The Annals of Statistics
2019, Vol. 47, No. 4, 1995–2022
https://doi.org/10.1214/18-AOS1738

NONPARAMETRIC SCREENING UNDER CONDITIONAL STRICTLY CONVEX LOSS FOR ULTRAHIGH DIMENSIONAL SPARSE DATA

BY XU HAN

Temple University

Sure screening technique has been considered as a powerful tool to handle the ultrahigh dimensional variable selection problems, where the dimensionality p and the sample size n can satisfy the NP dimensionality $\log p = O(n^a)$ for some $a > 0$ [*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** (2008) 849–911]. The current paper aims to simultaneously tackle the "universality" and "effectiveness" of sure screening procedures. For the "universality," we develop a general and unified framework for nonparametric screening methods from a loss function perspective. Consider a loss function to measure the divergence of the response variable and the underlying nonparametric function of covariates. We newly propose a class of loss functions called conditional strictly convex loss, which contains, but is not limited to, negative log likelihood loss from one-parameter exponential families, exponential loss for binary classification and quantile regression loss. The sure screening property and model selection size control will be established within this class of loss functions. For the "effectiveness," we focus on a goodness-of-fit nonparametric screening (Goffins) method under conditional strictly convex loss. Interestingly, we can achieve a better convergence probability of containing the true model compared with related literature. The superior performance of our proposed method has been further demonstrated by extensive simulation studies and some real scientific data example.

1. Introduction. Ultrahigh-dimensional variable selection has become an important problem in modern statistical research due to the big data collection in a variety of scientific areas, such as genomics, bioinformatics, functional magnetic resonance imaging, high frequency finance, etc. In all of these problems, statisticians want to select the important covariates associated with the response variable from p covariates. However, the dimensionality p can grow much faster than the sample size n. More specifically, $\log p = O(n^a)$ for some $a > 0$, which is denoted as nonpolynomial order (NP) [Fan and Lv (2008)]. As Fan, Samworth and Wu (2009) has pointed out: existing variable selection methods based on penalized

Received June 2017; revised June 2018.

[1]Supported in part by the summer research funding from Fox Business School at Temple University.

MSC2010 subject classifications. 62G99.

Key words and phrases. Ultrahigh dimensional variable selection, sure screening property, goodness-of-fit nonparametric screening, conditional strictly convex loss.

pseudo likelihood estimation [e.g., Tibshirani (1996), Fan and Li (2001), Zou and Hastie (2005), Zou (2006), Candès and Tao (2007), Zou and Li (2008), Zhang (2010)] can suffer from the simultaneous challenges to computational expediency, statistical accuracy and algorithmic stability in ultrahigh dimensional problems.

To handle the challenges in the ultrahigh-dimensional problems, Fan and Lv (2008) introduced a new statistical framework, sure independence screening. Their original method focused on the Gaussian linear regression models, and the important predictors were selected via the marginal correlation ranking. Formally, let $M_\star$ be the set of true important variables, and $\widehat{M}_n$ be the selected variables based on some procedure, then

$$P(M_\star \subset \widehat{M}_n) \geq 1 - \varepsilon_n, \tag{1}$$

where $\varepsilon_n > 0$ and $\varepsilon_n \to 0$ as $n \to \infty$. This is called the "sure screening property." Furthermore, the model selection size can be controlled at a polynomial rate of sample size with probability approaching 1. Because of its powerful performance and computational convenience in ultrahigh dimensional problems, the sure screening framework has received increasing attention in the past few years. Existing literature in this framework have mainly focused on the "universality" of the screening procedures, that is, developing procedures for various scenarios which possess the "sure screening property," for example, generalized linear model by Fan and Song (2010), nonparametric additive model by Fan, Feng and Song (2011), rank based model-free feature screening by Zhu et al. (2011), Cox model by Zhao and Li (2012), robust rank correlation screening by Li et al. (2012), varying coefficient model by Fan, Ma and Dai (2014), empirical likelihood based screening by Chang, Tang and Wu (2013), quantile-adaptive screening by He, Wang and Hong (2013), censored rank independence screening by Song et al. (2014), fused Kolmogorov filter by Mai and Zou (2015). On the other hand, formal pursuit of "effectiveness" of sure screening procedures have been largely ignored. Intuitively, for the convergence probability $1 - \varepsilon_n$ in (1), if ε_n converges to 0 slower, the corresponding screening procedure will have larger possibility of not selecting the true important variables. More specifically, existing literature commonly show that

$$P(M_\star \subset \widehat{M}_n) \geq 1 - s_n\{b\exp(-cn^a)\}, \tag{2}$$

where a, b, c are positive values and s_n is the size of true model. The rate a controls how high dimensionality the screening procedure can handle. It will be illustrated in detail in later sections. For a larger a, the probability of containing true model converges to 1 faster. The effect of c can be negligible for a larger a and a sufficient large n. The constant b is not crucial in the asymptotic sense, but it is important for finite sample situations. With the same value of a, a larger value of b indicates that important variables can be mis-selected with higher probability. For some existing results, b can even grow as n increases. Therefore, the constant b and the convergence rate a can be viewed as a measure of effectiveness of a screening method.

Correspondingly, a sure screening procedure will be considered more effective with a larger a and a smaller b in the convergence probability of (2). Although the existing screening procedures have been proved to possess the "sure screening property," the established convergence of containing the true model can be slow subject to various specific model settings and conditions.

Our first goal in the current paper is to develop a general and unified framework for sure screening methods from a loss function perspective. Consider a response variable Y which distribution depends on parameter θ. Suppose θ is a function of p-dimensional covariate vector $\mathbf{X} = (X_1, \ldots, X_p)^T$. We are interested in selecting the covariates X_j's which are associated with the response variable Y through a nonparametric function $\theta = f(X_1, \ldots, X_p)$. For notational convenience, we will write as $\theta(\mathbf{X})$ to denote its dependence on the covariates $\mathbf{X}$. In later presentation, we sometimes simply write it as θ for the true function, and the readers should be reminded that the θ is a function on $\mathbf{X}$. This setting includes a variety of commonly used regression models:

EXAMPLE 1 (Gaussian regression). Assume that $Y|\mathbf{X} = \mathbf{x}$ is from $N(\theta(\mathbf{x}), \sigma^2)$ for some constant $\sigma > 0$.

EXAMPLE 2 (Logistic regression). Assume that $Y|\mathbf{X} = \mathbf{x}$ is from Bernoulli distribution and $\ln P(Y = 1|\mathbf{X} = \mathbf{x}) - \ln P(Y = 0|\mathbf{X} = \mathbf{x}) = \theta(\mathbf{x})$.

EXAMPLE 3 (Poisson regression). Assume that $Y|\mathbf{X} = \mathbf{x}$ is from Poisson distribution and $\ln E(Y|\mathbf{X} = \mathbf{x}) = \theta(\mathbf{x})$.

EXAMPLE 4 (Quantile regression). Let $Q_\alpha(Y|\mathbf{X} = \mathbf{x})$ be the αth quantile of the distribution for $Y|\mathbf{X} = \mathbf{x}$, then assume $Q_\alpha(Y|\mathbf{X} = \mathbf{x}) = \theta(\mathbf{x})$.

The above Examples 1–3 fall within the general framework of mean regression:

$$E(Y|\mathbf{X} = \mathbf{x}) = h\big(\theta(\mathbf{x})\big) = g^{-1}\big(f(x_1, \ldots, x_p)\big), \tag{3}$$

where h is some known function, f is a nonparametric function and g is called the link function. When g is the canonical link, that is, $g = (h)^{-1}$, we have $\theta(\mathbf{X}) = f(X_1, \ldots, X_p)$. However, Example 4 is different from the mean regression.

The above regression models are equivalent to considering a loss function $l(\omega, Y)$ for measuring the divergence between a generic variable ω and the response variable Y where ω is a function of $\mathbf{X}$, and assuming that the true model of θ will minimize $E[l(\omega, Y)|\mathbf{X} = \mathbf{x}]$ with respect to ω. For instance, in the above Examples 1–3, we can choose $l(\omega, Y)$ as the negative of the log-likelihood of $Y|\mathbf{X} = \mathbf{x}$; In the above Example 4, we can choose $l(\omega, Y) = (Y - \omega)[\alpha - \mathbf{I}(Y - \omega < 0)]$, where $\mathbf{I}$ is an indicator function. Therefore, we will select the important covariates X_j's associated with Y based on such a loss function.

In the current paper, we newly propose a definition of loss function called conditional strictly convex loss, which contains, but is not limited to, negative log-likelihood loss for one-parameter exponential families, exponential loss for binary classification and quantile regression loss for robust estimation. Our sure screening property is established within such a wide class of loss functions. Therefore, several existing screening methods automatically fall within our framework, including Fan, Feng and Song (2011) for nonparametric additive models and He, Wang and Hong (2013) for quantile regression, although their proposed screening procedures can be different from ours. In addition, many more screening methods are suggested by our framework, for example, generalized additive models, binary classification by exponential loss and so on.

Our second goal of the current paper is to develop screening methods under conditional strictly convex loss with better convergence probability of containing the true model. We treat the marginal regression as fitting the response variable with componentwise covariates via the loss function. We impose an additive model structure for the unknown nonparametric function approximated by B-spline basis. Interestingly, if we consider the goodness-of-fit statistics as the marginal utility to rank the importance of each covariate to the joint model, we can achieve a much better convergence probability of containing the true model compared with other related literature. Detailed comparison between our results with other related literature will be presented in Section 3. Furthermore, our selected model size can be controlled at the level of sample size n rather than the dimensionality p_n with high probability.

The major contribution of the current paper is to simultaneously tackle the issues of "universality" and "effectiveness." For the "universality," we establish the sure screening property within a unified framework through the introduction of a new class of loss functions: conditional strictly convex loss; For the "effectiveness," within this framework, we show that the goodness-of-fit nonparametric screening methods can achieve a better convergence probability of containing the true model compared with related literature.

Theoretical pursuit of "universality" and "effectiveness" for screening procedures in the current paper has shed new light on the choice of sure screening methods and greatly benefited the applications of screening methods in practice. For example, the superior performance of our proposed method compared with other existing screening procedures will be further demonstrated by extensive simulation studies and some real scientific data example. Our method is called goodness-of-fit nonparametric screening (Goffins). To stabilize the computation performance, we also provide an iterative screening procedure and an improved variant to handle the situations where covariates are possibly correlated.

The rest of this paper will be organized as follows: Section 2 introduces the conditional strictly convex loss, the B-spline approximation and the goodness of fit nonparametric screening; Section 3 establishes the exponential bound, the sure screening properties and the control of model selection size; Section 4 proposes

an iterative screening procedure and an improved variant; Section 5 provides simulation studies and real data analysis. All the technical proofs and some numerical results are relegated to the Supplementary Material [Han (2019)].

2. Nonparametric screening under convex loss.

2.1. *Conditional strictly convex loss.* Let $l(x, y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a function and assume the partial derivative $\partial l(x, y)/\partial x$ exists almost everywhere for x throughout the paper. We consider $l(\omega, Y)$ as a loss function to measure the divergence between a generic variable ω and the response variable Y. We assume the convexity of $l(\omega, Y)$ in the ω position, that is, $l(t_1\omega_1 + t_2\omega_2, Y) \geq t_1 l(\omega_1, Y) + t_2 l(\omega_2, Y)$ for any real values $t_1 + t_2 = 1$ and $t_1, t_2 > 0$. Here, ω is a function of covariates $\mathbf{X}$, and can be written as $\omega(\mathbf{X})$ to denote its dependence on $\mathbf{X}$. For notational convenience, we sometimes simply write it as ω. Suppose the distribution of Y depends on some parameter θ where θ is a nonparametric function of the covariates $\mathbf{X}$. Assume the true model of θ minimizes $E[l(\omega, Y)|\mathbf{X}]$ with respect to ω.

In the current paper, we will newly propose a definition of loss function called conditional strictly convex loss. Our sure screening method will be established within such a wide class of convex loss functions.

DEFINITION 1. If $\partial E[l(\omega, Y)|\mathbf{X}]/\partial\omega$ is continuously differentiable in ω and $\partial^2 E[l(\omega, Y)|\mathbf{X}]/\partial\omega^2 > 0$, then $l(\omega, Y)$ is called a conditional strictly convex loss function.

