} a_{ij} = \frac{1}{n!} \sum_{\pi \in \Pi_n} \frac{1}{[n/2]} \sum_{i=1}^{[n/2]} a_{\pi(i)\pi([\frac{n}{2}]+i)} \tag{B.0.2}$$

for any deterministic $(a_{ij})_{1 \leq i,j \leq n}$. Here $[x]$ refers to the largest integer that is no larger than x, $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is a permutation for $\{1, \ldots, n\}$ and Π_n is the set of all such permutations (so $|\Pi_n| = n!$). We will also frequently use the identity ((B.0.2)) in our proofs.

B.0.1 Proof of Theorem 3

Proof B.1 (Proof of Theorem 3)

Let $\Sigma = \mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^{\mathrm{T}}]$ be the covariance matrix of $\mathbf{x}$. Since the distribution of $\mathbf{x}$ is continuous, Σ is positive definite. For ease of notation let $S(\beta) = \mathbb{E}[L_h(y - y' - (\mathbf{x}-\mathbf{x}')^{\mathrm{T}}\beta)]$. Since $L_h(\cdot)$ is convex, we know $S(\cdot)$ is convex. By Lemma 1 and the dominated convergence theorem, we have $\nabla S(\beta) = -\mathbb{E}[L_h'(y-y'-(\mathbf{x}-\mathbf{x}')^{\mathrm{T}}\beta)(\mathbf{x}-\mathbf{x}')]$ and $\nabla^2 S(\beta) = \mathbb{E}[L_h''(y - y' - (\mathbf{x} - \mathbf{x}')^{\mathrm{T}}\beta)(\mathbf{x} - \mathbf{x}')(\mathbf{x} - \mathbf{x}')^{\mathrm{T}}]$. Plugging in β^* we have $\nabla S(\beta^*) = -\mathbb{E}[L_h'(y-y'-(\mathbf{x}-\mathbf{x}')^{\mathrm{T}}\beta^*)(\mathbf{x}-\mathbf{x}')] = -\mathbb{E}[L_h'(\epsilon-\epsilon')(\mathbf{x}-\mathbf{x}')] = -\mathbb{E}[L_h'(\epsilon - \epsilon')]\mathbb{E}[\mathbf{x} - \mathbf{x}'] = 0$, by independence between the errors and covariates. Also, we have $\nabla^2 S(\beta^*) = \mathbb{E}[L_h''(y-y'-(\mathbf{x}-\mathbf{x}')^{\mathrm{T}}\beta^*)(\mathbf{x}-\mathbf{x}')(\mathbf{x}-\mathbf{x}')^{\mathrm{T}}] = \mathbb{E}[L_h''(\epsilon-\epsilon')(\mathbf{x}-\mathbf{x}')(\mathbf{x}-\mathbf{x}')^{\mathrm{T}}] = \mathbb{E}[L_h''(\epsilon - \epsilon')]\mathbb{E}[(\mathbf{x} - \mathbf{x}')(\mathbf{x} - \mathbf{x}')^{\mathrm{T}}] = \frac{4}{h}\mathbb{E}[K(\frac{\epsilon-\epsilon'}{h})]\Sigma$. By definition, assumption 1 and assumption 2 we have $\mathbb{E}[K(\frac{\epsilon-\epsilon'}{h})] = \int_{-\infty}^{\infty} K(\frac{v}{h})g(v)\mathrm{d}v \geq \int_{-((h\delta_0)\wedge\delta_1)}^{(h\delta_0)\wedge\delta_1} K(\frac{v}{h})g(v)\mathrm{d}v \geq \kappa_l\mu_1((h\delta_0 \wedge \delta_1)) > 0$. Thus $\nabla^2 S(\beta^*)$ is a positive definite matrix. Therefore, β^* is the unique minimizer of $S(\beta)$, i.e. $\beta_h^* = \beta^*$. □

B.0.2 Proof of Lemma 1

Proof B.2 (Proof of Lemma 1)

Notice that $L_h'(t) = 2\int_{-\infty}^{t}\frac{1}{h}K(\frac{v}{h})\mathrm{d}v - 1$ and $\int_{-\infty}^{\infty} K(t)\mathrm{d}t = 1$, so we have $L_h'(-t) = 2\int_{-\infty}^{-t}\frac{1}{h}K(\frac{v}{h})\mathrm{d}v - 1 = 2(1-\int_{-t}^{\infty}\frac{1}{h}K(\frac{v}{h})\mathrm{d}v) - 1 = 1 - 2\int_{-t}^{\infty}\frac{1}{h}K(\frac{v}{h})\mathrm{d}v = 1 - 2\int_{-\infty}^{t}\frac{1}{h}K(\frac{v}{h})\mathrm{d}v = -L_h'(t)$, where the last equality is due to a change of variable and $K(-t) = K(t), \forall t$. So the first statement is proved.

By the property of the kernel function, we have $\int_{-\infty}^{\infty}\frac{1}{h}K(\frac{v}{h})\mathrm{d}v = 1$. Since $L_h'(t) = 2\int_{-\infty}^{t}\frac{1}{h}K(\frac{v}{h})\mathrm{d}v - 1$ and $K(t) \geq 0$ for all t, we know $-1 \leq L_h'(t) \leq 2\int_{-\infty}^{\infty}\frac{1}{h}K(\frac{v}{h})\mathrm{d}v - 1 = 1$, so $|L_h'(t)| \leq 1, \forall t \in \mathbb{R}$. The second statement then follows from the mean value theorem.

Similarly, by assumption 1 we have $0 \leq L_h''(t) = \frac{2}{h}K(\frac{t}{h}) \leq \frac{2}{h}\kappa_u$, which means $|L_h''(t)| \leq \frac{2}{h}\kappa_u$. Applying the mean value theorem once again gives the third statement. So the proof is finished. □

B.0.3 Proof of Theorem 1

Proof B.3 (Proof of Theorem 1)
By definition of the ℓ_1-penalized CRR estimator, we have

$$\begin{aligned}
&\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_h\big(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\tilde{\beta}^{\lambda_0}\big) \\
&- \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_h\big(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\beta^*\big) \\
&\leq \lambda_0(\|\beta^*\|_1 - \|\tilde{\beta}^{\lambda_0}\|_1) \leq \lambda_0(\|\beta^*_{\mathbb{A}} - \tilde{\beta}^{\lambda_0}_{\mathbb{A}}\|_1 + \|\tilde{\beta}^{\lambda_0}_{\mathbb{A}}\|_1 - \|\tilde{\beta}^{\lambda_0}_{\mathbb{A}}\|_1 - \|\tilde{\beta}^{\lambda_0}_{\mathbb{A}^{\mathrm{c}}} - \beta^*_{\mathbb{A}^{\mathrm{c}}}\|_1) \\
&= \lambda_0(\|\mathbf{u}_{\mathbb{A}}\|_1 - \|\mathbf{u}_{\mathbb{A}^{\mathrm{c}}}\|_1), \qquad \text{(B.0.3)}
\end{aligned}$$

where we denote $\mathbf{u} \coloneqq \tilde{\beta}^{\lambda_0} - \beta^*$. On the other hand, by convexity of $L_h(\cdot)$, we have

$$\begin{aligned}
&\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_h\big(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\tilde{\beta}^{\lambda_0}\big) \\
&- \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_h\big(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\beta^*\big) \\
&\geq -\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L'_h\big(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\beta^*\big)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\tilde{\beta}^{\lambda_0} - \beta^*) \\
&\geq -\Big\|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L'_h\big(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\beta^*\big)(\mathbf{x}_i - \mathbf{x}_j)\Big\|_\infty \|\tilde{\beta}^{\lambda_0} - \beta^*\|_1 \\
&= -\Big\|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L'_h(\epsilon_i - \epsilon_j)(\mathbf{x}_i - \mathbf{x}_j)\Big\|_\infty (\|\mathbf{u}_{\mathbb{A}}\|_1 + \|\mathbf{u}_{\mathbb{A}^{\mathrm{c}}}\|_1). \qquad \text{(B.0.4)}
\end{aligned}$$

Define event $\mathcal{E} = \{\|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L'_h(\epsilon_i - \epsilon_j)(\mathbf{x}_i - \mathbf{x}_j)\|_\infty \leq \frac{\lambda_0}{2}\}$. Then we have

$$\begin{aligned}
\mathrm{P}(\mathcal{E}^{\mathrm{c}}) &= \mathrm{P}\Big(\Big\|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L'_h(\epsilon_i - \epsilon_j)(\mathbf{x}_i - \mathbf{x}_j)\Big\|_\infty > \frac{\lambda_0}{2}\Big) \\
&\leq \sum_{k=1}^{p} \mathrm{P}\Big(\Big|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L'_h(\epsilon_i - \epsilon_j)(x_{ik} - x_{jk})\Big| > \frac{\lambda_0}{2}\Big) \qquad \text{(B.0.5)}
\end{aligned}$$

By Lemma 1 and the fact that the distribution of $\epsilon_i - \epsilon_j$ is symmetric about zero, we know $L_h'(\epsilon_i - \epsilon_j)$ is a bounded random variable with mean zero who takes its value in $[-1, 1]$. Also, we know $|x_{ik} - x_{jk}| \leq 2M$. Thus, changing the value of ϵ_i for some i will lead to a change in the value of $\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} L_h'(\epsilon_i - \epsilon_j)(x_{ik} - x_{jk})$ which is no greater than $\frac{8M}{n}$. By McDiarmid's inequality,

$$\mathrm{P}\Big(\Big|\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} L_h'(\epsilon_i - \epsilon_j)(x_{ik} - x_{jk})\Big| > \frac{\lambda_0}{2}\Big) \leq 2 \exp\Big\{ - \frac{\lambda_0^2 n}{128 M^2}\Big\}. \tag{B.0.6}$$

So combining ((B.0.5)) and ((B.0.6)) we have $\mathrm{P}(\mathcal{E}^{\mathrm{c}}) \leq 2p \exp\Big\{ - \frac{\lambda_0^2 n}{128 M^2}\Big\}$. Now, under $\mathcal{E}$, combining ((B.0.3)) and ((B.0.4)), we have $\lambda_0(\|\mathbf{u}_{\mathbb{A}}\|_1 - \|\mathbf{u}_{\mathbb{A}^{\mathrm{c}}}\|_1) \geq -\frac{\lambda_0}{2}(\|\mathbf{u}_{\mathbb{A}}\|_1 + \|\mathbf{u}_{\mathbb{A}^{\mathrm{c}}}\|_1)$, which is $\|\mathbf{u}_{\mathbb{A}^{\mathrm{c}}}\|_1 \leq 3\|\mathbf{u}_{\mathbb{A}}\|_1$. In other words, under $\mathcal{E}$, we have $\mathbf{u} \in \mathcal{S}_{\mathbb{A}}$.

Define $F(\mathbf{v}) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} L_h\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big)$ for any $\mathbf{v} \in \mathbb{R}^p$. Also, define $\mathbb{C}(r) = \Big\{\mathbf{w} \in \mathbb{R}^p : \|\mathbf{w}\|_2 = r\sqrt{\frac{s \log p}{n}}, \mathbf{w} \in \mathcal{S}_{\mathbb{A}}\Big\}$ for any $r > 0$. Let $G(\mathbf{v}) = F(\mathbf{v}) - F(\beta^*)$, and let $H(r) = \sup_{\mathbf{v} \in \beta^* + \mathbb{C}(r)} \big|G(\mathbf{v}) - \mathbb{E}[G(\mathbf{v})]\big|$.

We first give an upper bound of $H(r)$. Notice that for any $\mathbf{v} \in \beta^* + \mathbb{C}(r)$, by Lemma 1 and assumption 5,

$$\begin{aligned}
&\big|L_h\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big) - L_h\big(\epsilon_i - \epsilon_j\big)\big| \leq \big|(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big| \\
&\leq \|\mathbf{x}_i - \mathbf{x}_j\|_\infty \cdot \|\mathbf{v} - \beta^*\|_1 \\
&\leq 2M \cdot 4\|(\mathbf{v} - \beta^*)_{\mathbb{A}}\|_1 \leq 8M \cdot \sqrt{s}\|(\mathbf{v} - \beta^*)_{\mathbb{A}}\|_2 \\
&\leq 8M\sqrt{s} r \sqrt{\frac{s \log p}{n}} = 8Mrs\sqrt{\frac{\log p}{n}}.
\end{aligned} \tag{B.0.7}$$

Therefore, when viewing $H(r)$ as a function of $(\epsilon_1, \ldots, \epsilon_n)$, changing the value of a particular ϵ_i will lead to a difference in the value of $H(r)$ whose absolute value is no more than $\frac{1}{n} \cdot 2^3 \cdot 8Mrs\sqrt{\frac{\log p}{n}} = \frac{64Mrs\sqrt{\log p}}{n\sqrt{n}}$. So by McDiarmid's inequality, $\mathrm{P}\big(\big|H(r) - \mathbb{E}[H(r)]\big| > t\big) \leq 2\mathrm{e}^{-\frac{2n^2t^2}{4096M^2r^2s^2 \log p}}$. Taking $t = 64M\frac{rs \log p}{n}$, we obtain

$$\mathrm{P}\Big(\big|H(r) - \mathbb{E}[H(r)]\big| > 64M\frac{rs \log p}{n}\Big) \leq 2\mathrm{e}^{-2 \log p}. \tag{B.0.8}$$

We next derive an upper bound for $\mathbb{E}[H(r)]$. Let $(\epsilon_1', \ldots, \epsilon_n')$ be an i.i.d. copy of

$(\epsilon_1, \ldots, \epsilon_n)$. Let $(\sigma_1, \ldots, \sigma_n)$ be a random vector with its components being i.i.d. Rademacher random variables (i.e. $\mathrm{P}(\sigma_i = 1) = \mathrm{P}(\sigma_i = -1) = \frac{1}{2}$), which is independent from all the other random elements. By ((B.0.2)), for large enough n we have

$$
\begin{aligned}
\mathbb{E}[H(r)] = \mathbb{E}\Bigg[\sup_{\mathbf{v}\in\beta^*+\mathbb{C}(r)} \Bigg| \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i} \Big\{ & L_h\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big) \\
& - L_h\big(\epsilon_i - \epsilon_j\big) \\
& - \mathbb{E}\big[L_h\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big)\big] + \mathbb{E}\big[L_h\big(\epsilon_i - \epsilon_j\big)\big] \Big\} \Bigg| \Bigg] \\
= \mathbb{E}\Bigg[\sup_{\mathbf{v}\in\beta^*+\mathbb{C}(r)} \frac{1}{n!} \Bigg| \sum_{\pi\in\Pi_n} \frac{1}{[\frac{n}{2}]} \sum_{i=1}^{[\frac{n}{2}]} \Big\{ & L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)} \\
& - (\mathbf{x}_{\pi(i)} - \mathbf{x}_{\pi([\frac{n}{2}]+i)})^{\mathrm{T}}(\mathbf{v} - \beta^*)\big) \\
& - L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)}\big) \\
- \mathbb{E}\big[L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)} & - (\mathbf{x}_{\pi(i)} - \mathbf{x}_{\pi([\frac{n}{2}]+i)})^{\mathrm{T}}(\mathbf{v} - \beta^*)\big)\big] \\
+ \mathbb{E}\big[L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)}\big)\big] & \Big\} \Bigg| \Bigg] \\
\overset{\text{(i)}}{\leq} \frac{1}{n!} \sum_{\pi\in\Pi_n} \mathbb{E}\Bigg[\sup_{\mathbf{v}\in\beta^*+\mathbb{C}(r)} \Bigg| \frac{1}{[\frac{n}{2}]} \sum_{i=1}^{[\frac{n}{2}]} \Big\{ & L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)} \\
& - (\mathbf{x}_{\pi(i)} - \mathbf{x}_{\pi([\frac{n}{2}]+i)})^{\mathrm{T}}(\mathbf{v} - \beta^*)\big) \\
& - L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)}\big) \\
- \mathbb{E}\big[L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)} & - (\mathbf{x}_{\pi(i)} - \mathbf{x}_{\pi([\frac{n}{2}]+i)})^{\mathrm{T}}(\mathbf{v} - \beta^*)\big)\big] \\
+ \mathbb{E}\big[L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)}\big)\big] & \Big\} \Bigg| \Bigg] \\
\overset{\text{(ii)}}{\leq} \frac{1}{n!} \sum_{\pi\in\Pi_n} \mathbb{E}\Bigg[\sup_{\mathbf{v}\in\beta^*+\mathbb{C}(r)} \Bigg| \frac{1}{[\frac{n}{2}]} \sum_{i=1}^{[\frac{n}{2}]} \Big\{ & L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)} \\
& - (\mathbf{x}_{\pi(i)} - \mathbf{x}_{\pi([\frac{n}{2}]+i)})^{\mathrm{T}}(\mathbf{v} - \beta^*)\big) \\
& - L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)}\big) \\
- L_h\big(\epsilon'_{\pi(i)} - \epsilon'_{\pi([\frac{n}{2}]+i)} & - (\mathbf{x}_{\pi(i)} - \mathbf{x}_{\pi([\frac{n}{2}]+i)})^{\mathrm{T}}(\mathbf{v} - \beta^*)\big) \\
+ L_h\big(\epsilon'_{\pi(i)} - \epsilon'_{\pi([\frac{n}{2}]+i)}\big) & \Big\} \Bigg| \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&\overset{\text{(iii)}}{=} \frac{1}{n!}\sum_{\pi\in\Pi_n}\mathbb{E}\Bigg[\sup_{\mathbf{v}\in\beta^*+\mathbb{C}(r)} \\
&\Bigg|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i\Big\{L_h\big(\epsilon_{\pi(i)}-\epsilon_{\pi([\frac{n}{2}]+i)}-(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})^{\mathrm{T}}(\mathbf{v}-\beta^*)\big) \\
&\qquad\qquad - L_h\big(\epsilon_{\pi(i)}-\epsilon_{\pi([\frac{n}{2}]+i)}\big) \\
&- L_h\big(\epsilon'_{\pi(i)}-\epsilon'_{\pi([\frac{n}{2}]+i)}-(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})^{\mathrm{T}}(\mathbf{v}-\beta^*)\big)+L_h\big(\epsilon'_{\pi(i)}-\epsilon'_{\pi([\frac{n}{2}]+i)}\big)\Big\}\Bigg|\Bigg] \\
&\overset{\text{(iv)}}{\leq} \frac{2}{n!}\sum_{\pi\in\Pi_n}\mathbb{E}\Bigg[\sup_{\mathbf{v}\in\beta^*+\mathbb{C}(r)}\Bigg|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i\Big\{L_h\big(\epsilon_{\pi(i)}-\epsilon_{\pi([\frac{n}{2}]+i)} \\
&\qquad\qquad - (\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})^{\mathrm{T}}(\mathbf{v}-\beta^*)\big) \\
&\qquad\qquad - L_h\big(\epsilon_{\pi(i)}-\epsilon_{\pi([\frac{n}{2}]+i)}\big)\Big\}\Bigg|\Bigg] \\
&\overset{\text{(v)}}{\leq} \frac{4}{n!}\sum_{\pi\in\Pi_n}\mathbb{E}\Bigg[\sup_{\mathbf{v}\in\beta^*+\mathbb{C}(r)}\Bigg|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})^{\mathrm{T}}(\mathbf{v}-\beta^*)\Bigg|\Bigg] \\
&\leq \frac{4}{n!}\sum_{\pi\in\Pi_n}\mathbb{E}\Bigg[\sup_{\mathbf{v}\in\beta^*+\mathbb{C}(r)}\Big\|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})\Big\|_\infty\|\mathbf{v}-\beta^*\|_1\Bigg] \\
&\leq \frac{4}{n!}\sum_{\pi\in\Pi_n}\mathbb{E}\Bigg[\Big\|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})\Big\|_\infty\Bigg]\cdot 4\sqrt{sr}\sqrt{\frac{s\log p}{n}} \\
&\overset{\text{(vi)}}{\leq} 4\frac{1}{[\frac{n}{2}]}\cdot\sqrt{2\log 2p}\cdot\sqrt{[\frac{n}{2}]}\cdot 2M\cdot 4\sqrt{sr}\sqrt{\frac{s\log p}{n}} \leq \frac{128Mrs\log p}{n}, \qquad \text{(B.0.9)}
\end{aligned}
$$

where (i) is by triangle inequality, (ii) is by triangle inequality and Fubini's theorem, (iii) is by symmetry and independence, (iv) is by triangle inequality, (v) is because of comparison theorem (for instance, see Theorem 4.12 in Ledoux and Talagrand (2013)), and (vi) is by Lemma 14.14 in Bühlmann and Van De Geer (2011). Combining ((B.0.8)) and ((B.0.9)), we know for sufficiently large n, with probability at least $1-2\mathrm{e}^{-2\log p}$, $H(r)\leq 192M\frac{rs\log p}{n}$. Define $\mathcal{E}_1=\{H(r)\leq 192M\frac{rs\log p}{n}\}$ so that we have $\mathrm{P}(\mathcal{E}_1^{\mathrm{c}})\leq 2\mathrm{e}^{-2\log p}$.