The conditional strictly convex loss includes, but is not limited to, the following three major types of loss functions:

Type 1: Negative log-likelihood loss for exponential families. Suppose that the random variable Y is from a one-parameter exponential family with density function

$$f_{Y|X}(y; \theta) = \exp\big(y\theta - b(\theta) + c(y)\big) \tag{4}$$

for some known functions $b(\cdot)$ and $c(\cdot)$ where $b''(\cdot)$ exists. Consider the negative log-likelihood loss:

$$l(\omega, Y) = -\big[\omega Y - b(\omega) + c(Y)\big]. \tag{5}$$

Minimization of $E[l(\omega, Y)|\mathbf{X}]$ with respect to ω and letting θ be the minimizer leads to $E[Y|\mathbf{X}] = b'(\theta)$, which naturally belongs to the mean regression (3). This is the setting of generalized additive model in Stone (1986). Note that the second derivative of $l(\omega, Y)$ with respect to ω is $b''(\omega)$, and $b''(\theta)$ is the variance of Y from the exponential families.

The loss function (5) can be better understood by some popular regression models:

EXAMPLE 1 (Gaussian regression). $b(\theta) = \theta^2/2$, $c(y) = -y^2/2$ and $l(\omega, Y) = (Y - \omega)^2/2$.

EXAMPLE 2 (Logistic regression). $b(\theta) = \ln(1 + \exp(\theta))$, $c(y) = 0$ and $l(\omega, Y) = -\omega Y + \ln(1 + \exp(\omega))$.

EXAMPLE 3 (Poisson regression). $b(\theta) = \exp(\theta)$, $c(y) = -\ln(y!)$ and $l(\omega, Y) = -Y\omega + \exp(\omega) + \ln(Y!)$.

Type 2: Exponential loss for classification. In classification problems, suppose $Y \in \{-1, 1\}$ and $P(Y = 1|\mathbf{X} = \mathbf{x}) = p(\mathbf{x})$. The goal is to construct a classifier $\theta(\mathbf{x})$. When new covariates $\mathbf{X}$ are available, predict the corresponding class type Y as 1 if $\theta(\mathbf{X}) > c$ and as -1 if $\theta(\mathbf{X}) < c$ where c is some threshold. The exponential loss is defined as

$$l(\omega, Y) = \exp(-Y\omega), \tag{6}$$

which has been considered as a smooth approximation to the misclassification loss [Freund and Schapire (1997)]. Minimization of $E[l(\omega, Y)|\mathbf{X}]$ with respect to ω and letting θ be the minimizer leads to

$$\ln \frac{P(Y = 1|\mathbf{X})}{P(Y = -1|\mathbf{X})} = 2\theta.$$

Type 3: Quantile regression loss. For many practical problems, the distribution information of response variable Y is usually not available or complicated. Instead of imposing a full distribution, quantile regression framework assumes that the αth quantile of Y given $\mathbf{X}$, $Q_\alpha(Y|\mathbf{X})$, is some function of $\mathbf{X}$, thus the distribution assumption can be substantially relaxed [Koenker (2005)]. Correspondingly, consider the loss function

$$l(\omega, Y) = (Y - \omega)\{\alpha - \mathbf{I}(Y - \omega < 0)\} \tag{7}$$

for $0 < \alpha < 1$ where $\mathbf{I}$ is an indicator function. When $\alpha = 1/2$, this is proportional to the least absolute deviation loss $|Y - \omega|$, which is popularly used for robust regression. The loss function $l(\omega, Y)$ is not differentiable in ω. This is a key difference from the aforementioned loss functions. Minimization of $E[l(\omega, Y)|\mathbf{X}]$ with respect to ω yields $Q_\alpha(Y|\mathbf{X}) = \theta$ where θ is the minimizer.

The following Proposition 2.1 shows that with mild conditions, Types 1–3 belong to the conditional strictly convex loss.

PROPOSITION 2.1. *For Type* 1, *if* b'' *is strictly positive and is a continuous function, then* (5) *belongs to the conditional strictly convex loss*; *for Type* 2, (6) *belongs to the conditional strictly convex loss*; *for Type* 3, *if the conditional distribution of* $Y|\mathbf{X}$ *has a continuous density function* $f_{Y|X}$ *and* $f_{Y|X} > 0$ *on any bounded domain, then* (7) *belongs to the conditional strictly convex loss*.

Without any further investigation, one might simply group Types 1 and 2 in Proposition 2.1 as one class since the corresponding loss functions are second differentiable in ω. However, we will show in Section 3.4 that even for Types 1 and 2, the loss functions possess some fundamental differences in the underlying structures, which raises challenges for proving the model selection size control in Section 3.4.

The name of conditional strictly convex loss is borrowed from "strictly convex function." However, there are some major differences between the two concepts. If $l(x, y)$ is a strictly convex function in x and $l'(x, y)$ is continuously differentiable in x, then l is also a conditional strictly convex loss, but a conditional strictly convex loss might not be a strictly convex function; see Type 3 quantile regression loss as such a counterexample.

A class of convex loss, Bregman divergence, can also be considered here. For a given convex function $q(\cdot)$ with derivative $q'(\cdot)$, the Bregman divergence [Brègman (1967)] is defined as

$$l(\omega, Y) = q(\omega) - q(Y) + (Y - \omega)q'(Y). \tag{8}$$

Note that $l(\omega, Y)$ is not generally a symmetric function in ω and Y. Suppose $q'(\cdot)$ is continuously differentiable and $q''(\cdot) > 0$, it is easy to show that such Bregman divergence belongs to the conditional strictly convex loss. It is impossible for us to list all the possibilities here, thus we will not go any further in this direction. It is worth mentioning that the quantile regression loss (7) does not belong to Bregman divergence. More detailed discussions about Bregman divergence are referred to Zhang, Jiang and Shang (2009).

2.2. *Goodness-of-fit nonparametric screening.* To capture the nonparametric structure of $\theta(\mathbf{X})$, a powerful model for dimensionality reduction is the additive model:

$$\theta(\mathbf{X}) = m_1(X_1) + \cdots + m_p(X_p) + \mu, \tag{9}$$

where $m_j(\cdot)$ are the square integrable functions and μ is an unknown constant. For identifiability, we assume $E[m_j(X_j)] = 0$ for $j = 1, \ldots, p$. Let $M_\star = \{j : E[m_j(X_j)]^2 > 0\}$ be the true sparse model with nonsparsity size $s_n = |M_\star|$. Suppose we have observed data $\{(\mathbf{X}_i, Y_i)\}$ for $i = 1, \ldots, n$, which are independent copies of $\{(\mathbf{X}, Y)\}$. The dimensionality p is ultrahigh and satisfies $\log p = O(n^a)$ for some $a > 0$. Based on the sample data, we aim to select a subset of covariates which contains $M_\star$ with moderate size. We allow p to grow with n, and denote the dimensionality as p_n.

In this paper, we refer to marginal regression as fitting models with componentwise covariates through the loss function $l(\omega, Y)$. We define the population version of the minimizer of the componentwise regression as

$$f_j^M(X_j) \equiv \underset{f_j \in L_2(P)}{\arg\min}\, E\big[l\big(f_j(X_j), Y\big)\big], \tag{10}$$

where P denotes the joint distribution of $(\mathbf{X}, Y)$ and $L_2(P)$ is the class of square integrable functions under measure P. We use B-spline basis to approximate the marginal nonparametric regression function. Let S_n be the space of polynomial splines of degree $l \geq 1$. Stone (1986) has shown that under some smoothness conditions, the nonparametric functions can be well approximated by functions in S_n. Correspondingly, we define

$$f_{nj}^M(X_j) \equiv \arg\min_{f_j \in S_n} E\big[l\big(f_j(X_j), Y\big)\big]. \tag{11}$$

We also define the marginal minimum divergence estimator as

$$\widehat{f}_{nj}^M(X_j) \equiv \arg\min_{f_j \in S_n} \mathbb{P}_n l\big(f_j(X_j), Y\big), \tag{12}$$

where $\mathbb{P}_n g(\mathbf{X}, Y) = n^{-1} \sum_{i=1}^n g(\mathbf{X}_i, Y_i)$ is the empirical expectation for generic function $g(\cdot)$. Let $\{\Psi_k\}_{k=1}^{d_n}$ denote a normalized B-spline basis with $\|\Psi_k\|_\infty \leq 1$, where $\|\cdot\|_\infty$ is the sup norm. For any $f_{nj} \in S_n$, we have

$$f_{nj}(x) = \sum_{k=1}^{d_n} \Psi_k(x)\beta_{jk}, \qquad 1 \leq j \leq p \tag{13}$$

for some coefficients $\{\beta_{jk}\}_{k=1}^{d_n}$. The construction of the B spline basis can be found in the well-known books, for example, de Boor (1978). Let $\boldsymbol{\Psi}_j \equiv \boldsymbol{\Psi}_j(X_j) = (\Psi_1(X_j), \ldots, \Psi_{d_n}(X_j))^T$, therefore, we can express

$$f_{nj}^M(X_j) = \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, \qquad \widehat{f}_{nj}^M(X_j) = \boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M, \tag{14}$$

where $\boldsymbol{\beta}_j^M$ and $\widehat{\boldsymbol{\beta}}_j^M$ are the d_n dimensional coefficient vector for the minimizers of (11) and (12).

We will consider a sure screening procedure based on goodness-of-fit statistics. Formally, let

$$G_{n,j} = \mathbb{P}_n\big\{l\big(\widehat{\beta}_0^M, Y\big) - l\big(\boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M, Y\big)\big\}, \qquad j = 1, \ldots, p_n,$$

where $\widehat{\beta}_0^M \equiv \arg\min_{\beta_0 \in \mathbb{R}} \mathbb{P}_n l(\beta_0, Y)$. Correspondingly, for the population level,

$$G_j^\star = E\big\{l\big(\beta_0^M, Y\big) - l\big(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y\big)\big\}, \qquad j = 1, \ldots, p_n,$$

where $\beta_0^M \equiv \arg\min_{\beta_0 \in \mathbb{R}} El(\beta_0, Y)$. The goodness-of-fit statistics compare the marginal regression model with the null model (no variables included in the model). Intuitively, if the marginal contribution of an individual variable is significant to the response variable, the goodness-of-fit measure should be relatively large. We select model by $\widehat{M}_{\nu_n} = \{1 \leq j \leq p_n : G_{n,j} \geq \nu_n\}$ for a predetermined threshold ν_n. Our screening method is called goodness-of-fit nonparametric **s**creening (Goffins). We intentionally use the letter "G" in $G_{n,j}$ and $G_j^\star$ to denote the goodness-of-fit statistics.

When l is the squared error loss, since the term $\mathbb{P}_n l(\widehat{\beta}_0^M, Y)$ in $G_{n,j}$ is not affected by the index j, Goffins is equivalent to screening based on the sum of squared residuals, that is, select the model by $\{1 \le j \le p_n : \mathbb{P}_n(Y - \boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_{j1}^M)^2 \le \mu_n\}$ for some threshold $\mu_n > 0$. Note that, $\mathbb{P}_n(Y - \boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M)^2$ can be further expressed as $\mathbb{P}_n Y^2 - \mathbb{P}_n(\boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M)^2$. Therefore, Goffins under the squared error loss is equivalent to selecting the model by $\{1 \le j \le p_n : \mathbb{P}_n(\boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M)^2 \ge \gamma_n\}$ for some threshold $\gamma_n > 0$. More generally, when l is the negative log-likelihood loss for exponential families, Goffins is equivalent to screening based on the likelihood ratio statistics. For parametric model based likelihood ratio screening; see Fan and Song (2010).

3. Sure screening properties.

3.1. *Preliminaries.* In this paper, we will show that our goodness-of-fit nonparametric screening (Goffins) has the sure screening property, and the number of the selected variables has moderate size. Let $[a, b]$ be the support of covariates X_j. The following conditions are needed:

(A) The nonparametric marginal functions $\{f_j^M\}_{j=1}^p$ belong to a class of functions $\mathfrak{F}$ whose rth derivative $f^{(r)}$ exists and is Lipschitz of order α:

$$\mathfrak{F} = \{f(\cdot) : |f^{(r)}(s) - f^{(r)}(t)| \le K|s - t|^\alpha, \text{ for } s, t \in [a, b]\} \tag{15}$$

for some positive constant K, where r is a nonnegative integer and $\alpha \in (0, 1]$ such that $d = r + \alpha > 0.5$.

(B) The marginal density functions g_j of X_j satisfies $0 < K_1 \le g_j(X_j) \le K_2 < \infty$ on $[a, b]$ for $1 \le j \le p$ for some constants K_1 and K_2.

(C) The unknown nonparametric function $\theta(\mathbf{X})$ satisfies that $\sup_{X \in \mathbb{R}^{p_n}} |\theta(\mathbf{X})| < M$ from some positive constant M.

Conditions A, B and C are standard regularity assumptions for nonparametric regression in Stone (1986), Fan, Feng and Song (2011), He, Wang and Hong (2013), etc.

The following Lemma 3.1 shows that the approximation error of marginal regression f_{nj}^M in (11) to marginal nonparametric projection f_j^M in (10) is negligible.