Next, for any $\mathbf{v}\in\beta^*+\mathbb{C}(r)$, we derive a lower bound for $\mathbb{E}[G(\mathbf{v})]$. First note that for any $\mathbf{v}\in\beta^*+\mathbb{C}(r)$, $|(\mathbf{x}_i-\mathbf{x}_j)^{\mathrm{T}}(\mathbf{v}-\beta^*)|\leq 8Mrs\sqrt{\frac{\log p}{n}}=o(1)$. Thus for large

enough n, for any $\mathbf{v} \in \beta^* + \mathbb{C}(r)$, by Taylor's theorem, there exists $a \in [0,1]$ such that

$$
\begin{aligned}
\mathbb{E}[G(\mathbf{v})] &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{E}\big[L_h\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big) - L_h\big(\epsilon_i - \epsilon_j\big)\big] \\
&= -\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{E}\big[L_h'\big(\epsilon_i - \epsilon_j\big)\big](\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*) \\
&+ \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{E}\big[L_h''\big(\epsilon_i - \epsilon_j - a(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big)\big]\big((\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big)^2 \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{h} \int_{-\infty}^{\infty} K\Big(\frac{e - a(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)}{h}\Big) g(e) \mathrm{d}e \\
&\qquad \big((\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big)^2 \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \int_{-\infty}^{\infty} K(e) g\big(he + a(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big) \mathrm{d}e \\
&\qquad \big((\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big)^2 \\
&\geq \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \int_{-(\delta_0 \wedge \frac{\delta_1}{2h})}^{\delta_0 \wedge \frac{\delta_1}{2h}} K(e) g\big(he + a(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big) \mathrm{d}e \\
&\qquad \big((\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big)^2 \\
&\overset{\text{(i)}}{\geq} \frac{\kappa_l \mu_1 (2\delta_0 \wedge \frac{\delta_1}{h})}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \big((\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big)^2 \\
&\overset{\text{(ii)}}{=} 2\kappa_l \mu_1 (2\delta_0 \wedge \frac{\delta_1}{h})(\mathbf{v} - \beta^*)^{\mathrm{T}} \frac{\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}}{n-1} (\mathbf{v} - \beta^*) \\
&\geq 2\kappa_l \mu_1 (2\delta_0 \wedge \frac{\delta_1}{h})(\mathbf{v} - \beta^*)^{\mathrm{T}} \frac{\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}}{n} (\mathbf{v} - \beta^*) \\
&\overset{\text{(iii)}}{\geq} 2\kappa_l \mu_1 (2\delta_0 \wedge \frac{\delta_1}{h}) \rho \|\mathbf{v} - \beta^*\|_2^2 \\
&= \kappa_l \mu_1 (4\delta_0 \wedge \frac{2\delta_1}{h}) \rho r^2 \frac{s \log p}{n},
\end{aligned}
\tag{B.0.10}
$$

where (i) is by assumption 1 and 2, (ii) is by assumption 5, (iii) is by assumption 6.

On the other hand, for any $\mathbf{v} \in \beta^* + \mathbb{C}(r)$, we have

$$
\lambda_0 \big| \|\mathbf{v}\|_1 - \|\beta^*\|_1 \big| \leq \lambda_0 \|(\mathbf{v} - \beta^*)_{\mathbb{A}}\|_1 + \lambda_0 \|(\mathbf{v} - \beta^*)_{\mathbb{A}^{\mathrm{c}}}\|_1 \leq 4\lambda_0 \|(\mathbf{v} - \beta^*)_{\mathbb{A}}\|_1
$$

$$\leq 4\lambda_0\sqrt{s}\|(\mathbf{v}-\beta^*)_{\mathbb{A}}\|_2 \leq 4\lambda_0\sqrt{sr}\sqrt{\frac{s\log p}{n}} = 4c_0 sr\frac{\log p}{n}. \tag{B.0.11}$$

Thus, combining ((B.0.10)) and ((B.0.11)), under $\mathcal{E}_1$, we have for any $\mathbf{v} \in \beta^* + \mathbb{C}(r)$,

$$\begin{aligned}
&F(\mathbf{v}) + \lambda_0\|\mathbf{v}\|_1 - F(\beta^*) - \lambda_0\|\beta^*\|_1 \geq G(\mathbf{v}) - \lambda_0\big|\|\mathbf{v}\|_1 - \|\beta^*\|_1\big| \\
&\geq \mathbb{E}[G(\mathbf{v})] - H(r) - \lambda_0\big|\|\mathbf{v}\|_1 - \|\beta^*\|_1\big| \\
&\geq \mathbb{E}[G(\mathbf{v})] - 192M\frac{rs\log p}{n} - 4c_0 sr\frac{\log p}{n} \\
&\geq \Big(\kappa_l\mu_1(4\delta_0 \wedge \frac{2\delta_1}{h})\rho r - 192M - 4c_0\Big)\frac{rs\log p}{n}.
\end{aligned}$$

Denote $\mu_2 \coloneqq \kappa_l\mu_1(2\delta_0 \wedge \frac{\delta_1}{h})$. Now, choose $r = \frac{192M+4c_0}{\mu_2\rho}$, we have that under $\mathcal{E}_1$,

$$\inf_{\mathbf{v}\in\beta^*+\mathbb{C}(r)} F(\mathbf{v}) + \lambda_0\|\mathbf{v}\|_1 > F(\beta^*) + \lambda_0\|\beta^*\|_1. \tag{B.0.12}$$

Recall that under $\mathcal{E}$, $\tilde{\beta}^{\lambda_0} \in \beta^* + \mathcal{S}_{\mathbb{A}}$. We next claim that under $\mathcal{E}\cap\mathcal{E}_1$, $\|\tilde{\beta}^{\lambda_0} - \beta^*\|_2 \leq r\sqrt{\frac{s\log p}{n}}$. In fact, if $\|\tilde{\beta}^{\lambda_0} - \beta^*\|_2 > r\sqrt{\frac{s\log p}{n}}$, let $t_0 \coloneqq \frac{r\sqrt{s\log p/n}}{\|\tilde{\beta}^{\lambda_0}-\beta^*\|_2}$, then $0 < t_0 < 1$. Further define $\tilde{\beta}' \coloneqq t_0\tilde{\beta}^{\lambda_0} + (1-t_0)\beta^*$, then we have $\|\tilde{\beta}' - \beta^*\|_2 = r\sqrt{\frac{s\log p}{n}}$. Moreover, since $\tilde{\beta}^{\lambda_0} - \beta^* \in \mathcal{S}_{\mathbb{A}}$ under $\mathcal{E}$ and $\mathcal{S}_{\mathbb{A}}$ is a cone, we know $\tilde{\beta}' - \beta^* = t_0(\tilde{\beta}^{\lambda_0} - \beta^*) \in \mathcal{S}_{\mathbb{A}}$. This means that under $\mathcal{E}$, $\tilde{\beta}' \in \beta^* + \mathbb{C}(r)$. By convexity of $F(\cdot)$ and $\|\cdot\|_1$ and by ((B.0.12)), we further have

$$\begin{aligned}
&t_0\Big(F(\tilde{\beta}^{\lambda_0}) + \lambda_0\|\tilde{\beta}^{\lambda_0}\|_1\Big) + (1-t_0)\Big(F(\beta^*) + \lambda_0\|\beta^*\|_1\Big) \geq F(\tilde{\beta}') + \lambda_0\|\tilde{\beta}'\|_1 \\
&\geq \inf_{\mathbf{v}\in\beta^*+\mathbb{C}(r)} F(\mathbf{v}) + \lambda_0\|\mathbf{v}\|_1 > F(\beta^*) + \lambda_0\|\beta^*\|_1
\end{aligned}$$

under $\mathcal{E}\cap\mathcal{E}_1$. The above inequality implies $F(\tilde{\beta}^{\lambda_0}) + \lambda_0\|\tilde{\beta}^{\lambda_0}\|_1 > F(\beta^*) + \lambda_0\|\beta^*\|_1$, which is a contradiction with the definition of $\tilde{\beta}^{\lambda_0}$. So the claim is proved. By union bound and previous results, we have $\mathrm{P}\big((\mathcal{E}\cap\mathcal{E}_1)^{\mathrm{c}}\big) \leq \mathrm{P}(\mathcal{E}^{\mathrm{c}}) + \mathrm{P}(\mathcal{E}_1^{\mathrm{c}}) \leq 2p\exp\Big\{-\frac{\lambda_0^2 n}{128M^2}\Big\} + 2\mathrm{e}^{-2\log p} = 2p^{-\left(\frac{c_0^2}{128M^2}-1\right)} + 2p^{-2}$. By the claim and the above bound, the proof of Theorem 1 is finished. □

B.0.4 Proof of Theorem 5

Proof B.4 (Proof of Theorem 5)

For notational simplicity, let $F(\mathbf{v}) = \frac{1}{sn(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_h\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v} - \beta_{\mathbb{A}}^*)\big)$ for any $\mathbf{v} \in \mathbb{R}^s$. Also, define $\mathbb{B}(\Delta) = \{\mathbf{v} \in \mathbb{R}^s : \|\mathbf{v} - \beta_{\mathbb{A}}^*\|_2 = \Delta\sqrt{\frac{s}{n}}\}$ for any $\Delta > 0$. By convexity of $F(\cdot)$, it suffices to prove that $\lim_{\Delta \to \infty} \mathrm{P}\Big(\inf_{\mathbf{v} \in \mathbb{B}(\Delta)} F(\mathbf{v}) > F(\beta_{\mathbb{A}}^*)\Big) = 1$.

Let $G(\mathbf{v}) = F(\mathbf{v}) - F(\beta_{\mathbb{A}}^*)$, and let $H(\Delta) = \sup_{\mathbf{v} \in \mathbb{B}(\Delta)} \big|G(\mathbf{v}) - \mathbb{E}[G(\mathbf{v})]\big|$. First, by Lemma 1 and assumption 5, for any $\mathbf{v} \in \mathbb{B}(\Delta)$, we have

$$\begin{aligned} &\frac{1}{s}\big|L_h\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v} - \beta_{\mathbb{A}}^*)\big) - L_h\big(\epsilon_i - \epsilon_j\big)\big| \leq \frac{1}{s}\big|(\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v} - \beta_{\mathbb{A}}^*)\big| \\ &\leq \frac{1}{s} \cdot 2M\sqrt{s} \cdot \Delta\sqrt{\frac{s}{n}} = \frac{2M\Delta}{\sqrt{n}}. \end{aligned} \tag{B.0.13}$$

Consequently, when viewing $H(\Delta)$ as a function of $(\epsilon_1, \ldots, \epsilon_n)$, changing the value of a particular ϵ_i will lead to a difference in the value of $H(\Delta)$ whose absolute value is no more than $\frac{1}{n} \cdot 2^3 \cdot \frac{2M\Delta}{\sqrt{n}} = \frac{16M\Delta}{n\sqrt{n}}$. So by McDiarmid's inequality, $\mathrm{P}\big(\big|H(\Delta) - \mathbb{E}[H(\Delta)]\big| > t\big) \leq 2\mathrm{e}^{-\frac{2n^2t^2}{256M^2\Delta^2}}$. Taking $t = T\frac{\Delta}{n}$ with sufficiently large constant $T > 0$, the right hand side of this inequality can be made arbitrarily small. This concludes that $|H(\Delta) - \mathbb{E}[H(\Delta)]| = O_p\left(\frac{\Delta}{n}\right)$. We next derive an upper bound for $\mathbb{E}[H(\Delta)]$. Similar as ((B.0.9)), by ((B.0.2)), triangle inequality and comparison theorem we have

$$\begin{aligned} \mathbb{E}[H(\Delta)] = \mathbb{E}\Bigg[&\sup_{\mathbf{v} \in \mathbb{B}(\Delta)} \\ &\Bigg|\frac{1}{sn(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \Big\{ L_h\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v} - \beta_{\mathbb{A}}^*)\big) - L_h\big(\epsilon_i - \epsilon_j\big) \\ &\qquad - \mathbb{E}\big[L_h\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v} - \beta_{\mathbb{A}}^*)\big)\big] + \mathbb{E}\big[L_h\big(\epsilon_i - \epsilon_j\big)\big] \Big\}\Bigg|\Bigg] \\ \leq \frac{2}{sn!} \sum_{\pi \in \Pi_n} \mathbb{E}\Bigg[&\sup_{\mathbf{v} \in \mathbb{B}(\Delta)} \\ &\Bigg|\frac{1}{[\frac{n}{2}]} \sum_{i=1}^{[\frac{n}{2}]} \sigma_i \Big\{ L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)} - (\mathbf{x}_{\pi(i)} - \mathbf{x}_{\pi([\frac{n}{2}]+i)})_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v} - \beta_{\mathbb{A}}^*)\big) \\ &\qquad - L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)}\big) \Big\}\Bigg|\Bigg] \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{4}{sn!}\sum_{\pi\in\Pi_n}\mathbb{E}\Bigg[\sup_{\mathbf{v}\in\mathbb{B}(\Delta)}\Bigg|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\Bigg|\Bigg]\\
&\leq \frac{4}{sn!}\sum_{\pi\in\Pi_n}\mathbb{E}\Bigg[\Bigg\|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})_{\mathbb{A}}\Bigg\|_2\Bigg]\cdot\Delta\sqrt{\frac{s}{n}}\\
&\leq \frac{4}{sn!}\sum_{\pi\in\Pi_n}\sqrt{\mathbb{E}\Bigg[\Bigg\|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})_{\mathbb{A}}\Bigg\|_2^2\Bigg]}\cdot\Delta\sqrt{\frac{s}{n}}\\
&\leq \frac{4}{s}\Delta\sqrt{\frac{s}{n}}\cdot\frac{1}{\sqrt{[\frac{n}{2}]}}\cdot 2M\sqrt{s}=\frac{8M\Delta}{\sqrt{n}}\sqrt{\frac{1}{[\frac{n}{2}]}}=O\left(\frac{\Delta}{n}\right).
\end{aligned}
$$

Combining this and previous result we get $H(\Delta)=O_p(\frac{\Delta}{n})$. Let $\eta>0$ be an arbitrary number, then it follows that there exists a constant $C_1>0$ such that $\mathrm{P}\Big(H(\Delta)>C_1\frac{\Delta}{n}\Big)<\eta,\forall n$. Next, for any $\mathbf{v}\in\mathbb{B}(\Delta)$, we derive a lower bound for $\mathbb{E}[G(\mathbf{v})]$. First note that for any $\mathbf{v}\in\mathbb{B}(\Delta)$, $|(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)|\leq\frac{2M\Delta s}{\sqrt{n}}=o(1)$. Thus similar as before, for large enough n, for any $\mathbf{v}\in\mathbb{B}(\Delta)$, by Taylor's theorem, there exists $a\in[0,1]$ such that

$$
\begin{aligned}
&\mathbb{E}[G(\mathbf{v})]=\frac{1}{sn(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\big[L_h\big(\epsilon_i-\epsilon_j-(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)-L_h\big(\epsilon_i-\epsilon_j\big)\big]\\
&\geq\frac{1}{sn(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\int_{-(\delta_0\wedge\frac{\delta_1}{2h})}^{\delta_0\wedge\frac{\delta_1}{2h}}K(e)g\big(he+a(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)\mathrm{d}e\\
&\big((\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)^2\\
&\geq\frac{\kappa_l\mu_1(2\delta_0\wedge\frac{\delta_1}{h})}{sn(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\big((\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)^2\\
&\geq\frac{2}{s}\kappa_l\mu_1(2\delta_0\wedge\frac{\delta_1}{h})\rho\|\mathbf{v}-\beta_{\mathbb{A}}^*\|_2^2=\kappa_l\mu_1(4\delta_0\wedge\frac{2\delta_1}{h})\rho\frac{\Delta^2}{n}.
\end{aligned}
$$

Combining this and $\mathrm{P}\Big(H(\Delta)>C_1\frac{\Delta}{n}\Big)<\eta$, we see as long as Δ is large enough such that $\kappa_l\mu_1(4\delta_0\wedge\frac{2\delta_1}{h})\rho\Delta^2>2C_1\Delta$, we have $\mathrm{P}\big(\inf_{\mathbf{v}\in\mathbb{B}(\Delta)}G(\mathbf{v})>0\big)\geq\mathrm{P}\big(\inf_{\mathbf{v}\in\mathbb{B}(\Delta)}G(\mathbf{v})\geq\frac{C_1\Delta}{n}\big)\geq\mathrm{P}\big(H(\Delta)\leq C_1\frac{\Delta}{n}\big)\geq1-\eta$. Since η is arbitrary, $\lim_{\Delta\to\infty}\mathrm{P}\Big(\inf_{\mathbf{v}\in\mathbb{B}(\Delta)}F(\mathbf{v})>F(\beta_{\mathbb{A}}^*)\Big)=1$, so Theorem 5 is proved. □

B.0.5 Proofs for Theorem 6

To establish the strong oracle property, we first introduce the following proposition.