LEMMA 3.1. *If l is a conditional strictly convex loss, under Conditions* A–C, *assume that f_j^M is uniformly bounded for $j = 1, \ldots, p$, then there exists a positive constant C_1 such that $E(f_j^M - f_{nj}^M)^2 \le C_1 d_n^{-2d}$, where d is defined in Condition* A.

To show that for $j \in M_\star$, $G_j^\star$ has a nonvanishing signal, we also need the following conditions:

(D) $\min_{j\in M_\star} E[f_j^M(X_j) - Ef_j^M(X_j)]^2 \geq c_1 d_n n^{-2\kappa}$, for some $0 < \kappa < d/(2d+1)$ and $c_1 > 0$.

(E) $d_n^{-2} \leq c_1(1-\xi)^2 n^{-2\kappa}/4C_1$ for some $\xi \in (0,1)$.

Condition D requires that the marginal nonparametric projections are at a certain strength level separate from the noise. Therefore, we can select the significant covariates based on a threshold. Similar conditions also appear in related literature on nonparametric screening, for example, Fan, Feng and Song (2011) and He, Wang and Hong (2013). See detailed discussion in Section 2 of Supplementary Material [Han (2019)].

LEMMA 3.2. *Under conditions in Lemma 3.1, in addition, Condition* D *and* E *are also satisfied, then* $\min_{j\in M_\star} G_j^\star \geq \frac{b^\star}{2} c_1 \xi d_n n^{-2\kappa}$ *for some positive constant* $b^\star$.

As we will show in later sections, the sure screening property depends on the characteristics of a generalized definition for partial derivative of loss function $l(x,y) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ with respect to x. More specifically, let $\widetilde{l}(x,y) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be a Riemann integrable function with respect to x such that for any $x_1 > x_2$ and any y, $l(x_1,y) - l(x_2,y) = \int_{x_2}^{x_1} \widetilde{l}(s,y)\,ds$. Since $l(x,y)$ is differentiable in x almost everywhere, such $\widetilde{l}$ exists and is unique almost everywhere in x. For notational convenience, we simply use $l'(x,y)$ to denote one such $\widetilde{l}(x,y)$. When $l(x,y)$ is differentiable in x, $l'(x,y)$ is uniquely determined. When we consider the quantile regression loss $l(x,y) = (y-x)\{\alpha - \mathbf{I}(y-x<0)\}$, if $x > y$, then $l(x,y) = (y-x)(\alpha-1)$; if $x < y$, then $l(x,y) = (y-x)\alpha$. Except at $x = y$, $\partial l(x,y)/\partial x = \mathbf{I}(y-x<0) - \alpha$. Hence, for quantile regression loss, for any $x_1 > x_2$ and any y, we have $l(x_1,y) - l(x_2,y) = \int_{x_2}^{x_1} [\mathbf{I}(y-s<0) - \alpha]\,ds$. Therefore, we will use $l'(x,y) = \mathbf{I}(y-x<0) - \alpha$ for the quantile regression loss throughout the paper. The above argument motivates the following Definition 2.

DEFINITION 2. The notation $l'(\omega, Y)$ is defined as follows: for Type 1 and 2, $l'(\omega,Y) = \frac{\partial l(\omega,Y)}{\partial \omega}$; for Type 3, $l'(\omega,Y) = \mathbf{I}(Y-\omega<0) - \alpha$.

To simplify the discussion, for the loss function $l(x,y)$ which is not differentiable in x but is differentiable in x almost everywhere, we only focus on the quantile regression loss here. However, similar argument also applies to other loss functions beyond quantile regression loss.

To characterize $l'(\omega,Y)$ for the exponential tail bound in Section 3.2, we also need the following definition for sub-Gaussian random variables.

DEFINITION 3. A random variable X with mean $\mu = EX$ is called σ-sub-Gaussian if there is a positive number σ such that

$$E\exp\bigl(\lambda(X-\mu)\bigr) \leq \exp\biggl(\frac{\lambda^2\sigma^2}{2}\biggr) \qquad \forall \lambda \in \mathbb{R}.$$

Note that if $X \sim N(\mu, \sigma^2)$, then X is σ-sub-Gaussian. If random variable X is bounded such that $a \le X \le b$, then X is sub-Gaussian with $\sigma = (b-a)/2$; see Buldygin and Kozachenko (2000) for more details.

3.2. *Exponential bound for marginal minimum divergence estimator.* Since both $\widehat{f}_{nj}^{M}$ and f_{nj}^{M} can be expressed in terms of B-spline basis functions, it is crucial to establish an exponential bound for the tail probability of $\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|$. The sharpness of this exponential bound directly affect the convergence probability of the screening method.

The following Theorem 3.1 provides an exponential bound for the tail probability of marginal minimum divergence estimator for the B-spline coefficients. It will serve as the cornerstone for our later derivations of the other theorems. The following conditions are required for Theorem 3.1:

(F) $d_n = o(n^{1/3})$ and $d_n = O(n^{2\kappa})$.
(G) $E[l'(\omega, Y)|\mathbf{X}]$ is bounded for any bounded ω.

PROPOSITION 3.1. *For Types* 1–3, *under the conditions in Proposition* 2.1, *condition G is satisfied.*

The tail probability of $\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|$ depends on the properties of $l'(\omega, Y)$. More specifically, we will consider the following set of conditions:

(H1) $l'(\omega, Y)$ is bounded for any bounded ω;
(H2) $l'(\omega, Y)$ conditional on $\mathbf{X}$ is a σ-sub-Gaussian random variable where σ does not depend on $\mathbf{X}$;
(H3) For any bounded ω, $E[\exp(\lambda l'(\omega, Y))|\mathbf{X}] < \infty$ for all $|\lambda| \le c_0$ with some constant $c_0 > 0$.

The notation $l'(\omega, Y)$ in Condition G and H1–H3 is based on Definition 2. By Definition 3, if Condition H1 is satisfied, then Condition H2 is also satisfied; if Condition H2 is satisfied, then Condition H3 is also satisfied. In the following Theorem 3.1, we will show that with stronger assumption a better tail probability can be correspondingly achieved.

To better understand the wide applicability of Conditions H1–H3, let us consider some examples from Types 1–3 which satisfy these conditions. Some popular regression models can be summarized in the following Proposition 3.2.

PROPOSITION 3.2. *Types* 2 *and* 3 *satisfy Condition* H1. *For Type* 1, *if* $Y|\mathbf{X}$ *follows Bernoulli distribution*, *then* (5) *satisfies Condition* H1; *if* $Y|\mathbf{X}$ *follows Normal distribution*, *then* (5) *satisfies Condition* H2; *if* $Y|\mathbf{X}$ *follows Poisson distribution*, *then* (5) *satisfies Condition* H3.

Furthermore, if $Y|\mathbf{X}$ follows some other distributions in the exponential family, under some regularity conditions, it is possible that the corresponding loss function (5) also satisfy Condition H3. For example, if $Y|\mathbf{X} \sim \text{Laplace}(\mu(\mathbf{X}), b)$ with a known parameter b, then Condition H3 is satisfied. If $Y|\mathbf{X} \sim \text{Exponential}(\lambda(\mathbf{X}))$, and if there exists a positive constant c such that $\lambda(\mathbf{X}) \geq c$, then Condition H3 is satisfied. Similar arguments for verifying Condition H3 also apply to Chi-square distribution, negative binomial distribution, inverse-Gaussian distribution with a known shape parameter, Gamma distribution with a known scale parameter. To save space, we will not discuss in detail for these examples.

THEOREM 3.1. *For a convex loss $l(\omega, Y)$, if it is also a conditional strictly convex loss, for any constant $c_3 > 0$, under Conditions C, F, G, there exist positive constants c_4 and c_5 such that for sufficiently large n:*

if Condition H1 *is satisfied and $d_n = o(n^{1-2\kappa})$, then*

$$P\big(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|^2 \geq c_3 d_n n^{-2\kappa}\big) \leq \exp\big(-c_4 n^{1-2\kappa} d_n^{-1}\big); \tag{16}$$

if Condition H2 *is satisfied and $d_n = o(n^{(1-2\kappa)/2})$, then*

$$P\big(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|^2 \geq c_3 d_n n^{-2\kappa}\big) \leq \exp\big(-c_4 n^{1-2\kappa} d_n^{-1}\big) + \exp\big(-c_5 n^{1-2\kappa} d_n^{-2}\big); \tag{17}$$

if Condition H3 *is satisfied and $d_n = o(n^{(1-2\kappa)/3})$, then*

$$\begin{aligned} P\big(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|^2 \geq c_3 d_n n^{-2\kappa}\big) \leq{}& \exp\big(-c_4 n^{1-2\kappa} d_n^{-1}\big) \\ &+ 2\exp\big(-c_5 n^{1/2-\kappa} d_n^{-3/2}\big). \end{aligned} \tag{18}$$

In (16), when $d_n = o(n^{1-2\kappa})$, $n^{1-2\kappa} d_n^{-1}$ in the tail probability diverges to infinity as n increases, which implies that the tail probability converges to zero. Similar arguments also apply to (17) and (18). It is worth mentioning that Theorem 3.1 is proved based on a unified argument with some modifications according to each situation of Conditions H1–H3. The proof is different from the related literature and can be of independent research interest.

3.3. *Sure screening.* Based on Theorem 3.1 for estimation of B-spline coefficients, we are now ready to establish the sure screening property for our Goffins method. Different properties of loss functions can lead to different convergence probabilities of containing the true model.

THEOREM 3.2. *Under the conditions in Theorem 3.1 and Lemma 3.2:*

(i) *for Types 1 and 2, there exists a positive constant ζ, then by taking $\nu_n = \nu d_n n^{-2\kappa}$ with $0 < \nu \leq \zeta$, there exists positive constants c_4, c_5 and c_6 such that if Condition* H1 *is satisfied and $d_n = o(n^{1-2\kappa})$, then*

$$P(M_\star \subset \widehat{M}_{\nu_n}) \geq 1 - s_n\big[\exp\big(-c_4 n^{1-2\kappa} d_n^{-1}\big) + 6\exp\big(-c_5 n^{1-2\kappa}\big)\big]; \tag{19}$$

if Condition H2 *is satisfied and* $d_n = o(n^{(1-2\kappa)/2})$, *then*

$$P(M_\star \subset \widehat{M}_{\nu_n}) \geq 1 - s_n\big[\exp\big(-c_4 n^{1-2\kappa} d_n^{-1}\big) + \exp\big(-c_5 n^{1-2\kappa} d_n^{-2}\big) + 6\exp\big(-c_6 n^{1-2\kappa}\big)\big]; \tag{20}$$

if Condition H3 *is satisfied and* $d_n = o(n^{(1-2\kappa)/3})$, *then*

$$P(M_\star \subset \widehat{M}_{\nu_n}) \geq 1 - s_n\big[\exp\big(-c_4 n^{1-2\kappa} d_n^{-1}\big) + 2\exp\big(-c_5 n^{1/2-\kappa} d_n^{-3/2}\big) + 6\exp\big(-c_6 n^{1-2\kappa}\big)\big]; \tag{21}$$

(ii) *for Type* 3, *if* $d_n = o(n^{1-2\kappa})$, *take* $\nu_n = \nu d_n n^{-2\kappa}$ *with* $\nu \leq b^\star c_1 \xi/4$ *where* $b^\star$ *is defined in Lemma* 3.2, *there exist positive constants* c_4 *and* c_5 *such that*

$$P(M_\star \subset \widehat{M}_{\nu_n}) \geq 1 - s_n\big[\exp\big(-c_4 n^{1-2\kappa} d_n^{-1}\big) + 12\exp\big(-c_5 n^{1-2\kappa}\big)\big]. \tag{22}$$

Theorem 3.2 shows that our Goffins method corresponding to conditional strictly convex loss possesses the sure screening property. It follows from Theorem 3.2 that in (19) and (22) we can handle the NP-dimensionality: $\log p_n = o(n^{1-2\kappa} d_n^{-1})$. Under this condition, $P(M_\star \subset \widehat{M}_{\nu_n}) \to 1$ to achieve the sure screening property. For (20), the NP-dimensionality will be changed to $\log p_n = o(n^{1-2\kappa} d_n^{-2})$ and for (21) we can handle $\log p_n = o(n^{1/2-\kappa} d_n^{-3/2})$.

The proof of Theorem 3.2 in the Supplementary Material [Han (2019)] is not limited to Types 1–3. For example, let Class A be the loss functions such that $l''(\omega, Y) \equiv \partial^2 l(\omega, Y)/\partial\omega^2$ exists, $l''(\omega, Y)$ is continuous in ω, $l''(\omega, Y) > 0$ and $l''(\omega, Y)$ is bounded when ω is bounded, then the results corresponding to Types 1–2 in Theorem 3.2 are also valid for the loss functions in Class A. It is not difficult to verify that Types 1–2 are only special examples in Class A. When $l(\omega, Y)$ is not differentiable in ω, the discussion is more complicated. Let $\widetilde{l}(x, y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a Riemann integrable function with respect to x such that for any $x_1 > x_2$ and any y, $l(x_1, y) - l(x_2, y) = \int_{x_2}^{x_1} \widetilde{l}(s, y)\, ds$. Since we assume that the loss function $l(x, y)$ is differentiable in x almost everywhere, such $\widetilde{l}$ exists and is unique almost everywhere in x. Let Class B be the loss functions such that there exists a corresponding $\widetilde{l}$ where $\widetilde{l}(\omega, Y)$ is bounded and $\widetilde{l}(\omega, Y)$ is nondecreasing in ω, then the results corresponding to Type 3 is also valid for the loss functions in Class B. Our definition of $l'(\omega, Y)$ for quantile regression loss clearly satisfies the conditions in Class B.

3.4. *Controlling selection size.* The sure screening methods will not be informative unless the model selection size can be controlled at a reasonable level. The following Theorem 3.3 shows that our Goffins method can control the size of the selected variables at the level of the sample size n rather than the dimension p_n. However, for controlling the model size, our definition of conditional strictly convex loss is not sufficient for the discussion. We need a finer class of loss functions which possesses certain structures. This is the motivation of our following Definition 4.

DEFINITION 4. If a convex loss $l(\omega, Y)$ satisfies that

$$\partial E\big[l(\omega, Y)|\mathbf{X}\big]/\partial\omega = G(\omega) - H(\mathbf{X})K(\omega)$$

for some functions $G(\cdot)$, $H(\cdot)$ and $K(\cdot)$, then $l(\omega, Y)$ is called conditional derivative separable loss.

PROPOSITION 3.3. *Types* 1–3 *are conditional derivative separable loss functions*.

For Type 1, $\partial E[l(\omega, Y)|\mathbf{X}]/\partial\omega = b'(\omega) - \mu(\mathbf{X})$ where $\mu(\mathbf{X}) = E(Y|\mathbf{X})$. Therefore, $G(\omega) = b'(\omega)$, $H(\mathbf{X}) = \mu(\mathbf{X})$ and $K(\omega) = 1$.

For Type 2, $\partial E[l(\omega, Y)|\mathbf{X}]/\partial\omega = \exp(\omega) - [\exp(\omega) + \exp(-\omega)]p(\mathbf{X})$ where $p(\mathbf{X}) = E(Y|\mathbf{X})$. Therefore, $G(\omega) = \exp(\omega)$, $H(\mathbf{X}) = p(\mathbf{X})$ and $K(\omega) = \exp(\omega) + \exp(-\omega)$.

For Type 3, $\partial E[l(\omega, Y)|\mathbf{X}]/\partial\omega = F_{Y|X}(\omega) - \alpha$ where $F_{Y|X}$ is the conditional cumulative distribution function of $Y|\mathbf{X}$. Therefore, $G(\omega) = F_{Y|X}(\omega)$, $H(\mathbf{X}) = \alpha$ and $K(\omega) = 1$.

Detailed discussions reveal different structures of loss functions. For example, Types 1 and 3 both have a constant $K(\omega) = 1$ while Type 2 does not have such property. Furthermore, Types 1 and 2 are second differentiable in ω but Type 3 is not differentiable in ω. Fortunately, we can propose a unified proof to bound $\sum_{j=1}^{p_n} E(f_{nj}^M - Ef_{nj}^M)^2$ when the loss function $l(\omega, Y)$ is conditional derivative separable loss and conditional strictly convex loss, which will serve as a major step for controlling the model selection size.

THEOREM 3.3. *Let* $\mathbf{\Psi} = (\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_{p_n})^T$, $\boldsymbol{\beta}^\star$ *be the coefficient vector of basis functions for the joint regression model of* $\theta(\mathbf{X})$ *on* $\mathbf{X}$, $\beta_0^\star$ *be the intercept term in the joint regression model and* $\mathbf{\Sigma} = E\mathbf{\Psi}\mathbf{\Psi}^T$. *If* l *is a conditional derivative separable loss and conditional strictly convex loss, under conditions in Theorem* 3.2, *in addition,* $E(\mathbf{\Psi}^T\boldsymbol{\beta}^\star)^2 = O(1)$ *and* $K(\beta_0^\star) \neq 0$, *then we have*:

(i) $\sum_{j=1}^{p_n} E(f_{nj}^M - Ef_{nj}^M)^2 = O(d_n\lambda_{\max}(\mathbf{\Sigma}))$;

(ii) *with* ν_n *described in Theorem* 3.2, *there exist constants* c_4, c_5, c_6 *such that*

Types 1 and 2: *if Condition* H1 *is satisfied and* $d_n = o(n^{1-2\kappa})$, *then*

$$P\big(|\widehat{M}_{\nu_n}| \le O\big(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\big)\big) \ge 1 - p_n\big[\exp\big(-c_4 n^{1-2\kappa}d_n^{-1}\big) + 6\exp\big(-c_5 n^{1-2\kappa}\big)\big];$$

if Condition H2 *is satisfied and* $d_n = o(n^{(1-2\kappa)/2})$, *then*

$$P\big(|\widehat{M}_{\nu_n}| \le O\big(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\big)\big) \ge 1 - p_n\big[\exp\big(-c_4 n^{1-2\kappa}d_n^{-1}\big) + \exp\big(-c_5 n^{1-2\kappa}d_n^{-2}\big) + 6\exp\big(-c_6 n^{1-2\kappa}\big)\big];$$

if Condition H3 *is satisfied and* $d_n = o(n^{(1-2\kappa)/3})$, *then*

$$P\big(|\widehat{M}_{\nu_n}| \le O\big(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\big)\big) \ge 1 - p_n\big[\exp\big(-c_4 n^{1-2\kappa}d_n^{-1}\big) + 2\exp\big(-c_5 n^{1/2-\kappa}d_n^{-3/2}\big) + 6\exp\big(-c_6 n^{1-2\kappa}\big)\big];$$

Type 3: *if* $d_n = o(n^{1-2\kappa})$, *then*

$$P\big(|\widehat{M}_{\nu_n}| \le O\big(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\big)\big) \ge 1 - p_n\big[\exp\big(-c_4 n^{1-2\kappa} d_n^{-1}\big) + 12\exp\big(-c_5 n^{1-2\kappa}\big)\big].$$

The tail probabilities directly follow from Theorem 3.2 and have been explained in Section 3.3. The selected model size depends on the matrix $\mathbf{\Sigma}$ which involves the dependence structure of the covariates. As discussed in Fan, Feng and Song (2011), it can be assumed that $\lambda_{\max}(\mathbf{\Sigma}) = n^{\tau}$ for some $\tau > 0$. Correspondingly, the model size will be controlled at a reasonable rate of n. With iterative Goffins method described in the next Section 4, we will show in the simulation studies that the number of false positives can be very small while the true important variables are all selected even when the covariates are correlated.

3.5. *Connection and comparison with related literature.* When l is the squared error loss, Fan, Feng and Song's (2011) screening for nonparametric additive models is based on $n^{-1}\sum_{i=1}^{n}(\widehat{f}_{nj}^{M}(X_{i,j}))^2$. They presented a similar result to Theorem 3.2 here but with a different convergence probability as

$$P(M_\star \subset \widehat{M}_{\nu_n}) \ge 1 - s_n d_n\big\{(8+2d_n)\exp\big(-c_4^* n^{1-4\kappa} d_n^{-3}\big) + 6d_n\exp\big(-c_5^* n d_n^{-3}\big)\big\},$$

and they can handle the NP dimensionality $\log p_n = o(n^{1-4\kappa} d_n^{-3})$. Compared with their result, our convergence probability in (20) for Gaussian regression not only improve on the convergence rate, but also improve significantly on those coefficient terms. We can achieve NP-dimensionality $\log p_n = o(n^{1-2\kappa} d_n^{-2})$. It should be noted that Fan, Feng and Song (2011) established the result under a weaker assumption than our Condition H2. More specifically, they assume $Y = \theta(\mathbf{X}) + \varepsilon$, $E(\varepsilon|\mathbf{X}) = 0$ and for any $B_1 > 0$, $E[\exp(B_1|\varepsilon|)|\mathbf{X}] \le B_2$ for some constant B_2. On the other hand, this condition is stronger than our Condition H3. If we use (21) for the comparison here in favor of Fan, Feng and Song's (2011) result, then we can handle the NP-dimensionality $\log p_n = o(n^{1/2-\kappa} d_n^{-3/2})$. When $n^{1/3-2\kappa} = O(d_n)$, we can handle a higher dimensionality. Otherwise, their result is better.

When l is the quantile regression loss, He, Wang and Hong's (2013) screening is based on $n^{-1}\sum_{i=1}^{n}(\widehat{f}_{nj}^{M}(X_{i,j}))^2$. They have shown that for positive constants c_6^*, c_7^* and c_8^*,

$$P\big(\|\widehat{\boldsymbol{\beta}}_j^{M} - \boldsymbol{\beta}_j^{M}\|^2 \ge c_6^* d_n n^{-4\kappa}\big) \le 2\exp\big(-c_7^* n^{1-8\kappa}\big) + \exp\big(-c_8^* n^{1-4\kappa} d_n^{-2}\big).$$

Correspondingly, they presented a convergence probability as

$$P(M_\star \subset \widehat{M}_{\nu_n}) \ge 1 - s_n\big\{11\exp\big(-c_9^* n^{1-8\kappa}\big) + 12 d_n^2\exp\big(-c_{10}^* n^{1-4\kappa} d_n^{-3}\big)\big\}.$$

Note that He, Wang and Hong (2013) consider a signal strength in Condition D as $c_1 n^{-2\kappa}$, and the parameter τ in Theorem 3.3 of He, Wang and Hong (2013) is equivalent to 2κ in our paper here. If we reset the minimum signal strength

in Condition D the same as that of He, Wang and Hong (2013), our result for Theorem 3.1 will be modified as

$$P(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|^2 \geq c_3 d_n n^{-4\kappa}) \leq \exp(-c_4 n^{1-4\kappa} d_n^{-1})$$

under the conditions of He, Wang and Hong (2013). Correspondingly, our result for Theorem 3.2 will be modified as

$$P(M_\star \subset \widehat{M}_{\nu_n}) \geq 1 - s_n[\exp(-c_4 n^{1-4\kappa} d_n^{-1}) + 12\exp(-c_5 n^{1-4\kappa})].$$

Therefore, He, Wang and Hong's (2013) convergence probability also indicates larger possibility of not selecting important covariates for a nonasymptotic setting.

When l is the negative log-likelihood loss for one-parameter exponential families, Fan and Song (2010) constructed sure screening for generalized linear models. If the B-spline approximation is treated as a type of group variable selection, then our Theorem 3.1 has some connections with Fan and Song's (2010) result. Compared with Fan and Song (2010), the tail probability in our Theorem 3.1 does not have the extra term $nP(\Omega_n^c)$ in their paper where n is the sample size and Ω_n is the region such that the loss function satisfies some Lipschitz condition. In Fan and Song (2010), their exponential bound also involves a Lipschitz constant. When the response variable is not bounded (e.g., most of the exponential families), this Lipschitz constant diverges to infinity, which results in a slower convergence rate for the tail bound, in contrast with our result. For example, when considering the squared error loss, Fan and Song (2010) Theorem 4 will have

$$P(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \geq c_3 n^{-\kappa}) \leq \exp(-c_4 n^{(1-2\kappa)/3}) + nm_1 \exp(-c_4 n^{(1-2\kappa)/3})$$

for bounded covariates. For our Theorem 3.1 (under Condition H2), let $d_n = 2$ for a fair comparison, then we have

$$P(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \geq c_3 n^{-\kappa}) \leq 2\exp(-c_4 n^{1-2\kappa}).$$

It is clear that we have a much better result here. When considering the Poisson regression loss, the corresponding convergence rate in the tail probability bound can be much slower than $(1-2\kappa)/3$. For our Theorem 3.1 (under Condition H3), let $d_n = 2$ for a fair comparison, we have

$$P(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \geq c_3 n^{-\kappa}) \leq \exp(-c_4 n^{1-2\kappa}) + 2\exp(-c_5 n^{1/2-\kappa}).$$

It is still a better result than Fan and Song (2010).