Proposition 10 Choose the tuning parameters such that $\min_{j\in\mathbb{A}}|\beta_j^*| > (a+1)\lambda$. Then, the LLA algorithm in ((3.3.3)) initialized by the ℓ_1-penalized CRR estimator $\tilde{\beta}^{\lambda_0}$ converges to $\hat{\beta}^{\text{ora}}$ after two iterations with probability at least $1-p_1-p_2-p_3$, where $p_1 := \mathrm{P}(\|\tilde{\beta}^{\lambda_0}-\beta^*\|_\infty > a_0\lambda)$, $p_2 := \mathrm{P}(\|\nabla_{\mathbb{A}^c}Q_n(\hat{\beta}^{\text{ora}})\|_\infty \geq a_1\lambda)$, $p_3 := \mathrm{P}(\min_{j\in\mathbb{A}}|\hat{\beta}_j^{\text{ora}}| \leq a\lambda)$. Here $Q_n(\beta) := \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}L_h\big(y_i-y_j-(\mathbf{x}_i-\mathbf{x}_j)^{\mathrm{T}}\beta\big)$. □

The proof of Proposition 10 directly follows from Theorem 1 and Theorem 2 of Fan et al. (2014b) and thus is omitted here.

Proof B.5 (Proof of Theorem 6)
Notice that for SCAD and MCP penalty, $p'_\lambda(0) = \lambda$. Therefore, with these penalty functions and the zero vector as initial value, the first iteration of LLA algorithm gives the ℓ_1-penalized CRR estimator with tuning parameter λ. Thus, the proof for case (ii) reduces to the proof of case (i) with the tuning parameter in the ℓ_1-penalization being the same as λ.

By Proposition 10, it suffices to upper bound p_1, p_2 and p_3 and show that they all converge to zero as $n\to\infty$. First, for p_1, in case (i), we have

$$p_1 = \mathrm{P}(\|\tilde{\beta}^{\lambda_0}-\beta^*\|_\infty > a_0\lambda) \leq \mathrm{P}\Big(\|\tilde{\beta}^{\lambda_0}-\beta^*\|_2 > \frac{192M+4c_0}{\mu_2\rho}\sqrt{\frac{s\log p}{n}}\Big)$$
$$\leq 2p^{-\left(\frac{c_0^2}{128M^2}-1\right)} + 2p^{-2},$$

where the second inequality is because $\|\cdot\|_\infty \leq \|\cdot\|_2$ and our choice for tuning parameter, and the last inequality is by Theorem 1. This implies $p_1\to 0$ as $n\to\infty$. In case (ii), we have

$$p_1 = \mathrm{P}(\|\tilde{\beta}^{\lambda}-\beta^*\|_\infty > a_0\lambda) \leq \mathrm{P}\Big(\|\tilde{\beta}^{\lambda}-\beta^*\|_2 > \frac{192M+4c_1}{\mu_2\rho}\sqrt{\frac{s\log p}{n}}\Big)$$
$$\leq 2p^{-\left(\frac{c_1^2}{128M^2}-1\right)} + 2p^{-2},$$

where the second inequality is because $\|\cdot\|_\infty \leq \|\cdot\|_2$ and our choice for tuning

parameter, and the last inequality is by Theorem 1. Once again, this implies $p_1 \to 0$ as $n \to \infty$. Below, our bounds for p_2 and p_3 are unified treatments for both case (i) and case (ii).

For p_2, first by Theorem 5, we know for any $\delta > 0$, for sufficiently large $r > 0$ we have $\mathrm{P}\Big(\|\hat{\beta}^{\mathrm{ora}} - \beta^*\|_2 > r\sqrt{\frac{s}{n}}\Big) \le \delta,\ \forall n$. For notational simplicity, define event $\mathcal{E} = \{\|\hat{\beta}^{\mathrm{ora}} - \beta^*\|_2 \le r\sqrt{\frac{s}{n}}\}$, so that $\mathrm{P}(\mathcal{E}^{\mathrm{c}}) \le \delta$. Also, let $\mathbb{B}(r) = \{\mathbf{v} \in \mathbb{R}^s : \|\mathbf{v} - \beta^*_{\mathbb{A}}\|_2 \le r\sqrt{\frac{s}{n}}\}$. We have

$$
\begin{aligned}
p_2 &= \mathrm{P}\Big(\Big\|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_h'\big(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\hat{\beta}^{\mathrm{ora}}\big)(\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{A}^{\mathrm{c}}}\Big\|_\infty \ge a_1\lambda\Big) \\
&\le \mathrm{P}\Big(\Big\|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_h'\big(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\hat{\beta}^{\mathrm{ora}}\big)(\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{A}^{\mathrm{c}}}\Big\|_\infty \ge a_1\lambda, \mathcal{E}\Big) \\
&\quad + \mathrm{P}(\mathcal{E}^{\mathrm{c}}) \\
&\le \sum_{k\in\mathbb{A}^{\mathrm{c}}} \mathrm{P}\Big(\sup_{\mathbf{v}\in\mathbb{B}(r)} \\
&\quad \Big|\frac{1}{n(n-1)}\sum_{i\neq j} L_h'\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v} - \beta^*_{\mathbb{A}})\big)(x_{ik} - x_{jk})\Big| \ge a_1\lambda\Big) + \delta.
\end{aligned}
\tag{B.0.14}
$$

We provide an upper bound for each term in the above summation. We have

$$
\begin{aligned}
&\mathrm{P}\Big(\sup_{\mathbf{v}\in\mathbb{B}(r)} \Big|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_h'\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v} - \beta^*_{\mathbb{A}})\big)(x_{ik} - x_{jk})\Big| \ge a_1\lambda\Big) \\
&\le \mathrm{P}\Big(\sup_{\mathbf{v}\in\mathbb{B}(r)} \Big|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} \Big(L_h'\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v} - \beta^*_{\mathbb{A}})\big) \\
&\qquad\qquad - \mathbb{E}\big[L_h'\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v} - \beta^*_{\mathbb{A}})\big)\big]\Big)(x_{ik} - x_{jk})\Big| \ge \frac{a_1\lambda}{2}\Big) \\
&\quad + \mathrm{P}\Big(\sup_{\mathbf{v}\in\mathbb{B}(r)} \\
&\quad \Big|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}\big[L_h'\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v} - \beta^*_{\mathbb{A}})\big)\big](x_{ik} - x_{jk})\Big| \ge \frac{a_1\lambda}{2}\Big) \\
&\quad \triangleq P_1 + P_2.
\end{aligned}
\tag{B.0.15}
$$

Let $F(r) \coloneqq \sup_{\mathbf{v}\in\mathbb{B}(r)} \Big|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\Big(L_h'\big(\epsilon_i-\epsilon_j-(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)-\mathbb{E}\big[L_h'\big(\epsilon_i-\epsilon_j-(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)\big]\Big)(x_{ik}-x_{jk})\Big|$. Then for P_1, we have

$$\begin{aligned} P_1 &= \mathrm{P}\Big(F(r)\geq\frac{a_1\lambda}{2}\Big)\leq \mathrm{P}\Big(|F(r)-\mathbb{E}[F(r)]|\geq\frac{a_1\lambda}{4}\Big)+\mathrm{P}\Big(\mathbb{E}[F(r)]\geq\frac{a_1\lambda}{4}\Big)\\ &\triangleq P_{11}+P_{12}. \end{aligned} \tag{B.0.16}$$

Notice that for any $\mathbf{v}\in\mathbb{B}(r)$,

$$\begin{aligned} &\Big|\Big(L_h'\big(\epsilon_i-\epsilon_j-(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)-\mathbb{E}\big[L_h'\big(\epsilon_i-\epsilon_j-(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)\big]\Big)\\ &(x_{ik}-x_{jk})\Big|\\ &\leq 2\cdot 2M = 4M, \end{aligned}$$

since $|L_h'(\cdot)|\leq 1$. Therefore, changing the value of a particular ϵ_i will lead to a difference in the value of $F(r)$ that is no greater than $\frac{16M}{n}$. By McDiarmid's inequality,

$$P_{11}=\mathrm{P}\Big(\big|F(r)-\mathbb{E}[F(r)]\big|>\frac{a_1\lambda}{4}\Big)\leq 2\mathrm{e}^{-\frac{na_1^2\lambda^2}{2048M^2}}\leq 2\mathrm{e}^{-\frac{a_1^2c_1^2\log p}{2048M^2}}=2p^{-\frac{a_1^2c_1^2}{2048M^2}}. \tag{B.0.17}$$

To upper bound P_{12}, we first give an upper bound for $\mathbb{E}[F(r)]$. Similar as ((B.0.9)), by ((B.0.2)), triangle inequality, comparison theorem and Lemma 1, for sufficiently large n we have

$$\begin{aligned} &\mathbb{E}[F(r)]\\ &=\mathbb{E}\Big[\sup_{\mathbf{v}\in\mathbb{B}(r)}\frac{1}{n!}\Big|\sum_{\pi\in\Pi_n}\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\Big(L_h'\big(\epsilon_{\pi(i)}-\epsilon_{\pi([\frac{n}{2}]+i)}-(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)\\ &\qquad -\mathbb{E}\big[L_h'\big(\epsilon_{\pi(i)}-\epsilon_{\pi([\frac{n}{2}]+i)}-(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)\big]\Big)\\ &\qquad (x_{\pi(i)k}-x_{\pi([\frac{n}{2}]+i)k})\Big|\Big]\\ &\leq\frac{2}{n!}\sum_{\pi\in\Pi_n}\mathbb{E}\Big[\sup_{\mathbf{v}\in\mathbb{B}(r)}\Big|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i\,L_h'\big(\epsilon_{\pi(i)}-\epsilon_{\pi([\frac{n}{2}]+i)}-(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)\end{aligned}$$

$$\cdot (x_{\pi(i)k} - x_{\pi([\frac{n}{2}]+i)k})\Big|\Big]$$

$$\leq \frac{4}{n!}\cdot\frac{2}{h}\cdot\kappa_u\cdot 2M\cdot\sum_{\pi\in\Pi_n}\mathbb{E}\Big[\sup_{\mathbf{v}\in\mathbb{B}(r)}\Big|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\Big|\Big]$$

$$\leq \frac{4}{n!}\cdot\frac{2}{h}\cdot\kappa_u\cdot 2M\cdot\sum_{\pi\in\Pi_n}\sqrt{\mathbb{E}\Big[\Big\|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i(\mathbf{x}_{\pi(i)}-\mathbf{x}_{\pi([\frac{n}{2}]+i)})_{\mathbb{A}}\Big\|_2^2\Big]}\cdot r\sqrt{\frac{s}{n}}$$

$$\leq 16M\kappa_u\cdot\frac{1}{h}r\sqrt{\frac{s}{n}}\cdot\frac{1}{\sqrt{[\frac{n}{2}]}}\cdot 2M\sqrt{s}\leq\frac{64M^2\kappa_u rs}{nh}.$$

By the conditions of the theorem, we have $\frac{a_1\lambda}{4} > \frac{64M^2\kappa_u rs}{nh}$ for sufficiently large n, and consequently by the above inequality, $P_{12} = \mathrm{P}\Big(\mathbb{E}[F(r)] \geq \frac{a_1\lambda}{4}\Big) = 0$. Next, we upper bound P_2. For large enough n, for any $\mathbf{v}\in\mathbb{B}(r)$, by the mean value theorem, there exists $a\in[0,1]$ such that

$$\Big|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\big[L_h'\big(\epsilon_i-\epsilon_j-(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)\big](x_{ik}-x_{jk})\Big|$$

$$=\Big|-\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\big[L_h''\big(\epsilon_i-\epsilon_j-a(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)\big]$$

$$(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)(x_{ik}-x_{jk})\Big|$$

$$=\Big|\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{h}\int_{-\infty}^{\infty}K\Big(\frac{e-a(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)}{h}\Big)g(e)\mathrm{d}e$$

$$(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)(x_{ik}-x_{jk})\Big|$$

$$=\Big|\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\int_{-\infty}^{\infty}K(e)g\big(he+a(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)\mathrm{d}e$$

$$(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)(x_{ik}-x_{jk})\Big|$$

$$\overset{\text{(i)}}{\leq}\frac{4M\mu_0}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}|(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)|$$

$$\leq 4M\mu_0\cdot 2M\sqrt{s}\cdot r\sqrt{\frac{s}{n}}=8M^2\mu_0 r\frac{s}{\sqrt{n}}=o(\lambda),$$

where (i) is by assumption 2.

So for sufficiently large n, $\sup_{\mathbf{v}\in\mathbb{B}(r)}\left|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\big[L_h'\big(\epsilon_i-\epsilon_j-(\mathbf{x}_i-\mathbf{x}_j)_{\mathbb{A}}^{\mathrm{T}}(\mathbf{v}-\beta_{\mathbb{A}}^*)\big)\big](x_{ik}-x_{jk})\right| < \frac{a_1\lambda}{2}$, and consequently, $P_2 = 0$. Combining this result with ((B.0.14)), ((B.0.15)), ((B.0.16)), ((B.0.17)) and $P_{12} = 0$, we have for sufficiently large n, $p_2 \leq 2(p-s)p^{-\frac{a_1^2c_1^2}{2048M^2}} + \delta \leq 2p^{-(\frac{a_1^2c_1^2}{2048M^2}-1)} + \delta$. Since $a_1c_1 > 32\sqrt{2}M$, we have $\frac{a_1^2c_1^2}{2048M^2} - 1 > 0$. So the previous inequality implies that $\limsup_{n\to\infty} p_2 \leq \delta$. By the arbitrariness of δ, we have $p_2 \to 0$ as $n \to \infty$.

Next, to upper bound p_3, we first have $p_3 = \mathrm{P}(\min_{j\in\mathbb{A}}|\hat{\beta}_j^{\mathrm{ora}}| \leq a\lambda) \leq \mathrm{P}(\max_{j\in\mathbb{A}}|\hat{\beta}_j^{\mathrm{ora}} - \beta_j^*| > \lambda) \leq \mathrm{P}(\|\hat{\beta}_{\mathbb{A}}^{\mathrm{ora}} - \beta_{\mathbb{A}}^*\|_2 > \lambda)$. By Theorem 5, we already have $\lim_{T\to\infty}\limsup_{n\to\infty}\mathrm{P}(\|\hat{\beta}_{\mathbb{A}}^{\mathrm{ora}} - \beta_{\mathbb{A}}^*\|_2 > T\sqrt{\frac{s}{n}}) = 0$. Since $\sqrt{\frac{s}{n}} = o(\lambda)$, $\limsup_{n\to\infty} p_3 \leq \lim_{T\to\infty}\limsup_{n\to\infty}\mathrm{P}\Big(\|\hat{\beta}_{\mathbb{A}}^{\mathrm{ora}} - \beta_{\mathbb{A}}^*\|_2 > T\sqrt{\frac{s}{n}}\Big) = 0$. Thus $p_3 \to 0$ as $n \to \infty$. The proof of Theorem 6 is finished. □

B.0.6 Proofs for Theorem 7

For any $\mathbf{v} \in \mathbb{R}^p$, denote

$$\begin{aligned} Q_n(\mathbf{v}) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_h\big(y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\mathbf{v}\big) \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_h\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big). \end{aligned}$$

Also, define $B(r) \coloneqq \big\{\mathbf{v} \in \mathbb{R}^p : \|\mathbf{v} - \beta^*\|_2 \leq r,\ \exists\mathbb{S} \subset \{1,\ldots,p\} \text{ s.t. } \mathbf{v}_{\mathbb{S}^{\mathrm{c}}} = \mathbf{0}, \mathbb{A} \subset \mathbb{S}, |\mathbb{S}| \leq 2K_n\big\}$ for any $r > 0$, and define $e(r) \coloneqq \sup_{\mathbf{v}\in B(r)}|Q_n(\mathbf{v}) - Q_n(\beta^*) - \mathbb{E}[Q_n(\mathbf{v})] + \mathbb{E}[Q_n(\beta^*)]|$. We first introduce the following lemma.

Lemma 6 For any $t, r > 0$, with probability at least $1 - 2\mathrm{e}^{-\frac{nt^2}{256M^2r^2K_n}}$, we have $|e(r) - \mathbb{E}[e(r)]| \leq t$ and $e(r) \leq \frac{32Mr\sqrt{K_n\log 2p}}{\sqrt{n}} + t$. □

Proof B.6 (Proof of Lemma 6)
First, for any $\mathbf{v} = (v_1,\ldots,v_p)^{\mathrm{T}} \in B(r)$, let $\mathbb{S}$ be the index set such that $\mathbb{A} \subset \mathbb{S}, |\mathbb{S}| \leq$

$2K_n, \mathbf{v}_{\mathbb{S}^c} = \mathbf{0}$. Then by Lemma 1 and assumption 5, we have

$$\begin{aligned}&\left|L_h\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{v} - \beta^*)\big) - L_h\big(\epsilon_i - \epsilon_j\big)\right| \\ &\leq \left|(\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{S}}^{\mathrm{T}}(\mathbf{v} - \beta^*)_{\mathbb{S}}\right| \leq 2M\sqrt{|\mathbb{S}|} \cdot r \leq 2Mr\sqrt{2K_n}.\end{aligned}$$

Consequently, when viewing $e(r)$ as a function of $(\epsilon_1, \ldots, \epsilon_n)$, changing the value of a particular ϵ_i will lead to a difference in the value of $e(r)$ whose absolute value is no more than $\frac{1}{n} \cdot 2^3 \cdot 2M\sqrt{2K_n} \cdot r = \frac{16M\sqrt{2K_n}r}{n}$. So by McDiarmid's inequality, for any $t > 0$,

$$\mathrm{P}\big(\big|e(r) - \mathbb{E}[e(r)]\big| > t\big) \leq 2\mathrm{e}^{-\frac{nt^2}{256M^2r^2K_n}}. \tag{B.0.18}$$

We next derive an upper bound for $\mathbb{E}[e(r)]$. Similar as ((B.0.9)), by ((B.0.2)), triangle inequality and contraction principle, for large enough n we have

$$\begin{aligned}\mathbb{E}[e(r)] &\leq \frac{2}{n!}\sum_{\pi\in\Pi_n}\mathbb{E}\bigg[\sup_{\mathbf{v}\in B(r)}\bigg|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i\Big\{L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)} \\ &\qquad - (\mathbf{x}_{\pi(i)} - \mathbf{x}_{\pi([\frac{n}{2}]+i)})^{\mathrm{T}}(\mathbf{v} - \beta^*)\big) \\ &\qquad - L_h\big(\epsilon_{\pi(i)} - \epsilon_{\pi([\frac{n}{2}]+i)}\big)\Big\}\bigg|\bigg] \\ &\leq \frac{4}{n!}\sum_{\pi\in\Pi_n}\mathbb{E}\bigg[\sup_{\mathbf{v}\in B(r)}\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i(\mathbf{x}_{\pi(i)} - \mathbf{x}_{\pi([\frac{n}{2}]+i)})^{\mathrm{T}}(\mathbf{v} - \beta^*)\bigg] \\ &\leq \sup_{\mathbf{v}\in B(r)}\|\mathbf{v} - \beta^*\|_1 \cdot \frac{4}{n!}\sum_{\pi\in\Pi_n}\mathbb{E}\bigg[\Big\|\frac{1}{[\frac{n}{2}]}\sum_{i=1}^{[\frac{n}{2}]}\sigma_i(\mathbf{x}_{\pi(i)} - \mathbf{x}_{\pi([\frac{n}{2}]+i)})\Big\|_\infty\bigg] \\ &\leq 4\sqrt{2K_n}r \cdot \frac{1}{\sqrt{[\frac{n}{2}]}} \cdot \sqrt{2\log 2p} \cdot 2M \leq \frac{32Mr\sqrt{K_n\log 2p}}{\sqrt{n}}.\end{aligned} \tag{B.0.19}$$

Combining ((B.0.18)) and ((B.0.19)), the proof of Lemma 6 is finished. □

We now turn to the proof of Theorem 7. Recall that $\Lambda = \{\lambda > 0 : |M_\lambda| \leq K_n\}$. Define $\Lambda_- := \{\lambda \in \Lambda : \mathbb{A} \not\subset M_\lambda\}$, which corresponds to underfitted models, and define $\Lambda_+ := \{\lambda \in \Lambda : \mathbb{A} \subset M_\lambda, \mathbb{A} \neq M_\lambda\}$, which corresponds to overfitted models. Meanwhile, for any index set $\mathbb{S} \in \big\{S \subset \{1, 2, \ldots, p\} : \mathbb{A} \subset S, |S| \leq 2K_n\big\}$ and any $r > 0$, let $P(\mathbb{S}, r) := \{\mathbf{v} \in \mathbb{R}^p : \|\mathbf{v} - \beta^*\|_2 = r, \mathbf{v}_{\mathbb{S}^c} = \mathbf{0}\}$.

Proof B.7 (Proof of Theorem 7)
By Theorem 6, there exists a correct tuning parameter $\lambda = \lambda_n$ such that $\mathrm{P}(\hat{\beta}^{\lambda_n} = \hat{\beta}^{\mathrm{ora}}) \to 1$ as $n \to \infty$. It suffices to show the following two formulae,

$$\mathrm{P}(\inf_{\lambda \in \Lambda_+} [\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\lambda_n)] > 0) \to 1, \tag{B.0.20a}$$

$$\mathrm{P}(\inf_{\lambda \in \Lambda_-} [\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\lambda_n)] > 0) \to 1. \tag{B.0.20b}$$

For any set $\mathbb{S} \in \{S \subset \{1, \ldots, p\} : |S| \le K_n\}$, define $\hat{\beta}^{\mathbb{S}} \coloneqq \arg\min_{\beta \in \mathbb{R}^p : \beta_{\mathbb{S}^{\mathrm{c}}} = \mathbf{0}} Q_n(\beta)$, and define $\hat{Q}_n^{\mathbb{S}} \coloneqq Q_n(\hat{\beta}^{\mathbb{S}})$.

First, we prove ((B.0.20a)). Note that for $\lambda \in \Lambda_+$, $\mathbb{A} \subset M_\lambda$ and thus $\hat{Q}_n^{M_\lambda} \le \hat{Q}_n^{\mathbb{A}} \le Q_n(\beta^*)$. Meanwhile, by definition, $\hat{Q}_n^{M_\lambda} \le Q_n(\hat{\beta}^{\lambda})$. Thus we have $\log\left(\frac{Q_n(\hat{\beta}^{\lambda})}{Q_n(\hat{\beta}^{\mathrm{ora}})}\right) \ge \log\left(\frac{\hat{Q}_n^{M_\lambda}}{\hat{Q}_n^{\mathbb{A}}}\right) = -\log\left(\frac{\hat{Q}_n^{\mathbb{A}}}{\hat{Q}_n^{M_\lambda}}\right) \ge -\frac{\hat{Q}_n^{\mathbb{A}} - \hat{Q}_n^{M_\lambda}}{\hat{Q}_n^{M_\lambda}}$ by the inequality $\log(1+x) \le x, \forall x \ge 0$. Then we have

$$\begin{aligned}
&\mathrm{P}(\inf_{\lambda \in \Lambda_+} [\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\lambda_n)] > 0) \\
&= \mathrm{P}(\inf_{\lambda \in \Lambda_+} [\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\lambda_n)] > 0, \hat{\beta}^{\lambda_n} = \hat{\beta}^{\mathrm{ora}}) \\
&+ \mathrm{P}(\inf_{\lambda \in \Lambda_+} [\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\lambda_n)] > 0, \hat{\beta}^{\lambda_n} \ne \hat{\beta}^{\mathrm{ora}}) \\
&\ge \mathrm{P}(\inf_{\lambda \in \Lambda_+} [\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\lambda_n)] > 0, \hat{\beta}^{\lambda_n} = \hat{\beta}^{\mathrm{ora}}) - \mathrm{P}(\hat{\beta}^{\lambda_n} \ne \hat{\beta}^{\mathrm{ora}}) \\
&= \mathrm{P}(\inf_{\lambda \in \Lambda_+} [\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\lambda_n)] > 0, \hat{\beta}^{\lambda_n} = \hat{\beta}^{\mathrm{ora}}) + o(1) \\
&= \mathrm{P}\Big(\inf_{\lambda \in \Lambda_+} \log\Big(\frac{Q_n(\hat{\beta}^{\lambda})}{Q_n(\hat{\beta}^{\mathrm{ora}})}\Big) + (|M_\lambda| - s)\frac{C_n \log p}{n} > 0, \hat{\beta}^{\lambda_n} = \hat{\beta}^{\mathrm{ora}}\Big) + o(1) \\
&\ge \mathrm{P}\Big(\inf_{\lambda \in \Lambda_+} \log\Big(\frac{Q_n(\hat{\beta}^{\lambda})}{Q_n(\hat{\beta}^{\mathrm{ora}})}\Big) + (|M_\lambda| - s)\frac{C_n \log p}{n} > 0\Big) - \mathrm{P}(\hat{\beta}^{\lambda_n} \ne \hat{\beta}^{\mathrm{ora}}) + o(1) \\
&\ge \mathrm{P}\Big(\inf_{\lambda \in \Lambda_+} \frac{C_n \log p}{n} - \frac{(\hat{Q}_n^{\mathbb{A}} - \hat{Q}_n^{M_\lambda})/(|M_\lambda| - s)}{\hat{Q}_n^{M_\lambda}} > 0\Big) + o(1).
\end{aligned} \tag{B.0.21}$$

Consider any $\lambda \in \Lambda_+$. For $\hat{Q}_n^{M_\lambda}$, we derive a lower bound. Let $r^* = C_1\sqrt{\frac{K_n \log 2p}{n}}$ and $t^* = C_2 r^* \sqrt{\frac{K_n}{n}}$, where C_1, C_2 are positive constants to be chosen. Consider the event $\mathcal{E} \coloneqq \big\{e(r^*) \le \frac{32 M r^* \sqrt{K_n \log 2p}}{\sqrt{n}} + t^*\big\}$. Then plugging $r = r^*$ and $t = t^*$ into Lemma 6, we obtain $\mathrm{P}(\mathcal{E}^{\mathrm{c}}) \le 2\mathrm{e}^{-\frac{nt^{*2}}{256 M^2 r^{*2} K_n}} = 2\mathrm{e}^{-\frac{C_2^2}{256 M^2}}$. Now for arbitrary $\delta > 0$, we choose

$C_2 \coloneqq 16M\sqrt{\log(\frac{2}{\delta})}$, so that the previous inequality gives $\mathrm{P}(\mathcal{E}^{\mathrm{c}}) \le \delta$.

For any $\mathbf{v} \in \mathbb{R}^p$, denote $F(\mathbf{v}) \coloneqq Q_n(\mathbf{v}) - Q_n(\beta^*)$. Under $\mathcal{E}$, for any $\beta \in P(M_\lambda, r^*)$, we have

$$\begin{aligned} F(\beta) &\ge \mathbb{E}[F(\beta)] - \sup_{\mathbf{v}\in P(M_\lambda, r^*)} |F(\mathbf{v}) - \mathbb{E}[F(\mathbf{v})]| \ge \mathbb{E}[F(\beta)] - e(r^*) \\ &\ge \mathbb{E}[F(\beta)] - \frac{32Mr^*\sqrt{K_n \log 2p}}{\sqrt{n}} - C_2 r^* \sqrt{\frac{K_n}{n}}. \end{aligned} \tag{B.0.22}$$

Meanwhile, for any $\beta \in P(M_\lambda, r^*)$, we have $|(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}_{M_\lambda}(\beta - \beta^*)_{M_\lambda}| \le 2M\sqrt{|M_\lambda|}r^* = O(\frac{K_n\sqrt{\log 2p}}{\sqrt{n}}) = o(1)$. Similar as before, by Taylor's theorem and the conditions of Theorem 7, for large enough n, for any $\beta \in P(M_\lambda, r^*)$, there exists $a \in [0,1]$ s.t.

$$\begin{aligned} &\mathbb{E}[F(\beta)] \\ &= \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j\ne i} \int_{-\infty}^{\infty} K(e) g\big(he + a(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}_{M_\lambda}(\beta - \beta^*)_{M_\lambda}\big) \mathrm{d}e \\ &\big((\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}_{M_\lambda}(\beta - \beta^*)_{M_\lambda}\big)^2 \\ &\ge \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j\ne i} \int_{-(\delta_0 \wedge \frac{\delta_1}{2h})}^{\delta_0 \wedge \frac{\delta_1}{2h}} K(e) g\big(he + a(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}_{M_\lambda}(\beta - \beta^*)_{M_\lambda}\big) \mathrm{d}e \\ &\qquad\qquad \cdot \big((\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}_{M_\lambda}(\beta - \beta^*)_{M_\lambda}\big)^2 \\ &\ge \frac{\kappa_l \mu_1 (2\delta_0 \wedge \frac{\delta_1}{h})}{n(n-1)} \sum_{i=1}^n \sum_{j\ne i} \big((\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}_{M_\lambda}(\beta - \beta^*)_{M_\lambda}\big)^2 \\ &\ge 2\kappa_l \mu_1 (2\delta_0 \wedge \frac{\delta_1}{h}) \phi \|(\beta - \beta^*)_{M_\lambda}\|_2^2 = 2\mu_2 \phi r^{*2}. \end{aligned} \tag{B.0.23}$$

Combining ((B.0.22)) and ((B.0.23)) we have, under $\mathcal{E}$, for any $\beta \in P(M_\lambda, r^*)$,

$$\begin{aligned} F(\beta) &\ge 2\mu_2 \phi r^{*2} - \frac{32Mr^*\sqrt{K_n \log 2p}}{\sqrt{n}} - C_2 r^* \sqrt{\frac{K_n}{n}} \\ &= (2\mu_2 \phi C_1 \sqrt{\frac{K_n \log 2p}{n}} - \frac{32M\sqrt{K_n \log 2p}}{\sqrt{n}} - C_2 \sqrt{\frac{K_n}{n}}) r^*. \end{aligned} \tag{B.0.24}$$

We now choose $C_1 \coloneqq \frac{32M + C_2}{\mu_2 \phi}$, so that $2\mu_2 \phi C_1 \sqrt{\frac{K_n \log 2p}{n}} - \frac{32M\sqrt{K_n \log 2p}}{\sqrt{n}} - C_2\sqrt{\frac{K_n}{n}} > 0$. By ((B.0.24)), this implies that under $\mathcal{E}$, $\inf_{\beta \in P(M_\lambda, r^*)} F(\beta) > 0$. Since $F(\cdot)$ is convex

and $F(\beta^*) = 0$, this further implies that under $\mathcal{E}$, $\|\hat{\beta}^{M_\lambda} - \beta^*\|_2 \le r^* = C_1\sqrt{\frac{K_n \log 2p}{n}}$. On the other hand, for any β such that $\|\beta - \beta^*\|_2 \le r^* = C_1\sqrt{\frac{K_n \log 2p}{n}}$, $\beta_{\mathbb{S}^c} = \mathbf{0}$ with some $\mathbb{S}$ satisfying $\mathbb{A} \subset \mathbb{S}$ and $|\mathbb{S}| \le K_n$, similarly as before, by Taylor's theorem, there exists $a \in [0,1]$ such that for large enough n,

$$
\begin{aligned}
&|\mathbb{E}[F(\beta)]| \\
&= \Big|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\int_{-\infty}^{\infty} K(e)g\big(he + a(\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{S}}^{\mathrm{T}}(\beta - \beta^*)_{\mathbb{S}}\big)\mathrm{d}e \\
&\big((\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{S}}^{\mathrm{T}}(\beta - \beta^*)_{\mathbb{S}}\big)^2\Big| \\
&\overset{\text{(i)}}{\le} \frac{\mu_0}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\big((\mathbf{x}_i - \mathbf{x}_j)_{\mathbb{S}}^{\mathrm{T}}(\beta - \beta^*)_{\mathbb{S}}\big)^2 \\
&\overset{\text{(ii)}}{=} 2\mu_0(\beta - \beta^*)_{\mathbb{S}}^{\mathrm{T}}\frac{\sum_{i=1}^{n}\mathbf{x}_{i,\mathbb{S}}\mathbf{x}_{i,\mathbb{S}}^{\mathrm{T}}}{n-1}(\beta - \beta^*)_{\mathbb{S}} \\
&\le 4\mu_0(\beta - \beta^*)_{\mathbb{S}}^{\mathrm{T}}\frac{\sum_{i=1}^{n}\mathbf{x}_{i,\mathbb{S}}\mathbf{x}_{i,\mathbb{S}}^{\mathrm{T}}}{n}(\beta - \beta^*)_{\mathbb{S}} \le 4\mu_0\frac{1}{n}\Big\|\sum_{i=1}^{n}\mathbf{x}_{i,\mathbb{S}}\mathbf{x}_{i,\mathbb{S}}^{\mathrm{T}}\Big\| \cdot \|(\beta - \beta^*)_{\mathbb{S}}\|_2^2 \\
&\le 4\mu_0(\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_{i,\mathbb{S}}\|_2^2) \cdot \|(\beta - \beta^*)_{\mathbb{S}}\|_2^2 \le 4\mu_0 M^2 K_n r^{*2} = 4\mu_0 M^2 C_1^2\frac{K_n^2 \log 2p}{n},
\end{aligned}
$$

where (i) is by assumption 2 and (ii) is by assumption 5. This further implies that under $\mathcal{E}$, for any β such that $\|\beta - \beta^*\|_2 \le r^* = C_1\sqrt{\frac{K_n \log 2p}{n}}$, $\beta_{\mathbb{S}^c} = \mathbf{0}$ with some $\mathbb{S}$ satisfying $\mathbb{A} \subset \mathbb{S}$ and $|\mathbb{S}| \le K_n$,

$$
\begin{aligned}
&\big|F(\beta)\big| \le \big|\mathbb{E}[F(\beta)]\big| + e(r^*) \le 4\mu_0 M^2 C_1^2\frac{K_n^2 \log 2p}{n} + \frac{32Mr^*\sqrt{K_n \log 2p}}{\sqrt{n}} + t^* \\
&= 4\mu_0 M^2 C_1^2\frac{K_n^2 \log 2p}{n} + \frac{32MC_1 K_n \log 2p}{n} + \frac{C_1 C_2 K_n\sqrt{\log 2p}}{n} \le C_4\frac{K_n^2 \log 2p}{n}
\end{aligned} \tag{B.0.25}
$$

where $C_4 > 0$ is some large constant. So under $\mathcal{E}$, by ((B.0.25)) and $\|\hat{\beta}^{M_\lambda} - \beta^*\|_2 \le r^*$,

$$
\sup_{\lambda\in\Lambda_+}|\hat{Q}_n^{M_\lambda} - Q_n(\beta^*)| \le C_4\frac{K_n^2 \log 2p}{n}. \tag{B.0.26}
$$

Since $\mathrm{P}(\mathcal{E}^{\mathrm{c}}) \leq \delta$ and δ is arbitrary positive number, this means

$$\sup_{\lambda \in \Lambda_+} |\hat{Q}_n^{M_\lambda} - Q_n(\beta^*)| = O_p\left(\frac{K_n^2 \log 2p}{n}\right) = o_p(1) \tag{B.0.27}$$

where the last equality is by the conditions of the theorem.