4. Iterative Goffins method and improved variant. In practice, unimportant variables can be correlated with the important variables, therefore, such variables can have significant marginal effects even though they are not significant in the joint true model. To improve the performance of our screening method, we consider an iterative version of Goffins. Given the data $\{(\mathbf{X}_i, Y_i)\}$, $i = 1, \ldots, n$, we choose the same truncation term $d_n = O(n^{1/5})$. In Theorem 3.2, the threshold ν_n

is chosen at the level $d_n n^{-2\kappa}$. In practice, the parameter κ is unknown, but we can determine a data-driven threshold. To achieve this, we extend the random permutation idea of Fan, Feng and Song (2011) and Zhao and Li (2012). Let $\mathbf{X}$ be the matrix with the ith row as $\mathbf{X}_i$. The algorithm works as follows:

Step 1. For every $j \in \{1, \ldots, p\}$, compute

$$\widehat{f}_{nj} = \arg\min_{f_{nj} \in S_n} \mathbb{P}_n l\big(f_{nj}(X_j), Y\big) \qquad 1 \le j \le p.$$

Randomly permute the rows of $\mathbf{X}$ and we have $\widetilde{\mathbf{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_p)$. Let $\omega_{(q)}$ be the qth quantile of $\{G^*_{n,j}, j = 1, \ldots, p\}$, where $\widehat{f}^*_{nj} = \arg\min_{f_{nj} \in S_n} \mathbb{P}_n l(f_{nj}(\widetilde{X}_j), Y)$. Then our method selects the following variables: $\mathcal{A}_1 = \{j : G_{n,j} \ge \omega_{(q)}\}$. In our numerical studies, we choose $q = 1$, the maximum value of the empirical norm of the permuted estimates.

Step 2. Apply penalized regression on the set $\mathcal{A}_1$ to select a subset $\mathcal{M}_1$. Specifically, when l is the negative log-likelihood loss, apply the penalized generalized additive model regression [e.g., penGAM in Meier, van de Geer and Bühlmann (2009)].

Step 3. For every $j \in \mathcal{M}_1^c = \{1, \ldots, p\}/\mathcal{M}_1$, minimize $\mathbb{P}_n l(f_0 + \sum_{i \in \mathcal{M}_1} f_{ni}(X_i) + f_{nj}(X_j), Y)$ with respect to $f_0 \in \mathbb{R}$, $f_{ni} \in S_n$ for all $i \in \mathcal{M}_1$ and $f_{nj} \in S_n$. For identifiability, we apply the B spline basis without the intercept for $j \in \mathcal{M}_1^c$ and for $i \in \mathcal{M}_1$. Apply the screening procedure with adaptive threshold determined by the new random permutation. Choose a set of indices $\mathcal{A}_2$. Then penalized regression is applied on the set $\mathcal{M}_1 \cup \mathcal{A}_2$ to select a subset $\mathcal{M}_2$.

Step 4. Iterate the process until $|\mathcal{M}_l| \ge s_0$ or $\mathcal{M}_l = \mathcal{M}_{l-1}$.

This iterative version of Goffins will be denoted as "I-Goffins" in our simulation studies. To further stabilize the performance, we can apply a "cap" to control the number of selected variables in each iteration. For example, in our simulation studies, we restrict to select 1 variable at each step. Since the chance of selecting unimportant variables in each step has been reduced, the probability of selecting important variables in the subsequent steps has been improved. This is the idea behind the greedy INIS method proposed by Fan, Feng and Song (2011) for additive modeling. To be consistent with Fan, Feng and Song (2011), we name this improved variant of our method as greedy iterative goodness-of-fit nonparametric screening (GI-Goffins).

5. Simulation studies. Similar to Fan, Feng and Song (2011), we set $n = 400$ but we consider $p = 1000, 2000, 5000$ for all examples to investigate the impact of high dimensionality on screening methods. Following Fan, Feng and Song (2011), we consider the number of spline basis functions as $d_n = \lceil n^{1/5} \rceil + 2 = 6$. Note that in this paper we consider the full B spline basis, and Fan, Feng and Song (2011) considered the B spline basis without the intercept. The goodness-of-fit screening methods under the two sets of basis are equivalent. Eight simulation examples

will be constructed according to the four major types of regressions: Gaussian regression, Logistic regression, Poisson regression and quantile regression. Define

$$f_1(x) = x, \qquad f_2(x) = (2x-1)^2, \qquad f_3(x) = \sin(2\pi x)/\big(2-\sin(2\pi x)\big),$$
$$f_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3,$$
$$f_5(x) = \exp(x-0.5), \qquad f_6(x) = 0.1\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3,$$
$$f_7(x) = \sin(x-1), \qquad f_8(x) = (x-1.5)^2, \qquad f_9(x) = 2\cos(x)/\big(2-\sin(x)\big).$$

- Model 1 (Linear regression): $Y|\mathbf{X} = 5f_1(X_1)+3f_2(X_2)+4f_3(X_3)+6f_4(X_4)+\sqrt{1.74}\varepsilon$. Each $X_i \sim \text{Uniform}(0,1)$ i.i.d. and $\varepsilon \sim N(0,1)$.
- Model 2 (Linear regression): The model is the same as Model 1 but the covariates $\mathbf{X} = (X_1,\ldots,X_p)^T$ are simulated according to the random effects model $X_j = (W_j + tU)/(1+t)$, $j = 1,\ldots,p$ where $W_1,\ldots,W_p$ and U are i.i.d. Uniform$(0,1)$ and $t = 0.4$.
- Model 3 (Logistic regression): $\ln(P(Y=1|\mathbf{X})/P(Y=0|\mathbf{X})) = 2f_1(X_1) + 3f_7(X_2) + 2f_8(X_3) + 3.5f_9(X_4)$. Each $X_i \sim \text{Uniform}(-2.5, 2.5)$ i.i.d.
- Model 4 (Logistic regression): The model is the same as Model 3 but the covariates $\mathbf{X} = (X_1,\ldots,X_p)^T$ are simulated according to the random effects model $X_j = (W_j + tU)/(1+t)$, $j = 1,\ldots,p$ where $W_1,\ldots,W_p$ are i.i.d. from Uniform$(-2.5, 2.5)$, independent of $U \sim \text{Uniform}(0,1)$ and $t = 0.4$.
- Model 5 (Poisson regression): $Y|\mathbf{X} \sim \text{Poisson}(\exp\{f_1(X_1)+f_3(X_2)+f_5(X_3)+f_6(X_4)\})$. Each $X_i \sim \text{Uniform}(0,1)$ i.i.d.
- Model 6 (Poisson regression): The model is the same as Model 5 and the covariates $\mathbf{X} = (X_1,\ldots,X_p)^T$ are simulated according to the same structure as Model 2.
- Model 7 (Heteroscedastic regression): $Y|\mathbf{X} = 5f_1(X_1)+3f_2(X_2)+4f_3(X_3)+4f_5(X_4)+0.5\exp(f_6(X_{20})+f_7(X_{21})+f_8(X_{22}))\varepsilon$, where $\mathbf{X} \sim N_p(0,\boldsymbol{\Sigma})$ independent of $\varepsilon \sim N(0,1)$ and the (i,j)th element of covariance matrix $\boldsymbol{\Sigma}$ is $0.8^{|i-j|}$.
- Model 8 (Heteroscedastic regression): The model is the same as Model 7 except that the random error $\varepsilon \sim \text{Laplace}(0,2)$.

Models 1 and 2 have been similarly considered in Meier, van de Geer and Bühlmann (2009) and Fan, Feng and Song (2011), while Models 3–8 are newly proposed in the current paper. The covariates are independent in Models 1, 3, 5 but correlated in Models 2, 4, 6, 7, 8. Note that Model 8 is different from Model 7 because the Laplace$(0,2)$ distribution for random error will emphasize more on the covariates X_{20}, X_{21}, X_{22}, thus making the heteroscedatic regression model more challenging.

Minimum model size. Following Fan and Song (2010), Fan, Feng and Song (2011) as well as later literature in sure screening field, we use the minimum model

size required to contain the true model $M_\star$ as a measure of the effectiveness of a screening method. The simulation round is 500 for all the examples. We compare our Goffins method with six other successful screening methods in the existing literature, including some recent model-free screening methods. More specifically, we consider fused Kolmogorov filter (Kfilter) by Mai and Zou (2015), quantile adaptive screening (QaSIS) by He, Wang and Hong (2013), SIS for generalized linear model by Fan and Song (2010), sure independent ranking and screening (SIRS) by Zhu, Li, Li and Zhu (2012), distance correlation learning (DC) by Li, Zhong and Zhu (2012) and empirical likelihood screening (EL) by Chang, Tang and Wu (2013). Note that when considering squared error loss, our Goffins is equivalent to NIS by Fan, Feng and Song (2011). Therefore, we will treat NIS as a special example of Goffins, and will not present NIS as a separate method for comparison here. However, when considering quantile regression loss, our Goffins method is different from QaSIS in He, Wang and Hong (2013), because Goffins is based on goodness-of-fit statistics while QaSIS is based on squared norm of fitted nonparametric function. We choose quantile 75% whenever a quantile regression loss is considered. The implementations of QaSIS and SIRS are based on http://users.stat.umn.edu/~wangx346/research/example1b.txt. The implementations of Kfilter, DC and EL are based on the R codes from the authors of related literature. The implementation of SIS is based on the R package "SIS."

In Tables 1–2 along with the Tables S1–S2 in the Supplementary Material [Han (2019)], we present the median, the interquartile range (IQR) and different quantiles of minimum model size. Following existing literature, if the median is closer to the true model size and the IQR is smaller, the corresponding screening method is considered as more effective. Overall, our Goffins method performs best among the seven screening methods. For Models 1, 3, 4, 5, 6, our medians of minimum model size are close to the true model size 4 and IQRs are the smallest. Our medians and IQRs will not increase significantly when the dimensionality p increases from 1000 to 5000. For comparison, other methods tend to select a much larger model to contain the true model, and the performance can deteriorate dramatically when p increases. Furthermore, 5%, 25%, 75% and 95% quantiles of our minimum model size are significantly smaller than the other methods. Models 7–8 are very challenging heteroscedastic regression models, but our method still performs better than the other methods, including Kfilter and QaSIS. Table S2 in the Supplementary Material [Han (2019)] also suggests that even when we consider quantile regression loss, Goffins is different from QaSIS. Model 2 turns out to be a difficult example for all the methods. However, our simulation in tables S3–S5 of Supplementary Material [Han (2019)] will show that an iterative version of Goffins (GI-Goffins or I-Goffins) can substantially reduce the false positives while selecting the true important variables.

TABLE 1
The median (IQR) and 5%, 25%, 75%, 95% quantiles of minimum model size

Model	p	Methods	Median (IQR)	5%	25%	75%	95%
Model 3	1000	Goffins	5 (7.25)	4	4	11.25	67.05
Size = 4		Kfilter	14 (28)	4	6	34	115.05
		QaSIS	NA	NA	NA	NA	NA
		SIS	393.5 (537)	38.9	158	695	941.05
		SIRS	393 (537)	38.95	158	695	941.05
		DC	16 (25)	4	9	34	112.05
		EL	451 (524.25)	55.75	211	725.25	948.05
	2000	Goffins	5 (6)	4	4	10	64.05
		Kfilter	13 (28.25)	4	7	35.25	119.65
		QaSIS	NA	NA	NA	NA	NA
		SIS	376.5 (521)	29.85	168.50	689.5	954.05
		SIRS	376 (521.5)	29.85	168	689.5	954.05
		DC	16 (24.25)	4	8	32.25	105.20
		EL	441 (508.25)	38.95	215.50	723.75	961.05
	5000	Goffins	9 (36.25)	4	4	40.25	301.20
		Kfilter	59 (151.5)	5	20	171.5	597
		QaSIS	NA	NA	NA	NA	NA
		SIS	2164 (2819.5)	165.1	887.5	3707.0	4814.4
		SIRS	2163.5 (2817.5)	164.20	888.75	3706.25	4814.40
		DC	67 (130.5)	8	27.75	158.25	504.15
		EL	2465.5 (2705.75)	229.85	1136.25	3842.00	4846.05
Model 4	1000	Goffins	5 (6)	4	4	10	55.1
Size = 4		Kfilter	16 (36)	4	6	42	163.1
		QaSIS	NA	NA	NA	NA	NA
		SIS	418.5 (566)	20.9	148.75	714.75	920.3
		SIRS	418.5 (566.25)	19.95	148.50	714.75	920.3
		DC	17 (33)	4	8	41	137.05
		EL	495 (543.75)	42	211.75	755.50	933.15
	2000	Goffins	6 (17)	4	4	21	131.1
		Kfilter	32.5 (72.25)	4	10	82.25	243.35
		QaSIS	NA	NA	NA	NA	NA
		SIS	858 (1079.25)	36.85	312.25	1391.50	1892
		SIRS	858 (1077)	36.9	314.5	1391.5	1892
		DC	34 (70)	5	14	84	240.15
		EL	1008 (1037.25)	83.85	446.75	1484	1911.05
	5000	Goffins	9 (32)	4	4	36	332.4
		Kfilter	59.5 (174)	5	16	190	770.75
		QaSIS	NA	NA	NA	NA	NA
		SIS	1857.5 (2778.5)	84.95	749.50	3528.00	4675.10
		SIRS	1855.5 (2781.5)	84.95	746.50	3528.00	4675.10
		DC	70 (160.5)	8	26	186.5	661.0
		EL	2261.5 (2701)	196.8	1054.5	3755.5	4726.5

TABLE 2
The median (IQR) and 5%, 25%, 75%, 95% quantiles of minimum model size

Model	***p***	**Methods**	**Median (IQR)**	**5%**	**25%**	**75%**	**95%**
Model 5	1000	Goffins	4 (0)	4	4	4	12.1
Size = 4		Kfilter	28 (63)	4	11	74	265.05
		QaSIS	NA	NA	NA	NA	NA
		SIS	447 (525.75)	22.95	174.25	700	944.15
		SIRS	491 (501)	46	242.5	743.5	951.05
		DC	11 (18)	4	6	24	71.05
		EL	461.5 (510)	32.9	194.75	704.75	949
	2000	Goffins	4 (1)	4	4	5	17
		Kfilter	60 (144)	5	17	161	504.1
		QaSIS	NA	NA	NA	NA	NA
		SIS	980 (1154.25)	48.95	371.75	1526	1924.05
		SIRS	1031 (1028)	112.75	516	1544	1904.30
		DC	21 (36)	4	9	45	132
		EL	1013 (1166.75)	54.9	393.75	1560.5	1923.05
	5000	Goffins	4 (2)	4	4	6	42
		Kfilter	147.5 (377)	7	41.75	418.75	1340.00
		QaSIS	NA	NA	NA	NA	NA
		SIS	2237.5 (2662.75)	163.80	978.25	3641.00	4748.45
		SIRS	2619.5 (2518.75)	266.80	1468.75	3987.50	4723.20
		DC	44 (98.25)	6	17	115.25	360.70
		EL	2320 (2701)	183.85	988.00	3689.00	4773.15
Model 6	1000	Goffins	4 (1.25)	4	4	5.25	19.05
Size = 4		Kfilter	35.5 (95)	4.95	12	107	345.05
		QaSIS	NA	NA	NA	NA	NA
		SIS	400.5 (571.75)	28	150.25	722	940.05
		SIRS	510.5 (514)	49.95	244.5	758.5	952.15
		DC	17 (36)	4	7	43	129
		EL	425 (567)	30.95	163	730	940.05
	2000	Goffins	4 (2)	4	4	6	31
		Kfilter	66.5 (159)	5	22	181	552.7
		QaSIS	NA	NA	NA	NA	NA
		SIS	930.5 (1067.25)	41.95	356	1423.25	1905.10
		SIRS	1027.5 (1051.75)	94.80	488.50	1540.25	1911.05
		DC	23 (47)	5	10	57	240.05
		EL	972.5 (1059.5)	50.80	383.25	1442.75	1912
	5000	Goffins	5 (7)	4	4	11	74.4
		Kfilter	182 (483.25)	6	52	535.25	1614.10
		QaSIS	NA	NA	NA	NA	NA
		SIS	2072.5 (2690.75)	198	825	3515.75	4709.70
		SIRS	2442 (2528.25)	231.70	1324	3852.25	4778.45
		DC	74 (173.5)	6	23.75	197.25	593.15
		EL	2130 (2627.5)	214.6	926.5	3554.0	4710.3

6. Data analysis. Classification between the malignant pleural mesothelioma (MPM) and the lung cancer adenocarcinoma (ADCA) has received increasing attention in both clinical studies and high dimensional statistical research. Gordon et al. (2002) studied the data from 181 tissue samples (31 MPM and 151 ADCA) with 12,533 gene expression levels for each sample. Among these 181 sample data, 16 MPM and 16 ADCA have been combined as the training set while the other 149 samples (15 MPM and 134 ADCA) are considered as the testing set. The goal of research is in two-fold as explained in Gordon et al. (2002): 1. Find the minimum number of predictor genes that are most importantly associated with the disease type; 2. Construct a classifier rule which can predict the future patients' disease type based on their gene expression levels with high statistical accuracy. Aspect 1 can substantially reduce the medical cost of obtaining patients' relevant gene data and the cost of potential scientific experiments on such genes. The performance of the classifier is usually evaluated based on the testing data.

Since the disease type is a categorical data, and the number of genes is extremely high ($p = 12{,}533$) compared with the small sample size ($n = 32$), we will apply our GI-Goffins method with respect to the logistic regression for the training data. We first standardize the gene expression data for each gene over the training samples such that the sample mean is 0 and the sample standard deviation is 1. Our method selects five genes that are importantly associated with the disease type: "31575-f-at," "37716-at," "39795-at," "41286-at" and "41402-at." We construct a generalized additive model (B spline basis without the intercept and the number of spline basis functions as $d_n = \lceil n^{1/5} \rceil + 1 = 3$) based on such five genes and apply the model to the training data. The fitted nonparametric functions corresponding to those five genes have also been plotted in Figure S1 in the Supplementary Material [Han (2019)]. Then we apply our constructed model to the test data. Among the 149 samples for the testing data, we make 144 correct predictions. For the 5 samples that we misclassified, one MPM sample has been predicted as ADCA while four ADCA samples has been predicted as MPM. ISIS for the generalized linear model has also been considered to select important variables. To be fair, we also apply a generalized additive model based on the selected genes for the training data and further use this fitted model for classification on the test data. However, this method will select fewer and different genes and the performance is much inferior to our method. I-EL is an iterative version of EL and penalized empirical likelihood regression described in Chang, Tang and Wu (2013). Its performance is even worse than ISIS.

This lung cancer data has also been analyzed by various statistical methods in the past literature. It is impossible and unnecessary for us to list all the relevant results here, and we only compare our method with some representative methods which have been shown superior performance. In Table 3, we will compare our GI-Goffins method with linear discriminant methods such as ROAD in Fan, Feng and Tong (2012) and FAIR in Fan and Fan (2008). Our GI-Goffins is a good balance between the testing error and the number of selected genes compared with other

TABLE 3
Performance of methods on lung cancer data. $p = 12{,}533$

Method	Training error	Testing error	Number of selected genes
GI-Goffins	0/32	5/149	5
ISIS	0/32	18/149	2
I-EL	0/32	40/149	5
ROAD	1/32	1/149	52
FAIR	0/32	7/149	31

methods. More selected genes will cause substantial cost in future diagnosis and experiments. Therefore, GI-Goffins is the method that we recommend for practice.

7. Further discussions.

7.1. *Optimality.* An interesting question is whether the convergence rate in the upper bound of the tail probability that we established in Theorem 3.1 is optimal. More specifically, if we have

$$b_1 \exp(-c_2 n^a) \le P\big(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|^2 \ge c_1 d_n n^{-2\kappa}\big) \le b_2 \exp(-c_3 n^a)$$

for some constants a, b_1, b_2, c_1, c_2 and c_3, then we can say that the convergence rate a in the upper bound of the tail probability is optimal, because the convergence rate a cannot be improved further.

When the loss function l is the negative log likelihood loss of one-parameter exponential families, under general regularity conditions, the maximum likelihood estimator has the asymptotic normality [Heyde (1997), Gao et al. (2008)], that is,

$$\big[I_j(\boldsymbol{\beta}_j^M)\big]^{1/2}\big(\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\big) - N(0, \mathbf{I}_{d_n}) \to 0 \qquad \text{in distribution,}$$

where I_j is the information matrix of the jth covariate. Plugging in the negative log likelihood loss and the B-spline basis functions, we have

$$n^{1/2}\big\{E\big[b''(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M)\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T\big]\big\}^{1/2}\big(\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\big) - N(0, \mathbf{I}_{d_n}) \to 0 \qquad \text{in distribution.}$$

Since $\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M$ is bounded based on the argument in the Supplementary Material [Han (2019)] and $b''(\cdot)$ is a continuous function, $b''(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M)$ is upper bounded by a positive constant. Furthermore, due to Lemma 3 in the Supplementary Material [Han (2019)],

$$\big(\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\big)^T n\big\{E\big[b''(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M)\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T\big\}\big(\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\big) \le D_2 n d_n^{-1} \|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|^2.$$

Therefore, asymptotically, we have

$$\begin{aligned} P\big(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|^2 \ge c_1 d_n n^{-2\kappa}\big) &= P\big(D_2 n d_n^{-1} \|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|^2 \ge D_2 c_1 n^{1-2\kappa}\big) \\ &= P\big(\chi_{d_n}^2 \ge D_2 c_1 n^{1-2\kappa}\big). \end{aligned}$$

Thus, we need to find a lower bound for the tail probability of $\chi^2_{d_n}$ distribution. When $d_n = 1$, it is well known that for any positive y,

$$P(\chi_1^2 \geq y) \geq 1 - \sqrt{1 - \exp\Big(-\frac{2y}{\pi}\Big)} = \frac{\exp(-\frac{2y}{\pi})}{1 + \sqrt{1 - \exp(-\frac{2y}{\pi})}} \geq \frac{1}{2}\exp\Big(-\frac{2y}{\pi}\Big).$$

Let $y = D_2 c_1 n^{1-2\kappa}$, comparing with our Theorem 3.1 under Condition H1 or H2, we have achieved the optimal convergence rate $n^{1-2\kappa}$ asymptotically. When $d_n = 2$, for any positive y, $P(\chi_2^2 \geq y) = \exp(-\frac{y}{2})$. Comparing with our Theorem 3.1 under Condition H1 or H2, we have also achieved the optimal convergence rate $n^{1-2\kappa}$ asymptotically. For more general d_n, we do not have a sharp lower bound of the tail probability of Chi-square distribution. Therefore, we will not discuss further here.

7.2. *Adaptive threshold.* Theorem 3.2 is established based on a threshold ν_n at the level of $d_n n^{-2\kappa}$. In practice, the parameter κ is unknown. Therefore, we need an adaptive threshold for the real data. Consider a threshold $\widehat{\nu}_n$ which is constructed based on the sample data, it will be interesting to derive a lower bound for $P(M_\star \subseteq \widehat{M}_{\widehat{\nu}_n})$. We have

$$P(M_\star \subseteq \widehat{M}_{\widehat{\nu}_n}) \geq 1 - \sum_{j \in M_\star} P(G_{n,j} < \widehat{\nu}_n).$$

Note that

$$P(G_{n,j} < \widehat{\nu}_n) \leq P(G_{n,j} < \nu_n) + P(\widehat{\nu}_n > \nu_n).$$

We have derived the upper bound of $P(G_{n,j} < \nu_n)$ is the proof of Theorem 3.2. Therefore, we need to derive an upper bound for the second term here.

Consider a permutation of the sample covariates $\{\mathbf{X}_i\}_{i=1}^n$. We can obtain the estimates of marginal regression based on the permuted data:

$$(\widehat{\boldsymbol{\beta}}_j^M)^\pi = \arg\min_{\beta_j \in \mathbb{R}^{d_n}} \mathbb{P}_n l\big(\boldsymbol{\Psi}_j^T(X_j^\pi)\boldsymbol{\beta}_j, Y\big),$$

where $\pi = (\pi_1, \ldots, \pi_n)$ is a permutation of the index $\{1, 2, \ldots, n\}$. Note that $(G_{n,j})^\pi$ is a statistical estimate of 0. We will derive an upper bound for $P(\widehat{\nu}_n > \nu)$ for a special case where the loss function $l(\cdot)$ is the squared error loss. Note that the least squares estimate follows

$$\sqrt{n}\widehat{\boldsymbol{\beta}}_j^\pi = \bigg(\frac{1}{n}\boldsymbol{\Psi}(X_j^\pi)\boldsymbol{\Psi}(X_j^\pi)^T\bigg)^{-1}\bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\Psi}^T(X_j^{\pi_i})Y_i\bigg).$$

By Anderson and Robinson (2001) Theorem 3.2, after some algebra, we have the asymptotic normality

$$\sqrt{n}\widehat{\boldsymbol{\beta}}_j^\pi - N(0, \widetilde{\boldsymbol{\Sigma}}) \to 0 \qquad \text{in distribution,}$$

where $\widetilde{\boldsymbol{\Sigma}} = (E\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}[E(\boldsymbol{\Psi} - E\boldsymbol{\Psi})(\boldsymbol{\Psi} - E\boldsymbol{\Psi})^T](E\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}$.