Next, we derive a lower bound for $Q_n(\beta^*)$. We first establish some useful inequalities regarding $L_h(\cdot)$. By ((C.1.1)), we have $L_h(u) = u\int_{-u}^{u} \frac{1}{h}K(\frac{v}{h})\mathrm{d}v - 2\int_{-\infty}^{u} \frac{v}{h}K(\frac{v}{h})\mathrm{d}v = u\int_{-\frac{u}{h}}^{\frac{u}{h}} K(v)\mathrm{d}v - 2h\int_{-\infty}^{\frac{u}{h}} vK(v)\mathrm{d}v = |u|\int_{-\frac{|u|}{h}}^{\frac{|u|}{h}} K(v)\mathrm{d}v - 2h\int_{-\infty}^{\frac{u}{h}} vK(v)\mathrm{d}v$. Let

$$L_{h,1}(u) = |u|\int_{-\frac{|u|}{h}}^{\frac{|u|}{h}} K(v)\mathrm{d}v$$

and

$$L_{h,2}(u) = -2h\int_{-\infty}^{\frac{u}{h}} vK(v)\mathrm{d}v.$$

We have $|L_{h,2}(u)| \leq 2h\int_{-\infty}^{\infty} |v|K(v)\mathrm{d}v = 2h\kappa_1$, and it is also straightforward to see that $0 \leq L_{h,1}(u) \leq |u|, \forall u \in \mathbb{R}$.

So by triangle inequality, we have

$$\begin{aligned}
&\Big|Q_n(\beta^*) - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}|\varsigma_{ij}|\Big| \\
&\leq \Big|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}(L_{h,1}(\varsigma_{ij}) - |\varsigma_{ij}|)\Big| + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}|L_{h,2}(\varsigma_{ij})| \\
&\leq \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\big(|\varsigma_{ij}| - L_{h,1}(\varsigma_{ij})\big) + 2h\kappa_1 \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}|\varsigma_{ij}|\int_{|v|\geq\frac{|\varsigma_{ij}|}{h}} K(v)\mathrm{d}v + 2h\kappa_1 \\
&\leq \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\int_{|v|\geq\frac{|\varsigma_{ij}|}{h}} h|v|K(v)\mathrm{d}v + 2h\kappa_1 \\
&\leq \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\int_{-\infty}^{\infty} h|v|K(v)\mathrm{d}v + 2h\kappa_1 \\
&= 3h\kappa_1.
\end{aligned} \tag{B.0.28}$$

Meanwhile, by ((C.1.1)), it can be directly calculated that $L_h(0) = h\kappa_1 > 0$. And because $L_h'(0) = 0$ and $L_h''(t) \geq 0, \forall t \in \mathbb{R}$, we have the trivial inequality $L_h(t) \geq L_h(0) = h\kappa_1,\ \forall t \in \mathbb{R}$.

Moreover, by the strong law of large numbers for U-statistics (Hoeffding, 1961), we know $\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} |\varsigma_{ij}| \xrightarrow{a.s.} \mathbb{E}[|\varsigma_{12}|] > 0$. So we have $\mathrm{P}\left(\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} |\varsigma_{ij}| > \frac{\mathbb{E}[|\varsigma_{12}|]}{2}\right) \to 1$. If $h\kappa_1 \leq \frac{\mathbb{E}[|\varsigma_{12}|]}{8}$, combining this result and ((B.0.28)) we obtain

$$\mathrm{P}\left(Q_n(\beta^*) > \frac{\mathbb{E}[|\varsigma_{12}|]}{8}\right) \to 1. \tag{B.0.29}$$

If however $h\kappa_1 > \frac{\mathbb{E}[|\varsigma_{12}|]}{8}$, using $L_h(t) \geq h\kappa_1$ we still get ((B.0.29)). So ((B.0.29)) always holds true.

((B.0.27)) and ((B.0.29)) together imply $\mathrm{P}\left(\inf_{\lambda \in \Lambda_+} \hat{Q}_n^{M_\lambda} > \frac{\mathbb{E}[|\varsigma_{12}|]}{16}\right) \to 1$. Combining this result and ((B.0.26)) and following ((B.0.21)), we have

$$\begin{aligned}
&\liminf_{n\to\infty}\ \mathrm{P}(\inf_{\lambda\in\Lambda_+}[\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\lambda_n)] > 0) \\
&\geq \liminf_{n\to\infty}\ \mathrm{P}\Big(\frac{C_n \log p}{n} - \sup_{\lambda\in\Lambda_+} \frac{(\hat{Q}_n^{\mathbb{A}} - \hat{Q}_n^{M_\lambda})/(|M_\lambda| - s)}{\hat{Q}_n^{M_\lambda}} > 0, \mathcal{E}\Big) - \delta \\
&\geq \liminf_{n\to\infty}\ \mathrm{P}\Big(\frac{C_n \log p}{n} - \sup_{\lambda\in\Lambda_+} \frac{\big(Q_n(\beta^*) - \hat{Q}_n^{M_\lambda}\big)/(|M_\lambda| - s)}{\hat{Q}_n^{M_\lambda}} > 0, \mathcal{E}\Big) - \delta \\
&\geq \liminf_{n\to\infty}\ \mathrm{P}\Big(\frac{C_n \log p}{n} - \frac{C_4 K_n^2 \log 2p/n}{\mathbb{E}[|\varsigma_{12}|]/16} > 0, \mathcal{E}\Big) - \delta \\
&\geq \liminf_{n\to\infty}\ \mathrm{P}\Big(\frac{C_n \log p}{n} - \frac{C_4 K_n^2 \log 2p/n}{\mathbb{E}[|\varsigma_{12}|]/16} > 0\Big) - 2\delta \\
&= 1 - 2\delta.
\end{aligned}$$

By the arbitrariness of δ, ((B.0.20a)) is proved.

Next, we prove ((B.0.20b)). First, similar to ((B.0.21)), we have

$$\begin{aligned}
&\mathrm{P}(\inf_{\lambda\in\Lambda_-}[\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\lambda_n)] > 0) \\
&\geq \mathrm{P}\Big(\inf_{\lambda\in\Lambda_-} \log\Big(\frac{Q_n(\hat{\beta}^\lambda)}{Q_n(\hat{\beta}^{\mathrm{ora}})}\Big) + (|M_\lambda| - s)\frac{C_n \log p}{n} > 0\Big) - \mathrm{P}(\hat{\beta}^{\lambda_n} \neq \hat{\beta}^{\mathrm{ora}}) + o(1)
\end{aligned}$$

$$\geq \mathrm{P}\Big(\inf_{\lambda\in\Lambda_-} \log\Big(\frac{\hat{Q}_n^{M_\lambda}}{Q_n(\beta^*)}\Big) + (|M_\lambda| - s)\frac{C_n \log p}{n} > 0\Big) + o(1). \tag{B.0.30}$$

Consider any $\lambda \in \Lambda_-$. Then $\|\hat{\beta}^{M_\lambda} - \beta^*\|_2 \geq \|\beta^*_{\mathbb{A}}\|_{\min}$.

Let $\mathcal{T} \coloneqq \{\beta \in \mathbb{R}^p : \|\beta - \beta^*\|_2 \geq \|\beta^*_{\mathbb{A}}\|_{\min}, \beta_{\mathbb{S}^{\mathrm{c}}} = \mathbf{0}$ with some $\mathbb{S}$ s.t. $\mathbb{A} \not\subset \mathbb{S}, |\mathbb{S}| \leq K_n\}$. Let $\tilde{r} = \|\beta^*_{\mathbb{A}}\|_{\min} \wedge \frac{\delta_1}{4M\sqrt{2K_n}}$. For any $\beta \in \mathcal{T}$ with its corresponding $\mathbb{S}$, define $\bar{\beta} \coloneqq a\beta + (1-a)\beta^*$ where $a = \frac{\tilde{r}}{\|\beta-\beta^*\|_2}$. Then $\|\bar{\beta}-\beta^*\|_2 = \tilde{r}$, $\bar{\beta}_{(\mathbb{S}\cup\mathbb{A})^{\mathrm{c}}} = \mathbf{0}$, and $|\mathbb{S}\cup\mathbb{A}| \leq 2K_n$. Define event $\mathcal{E}_1 \coloneqq \big\{e(\tilde{r}) \leq \frac{32M\tilde{r}\sqrt{K_n \log 2p}}{\sqrt{n}} + \tilde{t}\big\}$, where $\tilde{t} = C_5\tilde{r}\sqrt{\frac{K_n}{n}}$. Then plugging $r = \tilde{r}$ and $t = \tilde{t}$ into Lemma 6, we obtain $\mathrm{P}(\mathcal{E}_1^{\mathrm{c}}) \leq 2\mathrm{e}^{-\frac{n\tilde{t}^2}{256M^2\tilde{r}^2K_n}} = 2\mathrm{e}^{-\frac{C_5^2}{256M^2}}$. Now for arbitrary $\delta > 0$, we choose $C_5 \coloneqq 16M\sqrt{\log(\frac{2}{\delta})}$, so that the above inequality gives $\mathrm{P}(\mathcal{E}_1^{\mathrm{c}}) \leq \delta$.

Again, for any $\mathbf{v} \in \mathbb{R}^p$, denote $F(\mathbf{v}) \coloneqq Q_n(\mathbf{v}) - Q_n(\beta^*)$. Under $\mathcal{E}_1$, for any $\beta \in \mathcal{T}$ with its corresponding $\mathbb{S}$, by convexity of $F(\cdot)$, we have $F(\bar{\beta}) \leq aF(\beta) + (1-a)F(\beta^*) = aF(\beta)$. Consequently, we have, under $\mathcal{E}_1$, for any $\beta \in \mathcal{T}$,

$$\begin{aligned} F(\beta) &\geq \frac{1}{a}F(\bar{\beta}) = \frac{\|\beta-\beta^*\|_2}{\tilde{r}}F(\bar{\beta}) \geq \frac{\|\beta-\beta^*\|_2}{\tilde{r}}\big(\mathbb{E}[F(\bar{\beta})] - e(\tilde{r})\big) \\ &\geq \frac{\|\beta-\beta^*\|_2}{\tilde{r}}\big(\mathbb{E}[F(\bar{\beta})] - \frac{32M\tilde{r}\sqrt{K_n\log 2p}}{\sqrt{n}} - C_5\tilde{r}\sqrt{\frac{K_n}{n}}\big). \end{aligned} \tag{B.0.31}$$

Similar as before, by Taylor's theorem and conditions of Theorem 7, there exists $a \in [0,1]$ such that

$$\begin{aligned} \mathbb{E}[F(\bar{\beta})] &= \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}\mathbb{E}\big[L_h\big(\epsilon_i - \epsilon_j - (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}_{\mathbb{S}\cup\mathbb{A}}(\bar{\beta}-\beta^*)_{\mathbb{S}\cup\mathbb{A}}\big) \\ &\quad - L_h\big(\epsilon_i - \epsilon_j\big)\big] \\ &\geq \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}\int_{-(\delta_0\wedge\frac{\delta_1}{2h})}^{\delta_0\wedge\frac{\delta_1}{2h}} K(e)g\big(he + a(\mathbf{x}_i-\mathbf{x}_j)^{\mathrm{T}}_{\mathbb{S}\cup\mathbb{A}}(\bar{\beta}-\beta^*)_{\mathbb{S}\cup\mathbb{A}}\big)\mathrm{d}e \\ &\qquad\qquad \cdot\big((\mathbf{x}_i-\mathbf{x}_j)^{\mathrm{T}}_{\mathbb{S}\cup\mathbb{A}}(\bar{\beta}-\beta^*)_{\mathbb{S}\cup\mathbb{A}}\big)^2 \\ &\overset{\text{(i)}}{\geq} \frac{\kappa_l\mu_1(2\delta_0\wedge\frac{\delta_1}{h})}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}\big((\mathbf{x}_i-\mathbf{x}_j)^{\mathrm{T}}_{\mathbb{S}\cup\mathbb{A}}(\bar{\beta}-\beta^*)_{\mathbb{S}\cup\mathbb{A}}\big)^2 \\ &\geq 2\kappa_l\mu_1(2\delta_0\wedge\frac{\delta_1}{h})\phi\|(\bar{\beta}-\beta^*)_{\mathbb{S}\cup\mathbb{A}}\|_2^2 = 2\mu_2\phi\tilde{r}^2, \end{aligned} \tag{B.0.32}$$

where (i) is by $|(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}_{\mathbb{S}\cup\mathbb{A}}(\bar{\beta} - \beta^*)_{\mathbb{S}\cup\mathbb{A}}| \le 2M\sqrt{|\mathbb{S}\cup\mathbb{A}|}\tilde{r} \le 2M\sqrt{2K_n}\frac{\delta_1}{4M\sqrt{2K_n}} = \frac{\delta_1}{2}$. So under $\mathcal{E}$

$$
\begin{aligned}
F(\beta) &\ge \frac{\|\beta - \beta^*\|_2}{\tilde{r}}\big(\mathbb{E}[F(\bar{\beta})] - \frac{32M\tilde{r}\sqrt{K_n \log 2p}}{\sqrt{n}} - C_5\tilde{r}\sqrt{\frac{K_n}{n}}\big) \\
&\ge \frac{\|\beta - \beta^*\|_2}{\tilde{r}}\big(2\mu_2\phi\tilde{r}^2 - \frac{32M\tilde{r}\sqrt{K_n \log 2p}}{\sqrt{n}} - C_5\tilde{r}\sqrt{\frac{K_n}{n}}\big) \\
&= \frac{\|\beta - \beta^*\|_2}{\tilde{r}}\Big(2\mu_2\phi\big(\|\beta^*_{\mathbb{A}}\|_{\min} \wedge \frac{\delta_1}{4M\sqrt{2K_n}}\big) - \frac{32M\sqrt{K_n \log 2p}}{\sqrt{n}} - C_5\sqrt{\frac{K_n}{n}}\Big)\tilde{r} \\
&= \|\beta - \beta^*\|_2\Big(2\mu_2\phi\big(\|\beta^*_{\mathbb{A}}\|_{\min} \wedge \frac{\delta_1}{4M\sqrt{2K_n}}\big) - \frac{32M\sqrt{K_n \log 2p}}{\sqrt{n}} - C_5\sqrt{\frac{K_n}{n}}\Big)
\end{aligned}
$$

for any $\beta \in \mathcal{T}$, by ((B.0.31)) and ((B.0.32)). By the conditions of Theorem 7, $\sqrt{\frac{K_n \log 2p}{n}} = o(\sqrt{\frac{C_n \log p}{n}} \wedge \frac{\delta_1}{4M\sqrt{2K_n}}) = o\big(\|\beta^*_{\mathbb{A}}\|_{\min} \wedge \frac{\delta_1}{4M\sqrt{2K_n}}\big)$, so for large enough n, we have $F(\beta) \ge \mu_2\phi\|\beta - \beta^*\|_2\big(\|\beta^*_{\mathbb{A}}\|_{\min} \wedge \frac{\delta_1}{4M\sqrt{2K_n}}\big)$. Denote $\mathbb{S}^1(\beta) = \{j \in \mathbb{A} : \beta_j = 0\}$ and $\mathbb{S}^2(\beta) = \big\{j \in \{1, \ldots, p\} : \beta_j \ne 0, \beta^*_j = 0\big\}$ for any β, and note that $|\mathbb{S}^2(\beta)| - |\mathbb{S}^1(\beta)| = \|\beta\|_0 - s$. Then the previous inequality implies $F(\beta) \ge \mu_2\phi\sqrt{|\mathbb{S}^1(\beta)|} \cdot \|\beta^*_{\mathbb{A}}\|_{\min}\big(\|\beta^*_{\mathbb{A}}\|_{\min} \wedge \frac{\delta_1}{4M\sqrt{2K_n}}\big)$. Therefore, since $\hat{\beta}^{M_\lambda} \in \mathcal{T}$ for any $\lambda \in \Lambda_-$, under $\mathcal{E}_1$, we have for large enough n,

$$
\hat{Q}_n^{M_\lambda} - Q_n(\beta^*) \ge \mu_2\phi\sqrt{|\mathbb{S}^1(\hat{\beta}^{M_\lambda})|} \cdot \|\beta^*_{\mathbb{A}}\|_{\min}\big(\|\beta^*_{\mathbb{A}}\|_{\min} \wedge \frac{\delta_1}{4M\sqrt{2K_n}}\big). \tag{B.0.33}
$$

Moreover, we have $\mathrm{P}\Big(\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\ne i}|\varsigma_{ij}| < 2\mathbb{E}[|\varsigma_{12}|]\Big) \to 1$. So by ((B.0.28)), $\mathrm{P}\big(Q_n(\beta^*) < 2\mathbb{E}[|\varsigma_{12}|] + 3h\kappa_1\big) \to 1$. This combined with ((B.0.30)) and ((B.0.33)) gives

$$
\begin{aligned}
&\liminf_{n\to\infty} \mathrm{P}(\inf_{\lambda\in\Lambda_-}[\mathrm{HBIC}(\lambda) - \mathrm{HBIC}(\lambda_n)] > 0) \\
&\ge \liminf_{n\to\infty} \mathrm{P}\Big(\inf_{\lambda\in\Lambda_-} \log\Big(\frac{\hat{Q}_n^{M_\lambda}}{Q_n(\beta^*)}\Big) + (|M_\lambda| - s)\frac{C_n \log p}{n} > 0, \mathcal{E}_1\Big) - \mathrm{P}(\mathcal{E}_1^{\mathrm{c}}) \\
&\ge \liminf_{n\to\infty} \mathrm{P}\Big(\inf_{\lambda\in\Lambda_-} \min\{\log 2, \frac{\hat{Q}_n^{M_\lambda} - Q_n(\beta^*)}{2Q_n(\beta^*)}\} + (|M_\lambda| - s)\frac{C_n \log p}{n} > 0, \mathcal{E}_1\Big) \\
&- \delta
\end{aligned}
$$

$$
\begin{aligned}
&\geq \liminf_{n\to\infty} \mathrm{P}\Big(\inf_{\lambda\in\Lambda_-} \min\{\log 2, \frac{\mu_2\phi\sqrt{|\mathbb{S}^1(\hat{\beta}^{M_\lambda})|}\cdot\|\beta^*_{\mathbb{A}}\|_{\min}\big(\|\beta^*_{\mathbb{A}}\|_{\min}\wedge\frac{\delta_1}{4M\sqrt{2K_n}}\big)}{4\mathbb{E}[|\varsigma_{12}|]+6h\kappa_1}\} \\
&\qquad + (|\mathbb{S}^2(\hat{\beta}^{M_\lambda})| - |\mathbb{S}^1(\hat{\beta}^{M_\lambda})|)\frac{C_n\log p}{n} > 0, \mathcal{E}_1\Big) - \delta \\
&\geq \liminf_{n\to\infty} \mathrm{P}\Big(\inf_{\lambda\in\Lambda_-} \min\{\log 2, \frac{\mu_2\phi\sqrt{|\mathbb{S}^1(\hat{\beta}^{M_\lambda})|}\cdot\|\beta^*_{\mathbb{A}}\|_{\min}\big(\|\beta^*_{\mathbb{A}}\|_{\min}\wedge\frac{\delta_1}{4M\sqrt{2K_n}}\big)}{4\mathbb{E}[|\varsigma_{12}|]+6h\kappa_1}\} \\
&\qquad - |\mathbb{S}^1(\hat{\beta}^{M_\lambda})|\frac{C_n\log p}{n} > 0\Big) - 2\delta \\
&= 1 - 2\delta
\end{aligned}
$$

by the conditions of the theorem and the inequality $\log(1+x) \geq \min\{\frac{x}{2}, \log 2\}, \forall x > 0$. By the arbitrariness of δ, ((B.0.20b)) is proved. So we finish the proof of Theorem 7. □

Appendix C

Proof of Chapter 4

C.1 Proof of Theorem 1

We first give some general formula regarding the loss function L_h and its derivatives. Recall $L_h(u) = \int_{-\infty}^{\infty}(1-u+v)_+\frac{1}{h}K(\frac{v}{h})\mathrm{d}v, u \in \mathbb{R}$. A direct calculation gives

$$\begin{aligned}
L_h(t) &= \int_{-\infty}^{1} \frac{1-u}{h}K(\frac{t-u}{h})\mathrm{d}u, \\
L_h'(t) &= -\int_{-\infty}^{\frac{1-t}{h}} K(u)\mathrm{d}u, \\
L_h''(t) &= \frac{1}{h}K(\frac{1-t}{h}),\ \forall t \in \mathbb{R}.
\end{aligned} \tag{C.1.1}$$

It is important to note that $|L_h'(\cdot)| \leq 1$, since $K(t) \geq 0, \forall t$ and $\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1$.

Proof C.1 (Proof of Theorem 1)
By definition of the ℓ_1-penalized CRR estimator and triangle inequality, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} L_h\big(y_i(\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \hat{\beta}_0)\big) - \frac{1}{n}\sum_{i=1}^{n} L_h\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) + \lambda_0(\|\hat{\boldsymbol{\beta}}\|_2^2 - \|\boldsymbol{\beta}^*\|_2^2) \\
&\leq \lambda(\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1) \leq \lambda(\|\boldsymbol{\beta}_{\mathbb{A}}^* - \hat{\boldsymbol{\beta}}_{\mathbb{A}}\|_1 + \|\hat{\boldsymbol{\beta}}_{\mathbb{A}}\|_1 - \|\hat{\boldsymbol{\beta}}_{\mathbb{A}}\|_1 - \|\hat{\boldsymbol{\beta}}_{\mathbb{A}^{\mathrm{c}}} - \boldsymbol{\beta}_{\mathbb{A}^{\mathrm{c}}}^*\|_1) \\
&= \lambda(\|\mathbf{u}_{\mathbb{A}}\|_1 - \|\mathbf{u}_{\mathbb{A}^{\mathrm{c}}}\|_1),
\end{aligned} \tag{C.1.2}$$

where we denote $\mathbf{u} \coloneqq \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. On the other hand, by convexity of $L_h(\cdot)$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n L_h\big(y_i(\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \hat{\beta}_0)\big) - \frac{1}{n}\sum_{i=1}^n L_h\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) + \lambda_0(\|\hat{\boldsymbol{\beta}}\|_2^2 - \|\boldsymbol{\beta}^*\|_2^2) \\
&\geq \frac{1}{n}\sum_{i=1}^n L_h'\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y_i(\hat{\beta}_0 - \beta_0^*) \\
&+ \Big(2\lambda_0\boldsymbol{\beta}^{*\mathrm{T}} + \frac{1}{n}\sum_{i=1}^n L_h'\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y_i\mathbf{x}_i^{\mathrm{T}}\Big)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\
&\geq -\Big|\frac{1}{n}\sum_{i=1}^n L_h'\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y_i\Big| \cdot |\delta| \\
&- \Big\|2\lambda_0\boldsymbol{\beta}^* + \frac{1}{n}\sum_{i=1}^n L_h'\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y_i\mathbf{x}_i\Big\|_\infty (\|\mathbf{u}_{\mathbb{A}}\|_1 + \|\mathbf{u}_{\mathbb{A}^{\mathrm{c}}}\|_1),
\end{aligned} \tag{C.1.3}$$

where $\delta \coloneqq \hat{\beta}_0 - \beta_0^*$. Define event $\mathcal{E}_1 \coloneqq \{|\frac{1}{n}\sum_{i=1}^n L_h'\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y_i| \leq \frac{\lambda}{2}\}$ and $\mathcal{E}_2 \coloneqq \{\|2\lambda_0\boldsymbol{\beta}^* + \frac{1}{n}\sum_{i=1}^n L_h'\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y_i\mathbf{x}_i\|_\infty \leq \frac{\lambda}{2}\}$. Note that $\mathbb{E}\big[L_h'\big(y(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y\big] = 0$, and $\big|L_h'\big(y(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y\big| \leq 1$. So by Hoeffding's inequality,

$$\mathrm{P}(\mathcal{E}_1^{\mathrm{c}}) = \mathrm{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n L_h'\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y_i\Big| > \frac{\lambda}{2}\Big) \leq 2\exp\Big\{-\frac{n\lambda^2}{8}\Big\}. \tag{C.1.4}$$

Meanwhile, we have $\mathbb{E}\big[L_h'\big(y(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y\mathbf{x}\big] = \mathbf{0}$ by the definition of $\boldsymbol{\beta}^*$ and optimality condition. By the choice of tuning parameters we have

$$\begin{aligned}
\mathrm{P}(\mathcal{E}_2^{\mathrm{c}}) &= \mathrm{P}\Big(\Big\|2\lambda_0\boldsymbol{\beta}^* + \frac{1}{n}\sum_{i=1}^n L_h'\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y_i\mathbf{x}_i\Big\|_\infty > \frac{\lambda}{2}\Big) \\
&\leq \mathrm{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n L_h'\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y_i\mathbf{x}_i\Big\|_\infty > \frac{\lambda}{4}\Big) \\
&\leq \sum_{j=1}^p \mathrm{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n L_h'\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big) y_i x_{ij}\Big| > \frac{\lambda}{4}\Big).
\end{aligned} \tag{C.1.5}$$

Notice that by assumption 5 and $|L_h'(\cdot)| \leq 1$,

$$\mathbb{E}\Big[\mathrm{e}^{|L_h'(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)) y_i x_{ij}|/m_0}\Big] \leq \mathbb{E}\Big[\mathrm{e}^{\frac{|x_{ij}|}{m_0}}\Big] \leq 2.$$

This implies that $\|L_h'(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*))y_i x_{ij}\|_{\psi_1} \le m_0$, $\forall i \in \{1,\dots,n\}, \forall j \in \{1,\dots,p\}$. By Theorem 1.4 in Götze et al. (2021), there exists an absolute constant $\eta_0 > 0$ such that

$$\mathrm{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n L_h'\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big)y_i x_{ij}\Big| > \frac{\lambda}{4}\Big) \le 2\mathrm{e}^{-\frac{1}{\eta_0}(\frac{\lambda^2}{16m_0^2}\wedge\frac{\lambda}{4m_0})n}.$$

So following ((C.1.5)) we have $\mathrm{P}(\mathcal{E}_2^{\mathrm{c}}) \le 2p\mathrm{e}^{-\frac{1}{\eta_0}(\frac{\lambda^2}{16m_0^2}\wedge\frac{\lambda}{4m_0})n}$.

Now, under $\mathcal{E}_1 \cap \mathcal{E}_2$, combining ((C.1.2)) and ((C.1.3)) we have

$$-\frac{\lambda}{2}(|\delta| + \|\mathbf{u}_{\mathbb{A}}\|_1 + \|\mathbf{u}_{\mathbb{A}^{\mathrm{c}}}\|_1) \le \lambda(\|\mathbf{u}_{\mathbb{A}}\|_1 - \|\mathbf{u}_{\mathbb{A}^{\mathrm{c}}}\|_1),$$

which implies $\|\mathbf{u}_{\mathbb{A}^{\mathrm{c}}}\|_1 \le 3\|\mathbf{u}_{\mathbb{A}}\|_1 + |\delta|$, or $(\delta, \mathbf{u}) \in \mathcal{S}_{\mathbb{A}}$.

Define $F(\beta_0, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n L_h\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \beta_0)\big)$ for any $(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}\times\mathbb{R}^p$. Also, define $\mathbb{C}(r) = \Big\{(w, \mathbf{w}) \in \mathcal{S}_{\mathbb{A}} : |w|^2 + \|\mathbf{w}\|_2^2 = r^2\frac{s\log p}{n}\Big\}$ for any $r > 0$. Let $G(\beta_0, \boldsymbol{\beta}) = F(\beta_0, \boldsymbol{\beta}) - F(\beta_0^*, \boldsymbol{\beta}^*)$, and let $H(r) = \sup_{(\beta_0,\boldsymbol{\beta})\in(\beta_0^*,\boldsymbol{\beta}^*)+\mathbb{C}(r)} \big|G(\beta_0, \boldsymbol{\beta}) - \mathbb{E}[G(\beta_0, \boldsymbol{\beta})]\big|$.

We give an upper bound for $\mathbb{E}[H(r)]$. Let $\sigma_1, \dots, \sigma_n$ be i.i.d. Rademacher random variables (i.e. $\mathrm{P}(\sigma_i = 1) = \mathrm{P}(\sigma_i = -1) = \frac{1}{2}$), which is independent from all the other random elements. By the symmetrization inequality (see for instance, Lemma 2.3.1 in Van Der Vaart and Wellner (1996)) and contraction inequality (see for instance, Theorem 4.12 in Ledoux and Talagrand (2013)), $|L_h'(\cdot)| \le 1$ and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}[H(r)] &\le 2\mathbb{E}\Bigg[\sup_{(\beta_0,\boldsymbol{\beta})\in(\beta_0^*,\boldsymbol{\beta}^*)+\mathbb{C}(r)}\Bigg|\frac{1}{n}\sum_{i=1}^n \sigma_i\Big\{L_h\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \beta_0)\big) \\
&\quad - L_h\big(y_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big)\Big\}\Bigg|\Bigg] \\
&\le 4\mathbb{E}\Bigg[\sup_{(\beta_0,\boldsymbol{\beta})\in(\beta_0^*,\boldsymbol{\beta}^*)+\mathbb{C}(r)}\Bigg|\frac{1}{n}\sum_{i=1}^n \sigma_i y_i\big(\mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \beta_0 - \beta_0^*\big)\Bigg|\Bigg] \\
&\le \frac{4}{n}\,\mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^n \sigma_i y_i(1, \mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}\Bigg\|_\infty\Bigg]\Big(4\sqrt{s}\cdot r\sqrt{\frac{s\log p}{n}} + 2r\sqrt{\frac{s\log p}{n}}\Big). \qquad \text{(C.1.6)}
\end{aligned}$$

By assumption 5 and definition of Orlicz norm, we know $\|\sigma_i y_i x_{ij}\|_{\psi_1} = \|x_{ij}\|_{\psi_1} \le m_0$, $\forall i \in \{1,\dots,n\}, \forall j \in \{1,\dots,p\}$. Also, it is straightforward to see $\|\sigma_i y_i\|_{\psi_1} = \frac{1}{\log 2}$.

By Proposition 2.7.1 in Vershynin (2018), there exists a constant $c_1 > 0$ such that $\mathbb{E}[e^{t\sigma_i y_i x_{ij}}] \le e^{c_1^2 t^2}$ and $\mathbb{E}[e^{t\sigma_i y_i}] \le e^{c_1^2 t^2}$ for all $|t| < \frac{1}{c_1}$, $\forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, p\}$. By Jensen's inequality, we have for any $0 < t < \frac{1}{c_1}$,

$$
\begin{aligned}
&e^{t\mathbb{E}[\max\{\max_{1\le j\le p}|\sum_{i=1}^n \sigma_i y_i x_{ij}|, |\sum_{i=1}^n \sigma_i y_i|\}]} \\
&\le \mathbb{E}[e^{t\max\{\max_{1\le j\le p}|\sum_{i=1}^n \sigma_i y_i x_{ij}|, |\sum_{i=1}^n \sigma_i y_i|\}}] \\
&\le \mathbb{E}\Big[\max_{1\le j\le p}(e^{t\sum_{i=1}^n \sigma_i y_i x_{ij}} + e^{-t\sum_{i=1}^n \sigma_i y_i x_{ij}}) + e^{t\sum_{i=1}^n \sigma_i y_i} + e^{-t\sum_{i=1}^n \sigma_i y_i}\Big] \\
&\le \sum_{j=1}^p (\prod_{i=1}^n \mathbb{E}[e^{t\sigma_i y_i x_{ij}}] + \prod_{i=1}^n \mathbb{E}[e^{-t\sigma_i y_i x_{ij}}]) + \prod_{i=1}^n \mathbb{E}[e^{t\sigma_i y_i}] + \prod_{i=1}^n \mathbb{E}[e^{-t\sigma_i y_i}] \\
&\le 2pe^{c_1^2 t^2 n} + 2e^{c_1^2 t^2 n} \le 4pe^{c_1^2 t^2 n}.
\end{aligned}
$$

Consequently, for any $0 < t < \frac{1}{c_1}$,

$$
\mathbb{E}\Big[\Big\|\sum_{i=1}^n \sigma_i y_i (1, \mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}\Big\|_\infty\Big] \le \frac{\log p + \log 4}{t} + c_1^2 t n. \tag{C.1.7}
$$

By the condition of Theorem 1, we know $\frac{\sqrt{\log p + \log 4}}{c_1\sqrt{n}} = o(1)$, so for large enough n, $\frac{\sqrt{\log p + \log 4}}{c_1\sqrt{n}} < \frac{1}{c_1}$. Thus, choosing $t = \frac{\sqrt{\log p + \log 4}}{c_1\sqrt{n}}$ in ((C.1.7)) we obtain

$$
\mathbb{E}\Big[\Big\|\sum_{i=1}^n \sigma_i y_i (1, \mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}\Big\|_\infty\Big] \le 2c_1\sqrt{(\log p + \log 4)n} \tag{C.1.8}
$$

for large enough n. Thus, combining ((C.1.6)) and ((C.1.8)) we get

$$
\begin{aligned}
\mathbb{E}[H(r)] &\le \frac{4}{n} \cdot 2c_1\sqrt{(\log p + \log 4)n} \cdot \Big(4\sqrt{s} \cdot r\sqrt{\frac{s\log p}{n}} + 2r\sqrt{\frac{s\log p}{n}}\Big) \\
&\le \frac{96c_1 r s \log p}{n}.
\end{aligned}
$$

This implies that $H(r) = O_p(\frac{rs\log p}{n})$. Define event $\mathcal{G}_T := \{H(r) \le \frac{Trs\log p}{n}\}$ for any $T > 0$, then we have $\lim_{T\to\infty} \limsup_{n\to\infty} P(\mathcal{G}_T^{\mathrm{c}}) = 0$.

Next, for any $(\beta_0, \boldsymbol{\beta}) \in (\beta_0^*, \boldsymbol{\beta}^*) + \mathbb{C}(r)$, we derive a lower bound for $\mathbb{E}[G(\beta_0, \boldsymbol{\beta})]$. For large enough n, for any $(\beta_0, \boldsymbol{\beta}) \in (\beta_0^*, \boldsymbol{\beta}^*) + \mathbb{C}(r)$, by Taylor's theorem and as-

sumption 6, there exists $a \in [0,1]$ such that

$$\begin{aligned}
&\mathbb{E}[G(\beta_0, \boldsymbol{\beta})] = \mathbb{E}\big[L_h\big(y(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \beta_0)\big)\big] - \mathbb{E}\big[L_h\big(y(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \beta_0^*)\big)\big] \\
&= \frac{1}{2}(\beta_0 - \beta_0^*, (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}}) I\big(\beta_0^* + a(\beta_0 - \beta_0^*), \boldsymbol{\beta}^* + a(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\big) \\
&(\beta_0 - \beta_0^*, (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}})^{\mathrm{T}} \\
&\geq \frac{1}{2}\rho\big((\beta_0 - \beta_0^*)^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2\big) \\
&\geq \frac{1}{2}\rho r^2 \frac{s \log p}{n}. \qquad \text{(C.1.9)}
\end{aligned}$$

On the other hand, by our choice for tuning parameters, for any $(\beta_0, \boldsymbol{\beta}) \in (\beta_0^*, \boldsymbol{\beta}^*) + \mathbb{C}(r)$ we have

$$\begin{aligned}
&\lambda\big|\|\boldsymbol{\beta}\|_1 - \|\boldsymbol{\beta}^*\|_1\big| \\
&\leq \lambda\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathbb{A}}\|_1 + \lambda\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathbb{A}^{\mathrm{c}}}\|_1 \\
&\leq 4\lambda\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathbb{A}}\|_1 + \lambda|\beta_0 - \beta_0^*| \\
&\leq 4\lambda\sqrt{s}\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathbb{A}}\|_2 + \lambda r\sqrt{\frac{s \log p}{n}} \\
&\leq 4\lambda\sqrt{s}r\sqrt{\frac{s \log p}{n}} + \lambda r\sqrt{\frac{s \log p}{n}} \\
&\leq 5c_0 sr\frac{\log p}{n}, \qquad \text{(C.1.10)}
\end{aligned}$$

and we also have, by convexity of ℓ_2 norm,

$$\begin{aligned}
&\lambda_0(\|\boldsymbol{\beta}\|_2^2 - \|\boldsymbol{\beta}^*\|_2^2) \geq 2\lambda_0\boldsymbol{\beta}^{*\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \geq -2\lambda_0\|\boldsymbol{\beta}^*\|_{\max}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \\
&\geq -\frac{\lambda}{4}(4\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathbb{A}}\|_1 + |\beta_0 - \beta_0^*|) \\
&\geq -\lambda\sqrt{s}r\sqrt{\frac{s \log p}{n}} - \frac{\lambda}{4}r\sqrt{\frac{s \log p}{n}} \\
&\geq -\frac{2c_0 sr \log p}{n}. \qquad \text{(C.1.11)}
\end{aligned}$$

Thus, combining ((C.1.9)), ((C.1.10)) and ((C.1.11)), under $\mathcal{G}_T$, we have for any