If we let $\widehat{\nu}_{n,j} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\Psi}^T(X_j^{\pi_i})\widehat{\boldsymbol{\beta}}_j^{\pi})^2$, different thresholds for the marginal utilities of different covariates. As we have discussed in the Section 2.2, this screening based on $\mathbb{P}_n(\boldsymbol{\Psi}_j^T\widehat{\boldsymbol{\beta}}_j^M)^2$ is equivalent to our goodness-of-fit screening when the loss function is the squared error loss. We can show that asymptotically

$$P\big(\widehat{\nu}_{n,j} \geq c_1 d_n n^{-2\kappa}\big) \leq P\big(\chi^2_{d_n} \geq c_2 d_n n^{-2\kappa}\big) \leq \exp\big(-c_3 d_n n^{1-2\kappa}\big)$$

for some constants c_1, c_2 and c_3. The second inequality is by Laurent and Massart (2000) Lemma 3.1. Correspondingly, we have a lower bound for the convergence probability of containing the true model. In practice, the threshold $\widehat{\nu}_n$ will be chosen as the maximum value of $\{\widehat{\nu}_{n,j}\}_{j=1}^{p_n}$ under a number of permutations, and the loss function can be general. We do not have a theoretical result for such more complicated situations.

7.3. *Choice of loss function.* The framework of goodness-of-fit nonparametric screening includes many screening methods based on the choice of loss functions. An important question is how to choose loss function for practical data. When the response variable Y takes values $\{0, 1\}$, we suggest to consider the logistic regression loss: $l(\omega, Y) = -\omega Y + \ln(1 + \exp(\omega))$; when Y takes nonnegative integer values, we suggest to consider the Poisson regression loss: $l(\omega, Y) = -Y\omega + \exp(\omega) + \ln(Y!)$; When the distribution of Y is expected to be complicated, the quantile regression loss can be considered; when Y is a continuous variable, we suggest to start with the Gaussian regression loss: $l(\omega, Y) = (Y - \omega)^2/2$. This brief guideline could raise misspecification issue of loss functions.

7.4. *Iterative screening procedure.* The idea of iterative screening and penalization has been proposed since Fan and Lv (2008), and has achieved numerical success in practice. However, formal theoretical justification is still an open problem in the field. The first step is a marginal screening. To simplify the discussion, assume a fixed threshold γ_n is applied the selected variables $\mathrm{A}_1 = \{j : G_{n,j} \geq \gamma_n\}$ satisfies $M_\star \subseteq \mathrm{A}_1$ with high probability (sure screening property). For the second step, based on the set A_1, we apply some penalized regression and select a subset M_1. Ideally, we want to show sign consistency for M_1 under some regularity conditions. The difficulty is that the set A_1 is random, which is different from a conventional penalization regression. Fortunately, we can borrow the technique in Weng, Feng and Qiao (2017), which considers a two-step procedure for linear regression model (similar to screening + penalization). For the third step, it is a conditional marginal screening after penalization. Barut, Fan and Verhasselt (2016) has shown the sure screening property based on the conditional screening for generalized linear model. Therefore, if the sign consistency is achieved in step 2, then under some regularity conditions, sure screening property can be achieved in step 3. By mathematical induction, the iterative procedure can achieve sign consistency. We would like to explore the technical details as our future studies.

Acknowledgments. The author wants to thank the Joint Editor, the Associate Editor and the three anonymous referees for many insightful comments which significantly improve the presentation of the paper.

The author deeply appreciates Professor Jianqing Fan for his encouragement and constructive comments on this project. The author would like to thank Dr. Cheng Yong Tang and Professor Linda Zhao for helpful discussions on the paper. The author also thanks Dr. Shujie Ma and Dr. Lily Wang for the helpful discussion on B spline basis.

Special thanks go to the following researchers who kindly share their codes of numerical studies: Dr. Yang Feng for NIS, Dr. Qing Mai for Kfilter, Dr. Lukas Meier for penGAM, Dr. Cheng Yong Tang for EL and Dr. Wei Zhong for DC.

SUPPLEMENTARY MATERIAL

Supplement to "Nonparametric screening under conditional strictly convex loss for ultrahigh dimensional sparse data" (DOI: 10.1214/18-AOS1738SUPP; .pdf). Due to the space limit, all the technical proofs as well as some numerical results are relegated to the Supplementary Material [Han (2019)].

REFERENCES

ANDERSON, M. J. and ROBINSON, J. (2001). Permutation tests for linear models. *Aust. N. Z. J. Stat.* **43** 75–88. MR1837497

BARUT, E., FAN, J. and VERHASSELT, A. (2016). Conditional sure independence screening. *J. Amer. Statist. Assoc.* **111** 1266–1277. MR3561948

BRÈGMAN, L. M. (1967). A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Ž. Vyčisl. Mat. Mat. Fiz.* **7** 620–631. MR0215617

mr BULDYGIN, V.and KOZACHENKO, Y. (2000). Metric characterization of random variables and random processes.. *Translations of Mathematical Monographs* **188**.

CANDÈS, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2404.

CHANG, J., TANG, C. Y. and WU, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.* **41** 2123–2148. MR3127860

DE BOOR, C. (1978). *A Practical Guide to Splines. Applied Mathematical Sciences* **27**. Springer, New York. MR0507062

FAN, J. and FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009

FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557. MR2847969

FAN, J., FENG, Y. and TONG, X. (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 745–771. MR2965958

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322

FAN, J., MA, Y. and DAI, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *J. Amer. Statist. Assoc.* **109** 1270–1284. MR3265696

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **10** 2013–2038. MR2550099

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

FREUND, Y. and SCHAPIRE, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139. MR1473055

GAO, Q., WU, Y., ZHU, C. and WANG, Z. (2008). Asymptotic normality of maximum quasi-likelihood estimators in generalized linear models with fixed design. *J. Syst. Sci. Complex.* **21** 463–473. MR2425677

GORDON, G. et al. (2002). Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma. *Cancer Research* **62** 4963–4967.

HAN, X. (2019). Supplement to "Nonparametric screening under conditional strictly convex loss for ultrahigh dimensional sparse data." DOI:10.1214/18-AOS1738SUPP.

HE, X., WANG, L. and HONG, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.* **41** 342–369. MR3059421

HEYDE, C. C. (1997). *Quasi-Likelihood and Its Application*: *A General Approach to Optimal Parameter Estimation*. Springer, New York. MR1461808

KOENKER, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657

LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. MR1805785

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900

LI, G., PENG, H., ZHANG, J. and ZHU, L. (2012). Robust rank correlation based screening. *Ann. Statist.* **40** 1846–1877. MR3015046

MAI, Q. and ZOU, H. (2015). The fused Kolmogorov filter: A nonparametric model-free screening method. *Ann. Statist.* **43** 1471–1497. MR3357868

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443

SONG, R., LU, W., MA, S. and JENG, X. J. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika* **101** 799–814. MR3286918

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606. MR0840516

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

WENG, H., FENG, Y. and QIAO, X. (2017). Regularization after retention in ultrahigh dimensional linear regression models. *Statist. Sinica*. In press.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHANG, C., JIANG, Y. and SHANG, Z. (2009). New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Canad. J. Statist.* **37** 119–139. MR2509465

ZHAO, S. D. and LI, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Multivariate Anal.* **105** 397–411. MR2877525

ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

DEPARTMENT OF STATISTICAL SCIENCE
FOX BUSINESS SCHOOL
TEMPLE UNIVERSITY
PHILADELPHIA, PENNSYLVANIA 19122
USA
E-MAIL: hanxu3@temple.edu

Submitted to the Annals of Statistics

SUPPLEMENTARY MATERIALS TO "NONPARAMETRIC SCREENING UNDER CONDITIONAL STRICTLY CONVEX LOSS FOR ULTRAHIGH DIMENSIONAL SPARSE DATA"

By Xu Han

Temple University

1. Technical Proofs.

LEMMA 1. *If l is a conditional strictly convex loss, for any bounded functions $f_1(\mathbf{X})$ and $f_2(\mathbf{X})$ defined on the vector $\mathbf{X} = (X_1, \cdots, X_p)^T$, correspondingly there exists positive constants $b_1 < b_2$ such that*

$$\begin{aligned} & b_1(f_1(\mathbf{X}) - f_2(\mathbf{X}))^2 \\ \leq \; & (f_1(\mathbf{X}) - f_2(\mathbf{X}))\Big\{ \frac{\partial E[l(\omega, Y)|\mathbf{X}]}{\partial \omega}\Big|_{\omega = f_1(X)} - \frac{\partial E[l(\omega, Y)|\mathbf{X}]}{\partial \omega}\Big|_{\omega = f_2(X)} \Big\} \\ \leq \; & b_2(f_1(\mathbf{X}) - f_2(\mathbf{X}))^2 \end{aligned}$$

Note that the positive constants b_1 and b_2 depend on the bounds of $f_1(\mathbf{X})$ and $f_2(\mathbf{X})$. To simplify the notation, without loss of generality, we will use b_1 and b_2 whenever we apply Lemma 1 to the later proofs,

LEMMA 2 (Stone, 1986). *There exists a positive constant C_2 such that for any $1 \leq k \leq d_n$ and $1 \leq j \leq p$, $E\Psi_k^2(X_j) \leq C_2 d_n^{-1}$.*

LEMMA 3 (Zhou, Shen & Wolfe, 1998). *There exists some positive constants D_1 and D_2 such that*

$$D_1 d_n^{-1} \leq \lambda_{\min}(E\mathbf{\Psi}_j\mathbf{\Psi}_j^T) \leq \lambda_{\max}(E\mathbf{\Psi}_j\mathbf{\Psi}_j^T) \leq D_2 d_n^{-1}$$

where λ is the eigenvalue of a matrix.

LEMMA 4 (Boucheron, Lugosi and Massart, 2003). *Let $X_1, \cdots, X_n$ denote independent random variables taking values in some measurable set $\mathcal{X}$. Denote by X_1^n the vector of these n random variables. Let $F : \mathcal{X}^n \to \mathbb{R}$ be some measurable function. Let $Z = F(X_1, \cdots, X_n)$. Let $X_1', \cdots, X_n'$ denote independent copies of $X_1, \cdots, X_n$, and write $Z_i' = F(X_1, \cdots, X_{i-1}, X_i', X_{i+1},$*

$\cdots, X_n)$. *Define* $V^+ = E[\sum_{i=1}^n (Z - Z_i')_+^2 | X_1^n]$. *If* $V^+ \leq c$ *almost surely for some positive constant* c, *then for any* $\lambda > 0$,

$$E \exp\{\lambda(Z - EZ)\} \leq \exp(\lambda^2 c).$$

LEMMA 5 (Buldygin & Kozachenko (2000) Lemma 1.6). *Given any zero-mean random variable* X, *if there is a constant* σ *such that* $E \exp(\lambda X) \leq \exp(\lambda^2 \sigma^2/2)$ *for all* $\lambda \in \mathbb{R}$, *then we have* $E \exp(\frac{\xi X^2}{2\sigma^2}) \leq (1-\xi)^{-1/2}$ *for all* $\xi \in [0, 1)$.

LEMMA 6 (Buldygin & Kozachenko (2000) Theorem 3.1). *For a zero-mean random variable* X, *the following statements are equivalent:*

1. *There are non-negative numbers* (ν, b) *such that* $E \exp(\lambda X) \leq \exp(\nu^2 \lambda^2/2)$ *for all* $|\lambda| < b^{-1}$.
2. *There is a positive number* $c_0 > 0$ *such that* $E \exp(\lambda X) < \infty$ *for all* $|\lambda| \leq c_0$.

Proof of Proposition 2.1.1:
For Type 1,

$$\frac{\partial^2 E[l(\omega, Y)|\mathbf{X}]}{\partial \omega^2} = b''(\omega) > 0.$$

The proof for Type 1 is complete.

For Type 2, note that

$$\frac{\partial^2 E[l(\omega, Y)|\mathbf{X}]}{\partial \omega^2} = E[\exp(-Y\omega)|\mathbf{X}] > 0.$$

The proof for Type 2 is now complete.

For Type 3, we have

$$\frac{\partial^2 E[l(\omega, Y)|\mathbf{X}]}{\partial \omega^2} = \frac{\partial E[\mathbf{I}(Y < \omega) - \alpha|\mathbf{X}]}{\partial \omega} = \frac{\partial F_{Y|X}(\omega)}{\partial \omega} = f_{Y|X}(\omega) > 0.$$

The proof for Type 3 is now complete.