$(\beta_0, \boldsymbol{\beta}) \in (\beta_0^*, \boldsymbol{\beta}^*) + \mathbb{C}(r)$,

$$
\begin{aligned}
&F(\beta_0, \boldsymbol{\beta}) + \lambda_0\|\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 - F(\beta_0^*, \boldsymbol{\beta}^*) - \lambda_0\|\boldsymbol{\beta}^*\|_2^2 - \lambda\|\boldsymbol{\beta}^*\|_1 \\
&\geq G(\beta_0, \boldsymbol{\beta}) - \frac{7c_0 sr\log p}{n} \\
&\geq \mathbb{E}[G(\beta_0, \boldsymbol{\beta})] - H(r) - \frac{7c_0 sr\log p}{n} \\
&\geq \mathbb{E}[G(\beta_0, \boldsymbol{\beta})] - \frac{Trs\log p}{n} - 7c_0 sr\frac{\log p}{n} \\
&\geq \Big(\frac{1}{2}\rho r - T - 7c_0\Big)\frac{rs\log p}{n}.
\end{aligned}
$$

Now, choose $r = \frac{4T+28c_0}{\rho}$, we have that under $\mathcal{G}_T$,

$$
\inf_{(\beta_0,\boldsymbol{\beta})\in(\beta_0^*,\boldsymbol{\beta}^*)+\mathbb{C}(r)} F(\beta_0, \boldsymbol{\beta}) + \lambda_0\|\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 > F(\beta_0^*, \boldsymbol{\beta}^*) + \lambda_0\|\boldsymbol{\beta}^*\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_1. \tag{C.1.12}
$$

Recall that under $\mathcal{E}_1 \cap \mathcal{E}_2$, $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) \in (\beta_0, \boldsymbol{\beta}^*) + \mathcal{S}_{\mathbb{A}}$. We next claim that under $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{G}_T$, $|\hat{\beta}_0 - \beta_0^*|^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \leq r^2\frac{s\log p}{n}$. In fact, if $|\hat{\beta}_0 - \beta_0^*|^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 > r^2\frac{s\log p}{n}$, let $t_0 := \frac{r\sqrt{s\log p/n}}{\sqrt{|\hat{\beta}_0-\beta_0^*|^2+\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2^2}}$, then $0 < t_0 < 1$. Further define $(\beta_0', \boldsymbol{\beta}') := t_0(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) + (1 - t_0)(\beta_0^*, \boldsymbol{\beta}^*)$, then we have $|\beta_0' - \beta_0^*|^2 + \|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\|_2^2 = r^2\frac{s\log p}{n}$. Moreover, since $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) - (\beta_0, \boldsymbol{\beta}^*) \in \mathcal{S}_{\mathbb{A}}$ under $\mathcal{E}_1 \cap \mathcal{E}_2$ and $\mathcal{S}_{\mathbb{A}}$ is a cone, we know $(\beta_0', \boldsymbol{\beta}') - (\beta_0^*, \boldsymbol{\beta}^*) = t_0\big((\hat{\beta}_0, \hat{\boldsymbol{\beta}}) - (\beta_0^*, \boldsymbol{\beta}^*)\big) \in \mathcal{S}_{\mathbb{A}}$. This means that under $\mathcal{E}_1 \cap \mathcal{E}_2$, $(\beta_0', \boldsymbol{\beta}') \in (\beta_0^*, \boldsymbol{\beta}^*) + \mathbb{C}(r)$. By convexity of $F(\cdot)$ and norm functions and by ((C.1.12)), we further have

$$
\begin{aligned}
&t_0\Big(F(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) + \lambda_0\|\hat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1\Big) + (1 - t_0)\Big(F(\beta_0^*, \boldsymbol{\beta}^*) + \lambda_0\|\boldsymbol{\beta}^*\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_1\Big) \\
&\geq F(\beta_0', \boldsymbol{\beta}') + \lambda_0\|\boldsymbol{\beta}'\|_2^2 + \lambda\|\boldsymbol{\beta}'\|_1 \\
&\geq \inf_{(\beta_0,\boldsymbol{\beta})\in(\beta_0^*,\boldsymbol{\beta}^*)+\mathbb{C}(r)} F(\beta_0, \boldsymbol{\beta}) + \lambda_0\|\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 > F(\beta_0^*, \boldsymbol{\beta}^*) + \lambda_0\|\boldsymbol{\beta}^*\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_1
\end{aligned}
$$

under $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{G}_T$. The above inequality implies $F(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) + \lambda_0\|\hat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 > F(\beta_0^*, \boldsymbol{\beta}^*) + \lambda_0\|\boldsymbol{\beta}^*\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_1$, which is a contradiction with the definition of $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$. So the claim is proved. By union bound, previous results and choice of tuning pa-

rameters, we have

$$\begin{aligned}
&\mathrm{P}\big((\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{G}_T)^{\mathrm{c}}\big) \le \mathrm{P}(\mathcal{E}_1^{\mathrm{c}}) + \mathrm{P}(\mathcal{E}_2^{\mathrm{c}}) + \mathrm{P}(\mathcal{G}_T^{\mathrm{c}}) \\
&\le 2\exp\Big\{-\frac{n\lambda^2}{8}\Big\} + 2p\mathrm{e}^{-\frac{1}{\eta_0}(\frac{\lambda^2}{16m_0^2}\wedge\frac{\lambda}{4m_0})n} + \mathrm{P}(\mathcal{G}_T^{\mathrm{c}}) \\
&\le 2p^{-\frac{c_0^2}{8}} + 2p\mathrm{e}^{-\frac{1}{\eta_0}\frac{\lambda^2 n}{16m_0^2}} + 2p\mathrm{e}^{-\frac{1}{\eta_0}\frac{\lambda n}{4m_0}} + \mathrm{P}(\mathcal{G}_T^{\mathrm{c}}) \\
&\le 2p^{-\frac{c_0^2}{8}} + 2p^{-\left(\frac{1}{\eta_0}\frac{c_0^2}{16m_0^2}-1\right)} + 2\mathrm{e}^{-\sqrt{n\log p}\left(\frac{1}{\eta_0}\frac{c_0}{4m_0}-\sqrt{\frac{\log p}{n}}\right)} + \mathrm{P}(\mathcal{G}_T^{\mathrm{c}}).
\end{aligned}$$

Since $\frac{\log p}{n} = o(1)$, as long as c_0 is large enough (for instance $c_0 > 4\sqrt{2\eta_0}m_0$), we have

$$\lim_{T\to\infty}\limsup_{n\to\infty}\mathrm{P}\big((\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{G}_T)^{\mathrm{c}}\big) = 0.$$

Combining this result and the previous claim, the proof of Theorem 1 is finished. □

C.2 Proof of Lemma 1

It is seen that $L_h^G(v)$ is twice differentiable with

$$L_h^{G\prime\prime}(v) = \frac{1}{\sqrt{2\pi}h}\exp\left\{-\frac{(1-v)^2}{2h^2}\right\} \le \frac{1}{\sqrt{2\pi}h}. \tag{C.2.1}$$

Thus inequality ((4.4.1)) is obtained due to the mean value theorem.

We then prove inequality ((4.4.2)). The inequality is trivial when $v_1 < v_2 \le 1-h$ or $v_2 > v_1 \ge 1+h$. When $1-h < v_1 < v_2 < 1+h$, since L_h^E is twice differentiable between $1-h$ and $1+h$, we see

$$|L_h^{E\prime}(v_1) - L_h^{E\prime}(v_2)| < \sup_{v\in(1-h,1+h)}|L_h^{E\prime\prime}(v)||v_1 - v_2|,$$

and

$$\sup_{v\in(1-h,1+h)}|L_h^{E\prime\prime}(v)| = \sup_{v\in(1-h,1+h)}\left|\frac{3(h^2-(1-u)^2)}{4h^3}\right| < \frac{3}{4h}.$$

When $v_1 \le 1-h$ and $v_2 \ge 1+h$,

$$|L_h^{E\prime}(v_1) - L_h^{E\prime}(v_2)| < 1 < \frac{3}{4h}(2h) \le \frac{3}{4h}|v_1 - v_2|.$$

When $v_1 \le 1-h$ and $1-h < v_2 < 1+h$,

$$\begin{aligned}|L_h^{E\prime}(v_1) - L_h^{E\prime}(v_2)| =& \left|1 - \frac{(1-v_2+h)^2(2h-1+v_2)}{4h^3}\right| \\ <& \frac{3}{4h}|1-h-v_2| \\ \le& \frac{3}{4h}|v_1 - v_2|,\end{aligned}$$

where the second to the last inequality is due to

$$\sup_{v_2 \in (1-h,1+h)} \frac{\left|1 - \frac{(1-v_2+h)^2(2h-1+v_2)}{4h^3}\right|}{|1-h-v_2|} \le \frac{9}{16h} < \frac{3}{4h}.$$

When $1-h < v_1 < 1+h$ and $v_2 \ge 1+h$,

$$\begin{aligned}|L_h^{E\prime}(v_1) - L_h^{E\prime}(v_2)| =& \left|\frac{(1-v_1+h)^2(2h-1+v_1)}{4h^3}\right| \\ <& \frac{3}{4h}|v_1 - (1+h)| \\ \le& \frac{3}{4h}|v_1 - v_2|,\end{aligned}$$

where the second to the last inequality is due to

$$\sup_{v_2 \in (1-h,1+h)} \frac{\left|\frac{(1-v_1+h)^2(2h-1+v_1)}{4h^3}\right|}{|1-v_1+h|} \le \frac{9}{16h} < \frac{3}{4h}.$$

C.3 Iteration complexity analysis of the GCD algorithm

Notation. For a vector $\mathbf{v} = (v_1, \ldots, v_d)^{\mathrm{T}} \in \mathbb{R}^d$ and a univariate function $u(\cdot)$, we write $u(\mathbf{v}) = (u(v_1), \ldots, u(v_d))^{\mathrm{T}}$. Also, denote the subvector of $\mathbf{v}$ with its kth

component removed by $\mathbf{v}_{-k} = (v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_d)^{\mathrm{T}}$ and recover $\mathbf{v}$ from $\mathbf{v}_{-k}$ by $\mathbf{v} = [v_k, \mathbf{v}_{-k}]$. We also let ∂h be the sub-differential of a nonsmooth convex function h (see e.g., Bertsekas, 1999).

Iteration Complexity Analysis. Without loss of generality, we focus solely on the GCD algorithm for solving the weighted lasso penalized DCSVM

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} L_h(y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}) + \sum_{k=1}^{p} w_k |\beta_k|, \tag{C.3.1}$$

where $w_k \geq 0$ are the weights of the penalty. Indeed, this formulation covers all the sparsity patterns in Section 4.2.3. Also, the intercept term β_0 can be absorbed into the formulation by setting $x_{i1} = 1$ for $i = 1, \ldots, n$ and $w_1 = 0$. For ease of exposition, let us rewrite ((C.3.1)) as the following unconstrained optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} f(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + \sum_{k=1}^{p} h_k(\beta_k), \tag{C.3.2}$$

where $g(\boldsymbol{\beta}) = \sum_{i=1}^{n} L_h(y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})$ is smooth convex in $\boldsymbol{\beta} \in \mathbb{R}^p$, while $h_k(\beta_k) = w_k |\beta_k|$ is nonsmooth convex in β_k for each $k = 1, \ldots, p$. Let $h(\boldsymbol{\beta}) = \sum_{k=1}^{p} h_k(\beta_k)$. Note that $\nabla g(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i L_h'(y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}) \mathbf{x}_i$ with $\nabla_k g(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i L_h'(y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}) x_{ik}$ for $k = 1, \ldots, p$. Let $\rho_{\max} = \lambda_{\max}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = \lambda_{\max}(\mathbf{X}\mathbf{X}^{\mathrm{T}})$ and $\boldsymbol{\ell}(\boldsymbol{\beta}) = (\ell_1(\boldsymbol{\beta}), \ldots, \ell_n(\boldsymbol{\beta}))^{\mathrm{T}}$ with $\ell_i(\boldsymbol{\beta}) = L_h'(y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})$ for $i = 1, \ldots, n$. Denote by $\circ$ the Hadamard product. It follows that

$$\begin{aligned} &\|\nabla g(\boldsymbol{\beta}) - \nabla g(\boldsymbol{\beta}')\| = \|\mathbf{X}^{\mathrm{T}}[\mathbf{y} \circ (\boldsymbol{\ell}(\boldsymbol{\beta}) - \boldsymbol{\ell}(\boldsymbol{\beta}'))]\| \leq \rho_{\max}^{1/2} \|\boldsymbol{\ell}(\boldsymbol{\beta}) - \boldsymbol{\ell}(\boldsymbol{\beta}')\| \\ &\leq \rho_{\max}^{1/2} c_h \|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}')\| \leq c_h \rho_{\max} \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|, \end{aligned}$$

which implies that the gradient of $g(\cdot)$ is uniformly Lipschitz continuous with Lipschitz constant $L = c_h \rho_{\max}$. When restricted to each coordinate, we have

$$|\nabla_k g([\beta_k, \boldsymbol{\beta}_{-k}]) - \nabla_k g([\beta_k', \boldsymbol{\beta}_{-k}])| \leq c_h \|\mathbf{X}_k\|^2 |\beta_k - \beta_k'|, \; k = 1, \ldots, p,$$

which implies that the gradient of $g(\cdot)$ is coordinate-wise uniformly Lipschitz continuous with Lipschitz constants $L_k = c_h \|\mathbf{X}_k\|^2$, $k = 1, \ldots, p$.

In the GCD (cyclic coordinate descent) algorithm, let $\boldsymbol{\beta}^r$ be the update of $\boldsymbol{\beta}$ after

the rth cycle, $r \geq 0$. For ease of notation, denote

$$\mathbf{b}_k^{r+1} = (\beta_1^{r+1}, \ldots, \beta_{k-1}^{r+1}, \beta_k^r, \beta_{k+1}^r, \ldots, \beta_p^r)^{\mathrm{T}},\ k = 1, \ldots, p,$$
$$\mathbf{b}_{-k}^{r+1} = (\beta_1^{r+1}, \ldots, \beta_{k-1}^{r+1}, \beta_{k+1}^r, \ldots, \beta_p^r)^{\mathrm{T}},\ k = 1, \ldots, p.$$

Clearly, we have $\mathbf{b}_1^{r+1} = \boldsymbol{\beta}^r$ and $\mathbf{b}_{p+1}^{r+1} = \boldsymbol{\beta}^{r+1}$. Note that in the proximal gradient update,

$$\beta_k^{r+1} := \mathbf{prox}_{L_k^{-1} h_k}(\beta_k^r - L_k^{-1} \nabla_k g([\beta_k^r, \mathbf{b}_{-k}^{r+1}]))$$

is equivalent to

$$\beta_k^{r+1} := \arg\min_{\beta_k} u_k(\beta_k; [\beta_k^r, \mathbf{b}_{-k}^{r+1}]) + h_k(\beta_k),$$

where the proximity operator **prox** does the soft-thresholding (Parikh and Boyd, 2013) and

$$u_k(\beta_k; [\beta_k^r, \mathbf{b}_{-k}^{r+1}]) = g([\beta_k^r, \mathbf{b}_{-k}^{r+1}]) + \nabla_k g([\beta_k^r, \mathbf{b}_{-k}^{r+1}])(\beta_k - \beta_k^r) + \frac{L_k}{2}(\beta_k - \beta_k^r)^2$$

is a quadratic majorization function of $\hat{g}(\beta_k; \mathbf{b}_{-k}^{r+1}) := g([\beta_k, \mathbf{b}_{-k}^{r+1}])$ at β_k^r. It is easy to see that $u_k(\beta_k; [\beta_k^r, \mathbf{b}_{-k}^{r+1}])$ is strongly convex in β_k. By the optimality of β_k^{r+1}, there exists $\zeta_k^{r+1} \in \partial h_k(\beta_k^{r+1})$ such that

$$(\nabla u_k(\beta_k^{r+1}; [\beta_k^r, \mathbf{b}_{-k}^{r+1}]) + \zeta_k^{r+1})(\beta_k - \beta_k^{r+1}) \geq 0,\ \forall \beta_k. \tag{C.3.3}$$

Our analysis will be divided into three parts: the sufficient descent step, the cost-to-go estimate step, and the local error bound step. Similar techniques can be found in Luo and Tseng (1992), Luo and Tseng (1993), Zhang et al. (2013) and Hong et al. (2013).

Sufficient Descent. Consider the proximal gradient method applied to solving the following problem

$$\min_{\beta_k \in \mathbb{R}} f([\beta_k, \mathbf{b}_{-k}^{r+1}]) = g([\beta_k, \mathbf{b}_{-k}^{r+1}]) + h_k(\beta_k),$$

we have by ((C.3.3))

$$\begin{aligned}
&f(\mathbf{b}_k^{r+1}) - f(\mathbf{b}_{k+1}^{r+1}) = f([\beta_k^r, \mathbf{b}_{-k}^{r+1}]) - f([\beta_k^{r+1}, \mathbf{b}_{-k}^{r+1}]) \\
&\geq u_k(\beta_k^r; [\beta_k^r, \mathbf{b}_{-k}^{r+1}]) - u_k(\beta_k^{r+1}; [\beta_k^r, \mathbf{b}_{-k}^{r+1}]) + h_k(\beta_k^r) - h_k(\beta_k^{r+1}) \\
&= \nabla_k u_k(\beta_k^{r+1}; [\beta_k^r, \mathbf{b}_{-k}^{r+1}])(\beta_k^r - \beta_k^{r+1}) + h_k(\beta_k^r) - h_k(\beta_k^{r+1}) + \frac{L_k}{2}(\beta_k^r - \beta_k^{r+1})^2 \\
&\geq (\nabla_k u_k(\beta_k^{r+1}; [\beta_k^r, \mathbf{b}_{-k}^{r+1}]) + \zeta_k^{r+1})(\beta_k^r - \beta_k^{r+1}) + \frac{L_k}{2}(\beta_k^r - \beta_k^{r+1})^2 \\
&\geq \frac{L_k}{2}(\beta_k^r - \beta_k^{r+1})^2.
\end{aligned} \tag{C.3.4}$$

It follows that

$$f(\boldsymbol{\beta}^r) - f(\boldsymbol{\beta}^{r+1}) = \sum_{k=1}^{p} \left[f(\mathbf{b}_k^{r+1}) - f(\mathbf{b}_{k+1}^{r+1}) \right] \geq \frac{\underline{L}}{2} \|\boldsymbol{\beta}^r - \boldsymbol{\beta}^{r+1}\|^2, \tag{C.3.5}$$

where $\underline{L} = \min_{1 \leq k \leq p} L_k = c_h \min_{1 \leq k \leq p} \|\mathbf{x}_k\|^2$.

Cost-to-go Estimate. Let $\mathscr{X}^* := \{\boldsymbol{\beta}^* | f(\boldsymbol{\beta}^*) = \min_{\boldsymbol{\beta}} f(\boldsymbol{\beta})\}$ be the optimal solution set of problem ((C.3.2)). Let $\bar{\boldsymbol{\beta}}^r \in \mathscr{X}^*$ be the point in $\mathscr{X}^*$ such that $\mathrm{d}_{\mathscr{X}^*}(\boldsymbol{\beta}^r) := \min_{\boldsymbol{\beta} \in \mathscr{X}^*} \|\boldsymbol{\beta} - \boldsymbol{\beta}^r\| = \|\bar{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^r\|$. By optimality of

$$\beta_k^{r+1} = \arg\min_{\beta_k \in \mathbb{R}} u_k(\beta_k; [\beta_k^r, \mathbf{b}_{-k}^{r+1}]) + h_k(\beta_k),$$

one has

$$h(\beta_k^{r+1}) - h(\bar{\beta}_k^r) + \nabla_k g([\beta_k^r, \mathbf{b}_{-k}^{r+1}])(\beta_k^{r+1} - \bar{\beta}_k^r) \leq \frac{L_k}{2}(\bar{\beta}_k^r - \beta_k^r)^2.$$

By the mean value theorem, there exists $\lambda \in [0, 1]$ and $\boldsymbol{\xi}^r = \lambda \boldsymbol{\beta}^{r+1} + (1 - \lambda)\bar{\boldsymbol{\beta}}^r$ such that

$$g(\boldsymbol{\beta}^{r+1}) - g(\bar{\boldsymbol{\beta}}^r) = \langle \nabla g(\boldsymbol{\xi}^r), \boldsymbol{\beta}^{r+1} - \bar{\boldsymbol{\beta}}^r \rangle.$$

It follows that

$$
\begin{aligned}
f(\boldsymbol{\beta}^{r+1}) - f(\bar{\boldsymbol{\beta}}^r) &= g(\boldsymbol{\beta}^{r+1}) - g(\bar{\boldsymbol{\beta}}^r) + \sum_{k=1}^{p} \big[h_k(\beta_k^{r+1}) - h_k(\bar{\beta}_k^r)\big] \\
&= \sum_{k=1}^{p} \big[\nabla_k g(\boldsymbol{\xi}^r)(\beta_k^{r+1} - \bar{\beta}_k^r) + h_k(\beta_k^{r+1}) - h_k(\bar{\beta}_k^r)\big] \\
&= \sum_{k=1}^{p} \big[\nabla_k g([\beta_k^r, \mathbf{b}_{-k}^{r+1}])(\beta_k^{r+1} - \bar{\beta}_k^r) + h_k(\beta_k^{r+1}) - h_k(\bar{\beta}_k^r) \\
&\quad + \big(\nabla_k g(\boldsymbol{\xi}^r) - \nabla_k g([\beta_k^r, \mathbf{b}_{-k}^{r+1}])\big)(\beta_k^{r+1} - \bar{\beta}_k^r)\big] \\
&\le \sum_{k=1}^{p} \big[\frac{L_k}{2}(\bar{\beta}_k^r - \beta_k^r)^2 + \big(\nabla_k g(\boldsymbol{\xi}^r) - \nabla_k g([\beta_k^r, \mathbf{b}_{-k}^{r+1}])\big)(\beta_k^{r+1} - \bar{\beta}_k^r)\big].
\end{aligned}
$$

By the fact that $\nabla g(\cdot)$ is Lipschitz continuous, it is implied that

$$
\begin{aligned}
&\bigg(\sum_{k=1}^{p} \big(\nabla_k g(\boldsymbol{\xi}^r) - \nabla_k g([\beta_k^r, \mathbf{b}_{-k}^{r+1}])\big)(\beta_k^{r+1} - \bar{\beta}_k^r)\bigg)^2 \\
&\le \bigg(\sum_{k=1}^{p} \|\nabla g(\boldsymbol{\xi}^r) - \nabla g([\beta_k^r, \mathbf{b}_{-k}^{r+1}])\|^2\bigg)\bigg(\sum_{k=1}^{p} (\beta_k^{r+1} - \bar{\beta}_k^r)^2\bigg) \\
&\le \bigg(\sum_{k=1}^{p} L^2 \|\boldsymbol{\xi}^r - [\beta_k^r, \mathbf{b}_{-k}^{r+1}]\|^2\bigg)\|\boldsymbol{\beta}^{r+1} - \bar{\boldsymbol{\beta}}^r\|^2 \\
&= \bigg(\sum_{k=1}^{p} L^2 \|\lambda(\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r) + (1-\lambda)(\bar{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^r) + \boldsymbol{\beta}^r - [\beta_k^r, \mathbf{b}_{-k}^{r+1}]\|^2\bigg) \\
&\quad \cdot 2(\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|^2 + \|\boldsymbol{\beta}^r - \bar{\boldsymbol{\beta}}^r\|^2) \\
&\le 12(p+1)L^2\big[\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|^2 + \|\boldsymbol{\beta}^r - \bar{\boldsymbol{\beta}}^r\|^2\big]^2 \\
&\le 25pL^2\big[\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|^2 + \mathrm{d}^2_{\mathscr{X}^*}(\boldsymbol{\beta}^r)\big]^2.
\end{aligned}
$$

It follows that

$$
f(\boldsymbol{\beta}^{r+1}) - f(\bar{\boldsymbol{\beta}}^r) \le (5L\sqrt{p} + \bar{L})\big[\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|^2 + \mathrm{d}^2_{\mathscr{X}^*}(\boldsymbol{\beta}^r)\big], \tag{C.3.6}
$$

where $\bar{L} = \max_{1\le k\le p} L_k = c_h \max_{1\le k\le p} \|\mathbf{x}_k\|^2$.

Local error bound. Let $\mathbf{d}_{\mathscr{X}^*}(\boldsymbol{\beta}) \equiv \min_{\boldsymbol{\beta}^* \in \mathscr{X}^*} \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|$. Here we handle the Gaussian and Epanechnikov kernels separately. For the Gaussian kernel, that is, when $L_h(\cdot) = L_h^G(\cdot)$, according to ((C.3.4)) and ((C.3.5)), the GCD algorithm is descending along its iterations. We can thus restrict the domain of $\boldsymbol{\beta}$ to the sublevel set $\mathcal{L}_0 = \{\boldsymbol{\beta} : f(\boldsymbol{\beta}) \leq f(\mathbf{0})\}$. Let $\eta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ for $i = 1, \ldots, n$. It follows that the set $\mathcal{C}_0 = \{\boldsymbol{\eta} = (\eta_i, 1 \leq i \leq n)^{\mathrm{T}} : \boldsymbol{\beta} \in \mathcal{L}_0\}$ is convex compact. Therefore, for all $\boldsymbol{\beta} \in \mathcal{L}_0$, η_i is bounded by $\eta_{\max}$, where $\eta_{\max} = \max_{1 \leq i \leq n} \sup_{\boldsymbol{\beta} \in \mathcal{L}_0} |\eta_i| < \infty$. Note that the function $p(\mathbf{z}) = \sum_{i=1}^n L_h^G(y_i z_i)$ is strongly convex in $\mathbf{z} \in \mathcal{C}_0$ by ((C.2.1)). We can see that $g(\boldsymbol{\beta}) = p(\mathbf{X}\boldsymbol{\beta})$. It follows from Zhang et al. (2013) that for any $\xi \geq \min_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$, there exist $\kappa, \varepsilon > 0$ such that

$$\mathrm{d}_{\mathscr{X}^*}(\boldsymbol{\beta}) \leq \kappa \|\boldsymbol{\beta} - \mathbf{prox}_h(\boldsymbol{\beta} - \nabla g(\boldsymbol{\beta}))\|, \tag{C.3.7}$$

for all $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta} - \mathbf{prox}_h(\boldsymbol{\beta} - \nabla g(\boldsymbol{\beta}))\| \leq \varepsilon$ and $f(\boldsymbol{\beta}) \leq \xi$.

For the Epanechnikov kernel, that is, when $L_h(\cdot) = L_h^E(\cdot)$, one needs to add an additional ridge penalty $\mu\|\boldsymbol{\beta}\|^2$ for some small $\mu > 0$ in order to achieve strong optimality. Thus, when the Epanechnikov kernel is used, we instead consider the following problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n L_h^E(y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \sum_{k=1}^p w_k |\beta_k| + \mu\|\boldsymbol{\beta}\|^2$$

and solve it using the GCD algorithm.

As a summary, we show in the following theorem that the GCD algorithm converges at least linearly.

Theorem 2

The GCD algorithm converges at least linearly to a solution in $\mathscr{X}^*$. □

Proof C.2

We first show that there exists some $\sigma > 0$ such that

$$\|\boldsymbol{\beta}^r - \mathbf{prox}_h(\boldsymbol{\beta}^r - \nabla g(\boldsymbol{\beta}^r))\| \leq \sigma \|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|, \ \forall r \geq 1. \tag{C.3.8}$$

For any $r \geq 1$ and any $1 \leq k \leq p$, by the optimality of

$$\beta_k^{r+1} = \arg\min_{\beta_k} u_k(\beta_k; [\beta_k^r, \mathbf{b}_{-k}^{r+1}]) + h_k(\beta_k),$$

we have

$$\beta_k^{r+1} = \mathbf{prox}_{L_k^{-1} h_k}(\beta_k^{r+1} - L_k^{-1} \nabla u_k(\beta_k^{r+1}; [\beta_k^r, \mathbf{b}_{-k}^{r+1}])).$$

Let $\hat{L}_k = \max(1, L_k)$ and $\tilde{L}_k = \max(1, L_k^{-1})$. It follows from Lemma 4.3 of Kadkhodaie et al. (2014) that

$$\begin{aligned}
&|\beta_k^r - \mathbf{prox}_{h_k}(\beta_k^r - \nabla_k g(\boldsymbol{\beta}^r))| \leq \hat{L}_k |\beta_k^r - \mathbf{prox}_{L_k^{-1} h_k}(\beta_k^r - L_k^{-1} \nabla_k g(\boldsymbol{\beta}^r))| \\
&\leq \hat{L}_k \big[|\beta_k^{r+1} - \mathbf{prox}_{L_k^{-1} h_k}(\beta_k^r - L_k^{-1} \nabla_k g(\boldsymbol{\beta}^r))| + |\beta_k^{r+1} - \beta_k^r| \big] \\
&\leq \hat{L}_k \big[|\mathbf{prox}_{L_k^{-1} h_k}(\beta_k^{r+1} - L_k^{-1} \nabla u_k(\beta_k^{r+1}; [\beta_k^r, \mathbf{b}_{-k}^{r+1}])) \\
&\qquad - \mathbf{prox}_{L_k^{-1} h_k}(\beta_k^r - L_k^{-1} \nabla_k g(\boldsymbol{\beta}^r))| + |\beta_k^{r+1} - \beta_k^r| \big] \\
&\leq 2\hat{L}_k |\beta_k^{r+1} - \beta_k^r| + \hat{L}_k L_k^{-1} |\nabla u_k(\beta_k^{r+1}; [\beta_k^r, \mathbf{b}_{-k}^{r+1}]) - \nabla_k g(\boldsymbol{\beta}^r)| \\
&\leq 3\hat{L}_k |\beta_k^{r+1} - \beta_k^r| + \tilde{L}_k \|\nabla g([\beta_k^r, \mathbf{b}_{-k}^{r+1}]) - \nabla g(\boldsymbol{\beta}^r)\| \\
&\leq (3\hat{L}_k + L\tilde{L}_k) \|\boldsymbol{\beta}_k^{r+1} - \boldsymbol{\beta}_k^r\|.
\end{aligned}$$

It follows that

$$\|\boldsymbol{\beta}^r - \mathbf{prox}_h(\boldsymbol{\beta}^r - \nabla g(\boldsymbol{\beta}^r))\| \leq (3\hat{L} + L\tilde{L})\sqrt{p} \|\boldsymbol{\beta}_k^{r+1} - \boldsymbol{\beta}_k^r\|,$$

where $\hat{L} = \max(1, \bar{L})$ and $\tilde{L} = \max(1, \underline{L}^{-1})$. Therefore, when we take $\sigma = (3\hat{L} + L\tilde{L})\sqrt{p}$, we get the desired result in ((C.3.8)). Note that the sufficient descent property ((C.3.5)) implies that $\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\| \to 0$ as $r \to \infty$. It follows from ((C.3.8)) that $\|\boldsymbol{\beta}^r - \mathbf{prox}_h(\boldsymbol{\beta}^r - \nabla g(\boldsymbol{\beta}^r))\| \to 0$ as $r \to \infty$. Thus, by ((C.3.7)) we have $\mathrm{d}_{\mathscr{X}^*}(\boldsymbol{\beta}^r) \to 0$ as $r \to \infty$. Consequently, from ((C.3.6)) it implies that $f(\boldsymbol{\beta}^r) \to f^* := \min_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$, which shows that the GCD algorithm converges to the global minimum.

Now let $c_1 = \underline{L}(2B)^{-1}$, $c_2 = 5L\sqrt{p} + \bar{L}$, and $\Delta^r = f(\boldsymbol{\beta}^r) - f^*$. By the local error

bound ((C.3.7)) and the cost-to-go estimate ((C.3.6)), we obtain

$$\begin{aligned}
&\Delta^{r+1} \leq c_2\big[\mathrm{d}^2_{\mathscr{X}^*}(\boldsymbol{\beta}^r) + \|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|^2\big] \\
&\leq c_2\kappa^2\|\boldsymbol{\beta}^r - \mathbf{prox}_h(\boldsymbol{\beta}^r - \nabla g(\boldsymbol{\beta}^r))\|^2 + c_2\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|^2 \\
&\leq (c_2\kappa^2\sigma^2 + c_2)\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|^2 \\
&\leq (c_2\kappa^2\sigma^2 + c_2)c_1^{-1}[f(\boldsymbol{\beta}^r) - f(\boldsymbol{\beta}^{r+1})] \\
&= (c_2\kappa^2\sigma^2 + c_2)c_1^{-1}(\Delta^r - \Delta^{r+1}),
\end{aligned}$$

which implies that

$$\Delta^{r+1} \leq \frac{c_3}{1 + c_3}\Delta^r, \tag{C.3.9}$$

where $c_3 = (c_2\kappa^2\sigma^2 + c_2)c_1^{-1}$. We can see from ((C.3.9)) that $f(\boldsymbol{\beta}^r)$ approaches f^* with at least linear rate of convergence. From ((C.3.5)) again, this further implies that the sequence $\{\boldsymbol{\beta}^r\}$ converges at least linearly. □

C.4 Additional numeric results with Gaussian kernel

Under the same settings introduced in our simulation section, we compared the performance of lasso DCSVM and elastic-net DCSVM, using Gaussian kernel. The result is shown in Table C.1. Again, we can see that the elastic-net DCSVM outperforms lasso DCSVM. We also conducted elastic-net DCSVM with Gaussian kernel on the same real datasets that we introduced in our real data section, and compared its performance with the performance of elastic-net SVM and elastic-net logistic regression. The result is displayed in Table C.2. Overall, DCSVM still achieves the best performance.

Table C.1. *Comparison of prediction error (in percentage) and variable selection of density-convoluted SVM with Gaussian kernels using lasso and elastic-net (enet) penalties. Denote by C and IC the number of correctly and incorrectly selected variables, respectively. Under each simulation setting, the method with the lowest prediction error is marked by a black box. All the quantities are averaged over 50 independent runs and the standard errors of the prediction error are given in parentheses.*

		lasso-DCSVM				enet-DCSVM			
p	ρ	err (%)		C	IC	err (%)		C	IC
Example 1									
Example 1									
500		6.92	(0.14)	5	0	**6.84**	(0.14)	5	0
5000		7.22	(0.19)	5	0	**7.11**	(0.13)	5	0
Example 2									
500	0.2	13.96	(0.21)	5	0	**13.52**	(0.19)	5	1
	0.7	23.18	(0.26)	3	0	**22.65**	(0.25)	4	0
	0.9	24.83	(0.24)	2	0	**24.75**	(0.23)	4	0
5000	0.2	14.46	(0.23)	5	0	**13.78**	(0.18)	5	0
	0.7	23.57	(0.26)	3	0	**22.66**	(0.21)	4	0
	0.9	25.25	(0.25)	2	0	**24.70**	(0.25)	3	0
Example 3									
500	0.2	10.58	(0.21)	5	0	**10.27**	(0.15)	5	1
	0.7	19.78	(0.21)	4	0	**19.48**	(0.18)	4	0
	0.9	23.97	(0.22)	2	0	**23.49**	(0.21)	4	0
5000	0.2	10.70	(0.20)	5	0	**10.51**	(0.20)	5	0
	0.7	20.13	(0.24)	3	0	**19.70**	(0.21)	4	0
	0.9	24.34	(0.30)	2	0	**23.85**	(0.23)	4	0

Table C.2. *Comparison of prediction error (in percentage) and run time (in second) of elastic-net density-convoluted SVM with Gaussian kernel, elastic-net SVM, and elastic-net logistc regression. For each benchmark data, the method with the lowest prediction error is marked by a black box. All the quantities are averaged over 50 independent runs and the standard errors of the prediction error are given in parentheses.*

			enet-DCSVM			enet-SVM			enet-logistic		
data	n	p	err (%)		time	err (%)		time	err (%)		time
arcene	100	9920	**32.00**	(1.42)	454.36	37.09	(1.59)	8912.87	35.82	(1.65)	219.30
breast	42	22283	**24.86**	(1.79)	243.13	30.38	(2.05)	1946.98	30.76	(2.14)	227.88
colon	62	2000	**18.71**	(1.11)	91.70	18.90	(1.55)	722.48	23.87	(1.51)	27.33
leuk	72	7128	3.94	(0.51)	215.95	**3.89**	(0.51)	1863.23	4.33	(0.61)	115.00
LSVT	126	309	**15.74**	(0.62)	73.04	16.20	(0.68)	74.20	15.87	(0.68)	9.05
malaria	71	22283	**5.49**	(0.63)	818.98	7.60	(1.21)	12046.09	6.80	(0.98)	483.20
ovarian	253	15154	**0.67**	(0.13)	1491.25	4.87	(1.23)	14442.87	0.87	(0.14)	964.16
prostate	102	6033	9.69	(0.68)	199.85	**8.98**	(0.50)	2421.20	10.24	(0.61)	116.50