Proof of Lemma 1: Given $\mathbf{X}$, without loss of generality, assume $f_1(\mathbf{X}) > f_2(\mathbf{X})$. By Taylor's theorem,

$$\begin{aligned} & \frac{\partial E[l(\omega, Y)|\mathbf{X}]}{\partial \omega}\Big|_{\omega = f_1(X)} - \frac{\partial E[l(\omega, Y)|\mathbf{X}]}{\partial \omega}\Big|_{\omega = f_2(X)} \\ = \quad & \frac{\partial^2 E[l(\omega, Y)|\mathbf{X}]}{\partial \omega^2}\Big|_{\omega = \xi(X)} (f_1(\mathbf{X}) - f_2(\mathbf{X})) \end{aligned}$$

where $\xi(\mathbf{X}) \in (f_2(\mathbf{X}), f_1(\mathbf{X}))$. Since $\partial^2 E[l(\omega, Y)|\mathbf{X}]/\partial\omega^2$ is a continuous function in ω and is strictly positive, there exists a positive constant b_1 such that

$$\frac{\partial^2 E[l(\omega, Y)|\mathbf{X}]}{\partial \omega^2}\Big|_{\omega=\xi(X)} \geq b_1.$$

Due to the continuity on a bounded domain, the left hand side of the last line is also upper bounded by a positive constant b_2. Therefore, the proof of Lemma 1 is now complete.

Notations and Formula

For a generic matrix $\mathbf{A}$, we use $\|\mathbf{A}\|$ to denote the operator norm of $\mathbf{A}$. When $\mathbf{A}$ is a vector, $\|\mathbf{A}\|$ is simplified to be the L_2 norm.

Next we will present a very useful decomposition for loss function l. By the notation l' based on Definition 2, we have

$$l(b, y) = l(a, y) + l'(a, y)(b - a) + \int_a^b [l'(s, y) - l'(a, y)]ds. \tag{S1}$$

for any a, b, y. When $l(x, y)$ is differentiable in x, (S1) is equivalent to the fundamental theorem of calculus. When we consider the quantile regression loss, let $l'(x, y) = \mathbf{I}(y - x < 0) - \alpha$, (S1) is due to Definition 3. Knight's identity (Knight 1998) is equivalent to a special case of (S1) where $\alpha = 0.5$.

Proof of Lemma 3.1.1: The proof follows some similar argument in Stone (1986). By Lemma 5 of Stone (1985), there exists $f_{nj} \in S_n$ such that $E(f_{nj}(X_j) - f_j^M(X_j))^2 \leq a_1^2 d_n^{-2d}$, where a_1 is some positive constant.

In the general decomposition (S1), replace a by f_j^M, b by f_{nj} and y by Y. Since $d > 0.5$, and f_j^M is bounded, by Lemma 7 of Stone (1986), f_{nj} is also bounded. Apply expectation with respect to $\mathbf{X}$ and Y to (S1). For the second term in (S1), consider $E_{X,Y} l(tf_{nj} + (1-t)f_j^M)$ as a function of t, it will be minimized at $t = 0$ by the definition of f_j^M, therefore, $E_{X,Y}[l'(f_j^M, Y)(f_{nj} - f_j^M)] = 0$. For the third term in (S1), change variable and let $s = f_j^M + t(f_{nj} - f_j^M)$, then by Lemma 1 we have

$$E_{X,Y}\Big\{\int_0^1 [l'(f_j^M + t(f_{nj} - f_j^M), Y) - l'(f_j^M, Y)]dt(f_{nj} - f_j^M)\Big\} \leq \frac{1}{2} b_2 E_X (f_{nj} - f_j^M)^2. \tag{S2}$$

Therefore, recalling (S1), there exists a positive constant A_1 such that

$$E_{X,Y}[l(f_{nj}, Y) - l(f_j^M, Y)] \leq A_1 d_n^{-2d}$$

for $n \geq 1$. Let c denote a positive constant. Choose v where $v \in S_n$ with $E(v - f_j^M)^2 = c d_n^{-2d}$. Then $E(v - f_{nj})^2 \leq 2(c + a_1^2) d_n^{-2d}$. By Lemma 7 of

Stone (1986), v is bounded. By Lemma 1 there exists a positive constant A_2 such that

$$E_{X,Y}[l(v,Y) - l(f_j^M, Y)] \geq A_2 c d_n^{-2d}$$

for all $v \in S_n$ such that $E(v - f_j^M)^2 = cd_n^{-2d}$. Let c be chosen so that $A_2 c > A_1$. Therefore, $E_{X,Y} l(v,Y) > E_{X,Y} l(f_{nj}, Y)$ for all v such that $E(v - f_j^M)^2 = cd_n^{-2d}$. By the convexity of loss function l, the expectation of this loss function is also convex. Therefore, $E(f_{nj}^M - f_j^M)^2 < cd_n^{-2d}$. The proof of Lemma 3.1.1 is now complete. .

Proof of Lemma 3.1.2: Lemma 1 in Fan, Feng & Song (2011) can not be directly applied, since we don't have the orthogonality property here. Instead, by binomial expansion,

$$\begin{aligned} E(f_{nj}^M - Ef_{nj}^M)^2 &= E(f_{nj}^M - Ef_{nj}^M - f_j^M + Ef_j^M)^2 + E(f_j^M - Ef_j^M)^2 \\ &\quad +2E(f_j^M - Ef_j^M)(f_{nj}^M - Ef_{nj}^M - f_j^M + Ef_j^M). \end{aligned}$$

By Cauchy Schwartz inequality,

$$\begin{aligned} & E[(f_j^M - Ef_j^M)(f_{nj}^M - Ef_{nj}^M - f_j^M + Ef_j^M)] \\ \geq\ & -[E(f_j^M - Ef_j^M)^2]^{1/2}[E(f_{nj}^M - Ef_{nj}^M - f_j^M + Ef_j^M)^2]^{1/2}. \end{aligned}$$

Therefore,

$$\begin{aligned} E(f_{nj}^M - Ef_{nj}^M)^2 &\geq E(f_j^M - Ef_j^M)^2 \\ \text{(S3)} \qquad & -2[E(f_{nj}^M - Ef_{nj}^M - f_j^M + Ef_j^M)^2]^{1/2}[E(f_j^M - Ef_j^M)^2]^{1/2}. \end{aligned}$$

By Condition (D) and Lemma 3.1.1, for $j \in M_\star$, we have

$$[E(f_j^M - Ef_j^M)^2]^{1/2} \geq c_1^{1/2} d_n^{1/2} n^{-\kappa}; \qquad [E(f_{nj}^M - Ef_{nj}^M - f_j^M + Ef_j^M)^2]^{1/2} \leq C_1^{1/2} d_n^{-d}.$$

By Condition (E), it is easy to show that

$$\min_{j \in M_\star} E(f_{nj}^M - Ef_{nj}^M)^2 \geq c_1 \xi d_n n^{-2\kappa}.$$

By the general decomposition (S1), $G_j^\star$ can be expressed as

$$\begin{aligned} & El'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)(\beta_0^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M) \\ +\ & E\Big\{(\beta_0^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M)^T \int_0^1 [l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M + t(\beta_0^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M), Y) - l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)] dt\Big\}. \end{aligned}$$

By the definition of $\boldsymbol{\beta}_j^M$, the first term is 0. Due to the convexity of $l()$, $El(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j, Y)$ has a unique minimum over $\boldsymbol{\beta}_j \in \mathcal{B}$ at an interior point $\boldsymbol{\beta}_j^M$,

where each coordinate of $\boldsymbol{\beta}_j^M$ has absolute value bounded by a constant B. See Fan & Song (2010). Correspondingly, β_0^M is also bounded. Therefore, $\|\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M\|_\infty \leq (\sum_{i=1}^{d_n}\Psi_i)B = B$, where we have used the fact that $\|\Psi_i\|_\infty \leq 1$ and the partition of unity (de Boor 1978). By the definition of conditional strictly convex loss and Lemma 1, the second term is lower bounded. Formally,

$$\begin{aligned} G_j^\star &\geq \frac{b_1}{2}E(\beta_0^M - \boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M)^2 \\ &= \frac{b_1}{2}\big\{(\beta_0^M - E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M)^2 - 2E(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M - E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M)(\beta_0^M - E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M) \\ &\qquad +E(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M - E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M)^2\big\} \\ &\geq \frac{b_1}{2}E(f_{nj}^M - Ef_{nj}^M)^2 \end{aligned}$$

The proof of Lemma 3.1.2 is now complete.

Proof of Proposition 3.2.1: For any bounded ω,

For Type 1, since $l'(\omega, Y) = b'(\omega) - Y$, $E(Y|\mathbf{X}) = b'(\theta)$ and the fact that θ is bounded, the conclusion is correct.

For Type 2, $E[l'(\omega, Y)|\mathbf{X}] = E[-Y\exp(-Y\omega)|\mathbf{X}]$. Since Y only takes 1 or -1, and ω is bounded, the conclusion is correct.

For Type 3, $l'(\omega, Y) = I(Y < \omega) - \alpha$, we have $E[l'(\omega, Y)|\mathbf{X}] = F_{Y|X}(\omega) - \alpha$. The conclusion is correct.

The proof of Proposition 3.2.1 is now complete.

Proof of Proposition 3.2.2:

For Type 2, $l'(\omega, Y) = -Y\exp(-Y\omega)$. Since $Y \in \{-1, 1\}$, for a bounded ω, $l'(\omega, Y)$ is also bounded.

For Type 3, since by Definition 2, $l'(\omega, Y) = \mathbf{I}(Y - \omega < 0) - \alpha$, it is bounded.

For Type 1, $l'(\omega, Y) = b'(\omega) - Y$. If $Y|\mathbf{X}$ follows Bernoulli distribution, $Y \in \{0, 1\}$. Since ω is bounded and $b'()$ is continuous, $b'(\omega)$ is also bounded. Therefore, $l'(\omega, Y)$ is bounded. If $Y|\mathbf{X}$ follows Normal distribution, then $E(Y|\mathbf{X}) = b'(\theta)$ and $Var(Y|\mathbf{X}) = b''(\theta)$. Correspondingly, conditional on $\mathbf{X}$, $b'(\omega) - Y \sim N\big(b'(\omega) - b'(\theta), b''(\theta)\big)$. Since θ is bounded, and $b''()$ is a continuous function, then $b''(\theta)$ is bounded by some positive constant σ^2. By Definition 3, conditional on $\mathbf{X}$, $l'(\omega, Y)$ is $\sigma-$subgaussian and Condition H2 is satisfied. If $Y|\mathbf{X}$ follows Poisson distribution, $E[\exp(\lambda l'(\omega, Y))|\mathbf{X}] = \exp(\lambda b'(\omega))E[\exp(-\lambda Y)|\mathbf{X}]$. Since the moment generating function for Poisson always exists, Condition H3 is satisfied.

The proof of Proposition 3.2.2 is now complete.

The following Lemmas 7-13 will be useful for proving Theorem 3.2.1.

LEMMA 7. *Let $B(N) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_j^M\| \leq N\}$ for a positive constant N. For any $r > 0$,*

$$\begin{aligned} & E\Big[\exp\Big\{r \sup_{\boldsymbol{\beta}\in B(N)} \big|(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)^T(\mathbb{P}_n - E)\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)\big|\Big\}\Big] \\ \leq \quad & \exp\big\{(C_2)^{1/2} r N^2 n^{-1/2} + 4r^2 N^4 n^{-1}\big\} \end{aligned}$$

where C_2 is defined in Lemma 2.

Proof of Lemma 7: In Lemma 4, let $Z = N^2\|(\mathbb{P}_n - E)\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T\|$. For generic two matrices $\mathbf{A}$ and $\mathbf{B}$, the operator norm possesses triangle inequality: $\big|\|\mathbf{A}\| - \|\mathbf{B}\|\big| \leq \|\mathbf{A} - \mathbf{B}\|$. Therefore, in Lemma 7,

$$V^+ \leq n^{-2}N^4 E\Big[\sum_{l=1}^n \|\boldsymbol{\Psi}_j(X_{lj})\boldsymbol{\Psi}_j(X_{lj})^T - \boldsymbol{\Psi}_j(X'_{lj})\boldsymbol{\Psi}_j(X'_{lj})^T\|^2 \big| \{X_{lj}\}_{l=1}^n\Big]$$

where X'_{lj} is an independent copy of X_{lj}. Note that $\|\mathbf{A} - \mathbf{B}\|^2 \leq 2[\|\mathbf{A}\|^2 + \|\mathbf{B}\|^2]$. Since $\|\Psi_i\|_\infty \leq 1$,

$$\|\boldsymbol{\Psi}_j\| = (\sum_{i=1}^{d_n} \Psi_i^2(X_j))^{1/2} \leq (\sum_{i=1}^{d_n} \Psi_i(X_j))^{1/2} = 1, \tag{S4}$$

where we have used the partition of unity and all the B spline basis functions are in $(0, 1]$. See de Boor (1978). Therefore, $\|\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T\|^2 \leq \|\boldsymbol{\Psi}_j\|^4 \leq 1$ and correspondingly $V^+ \leq 4n^{-1}N^4$. Also note that

$$\sup_{\boldsymbol{\beta}\in B(N)} \big|(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)^T(\mathbb{P}_n - E)\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)\big| \leq N^2\|(\mathbb{P}_n - E)\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T\|.$$

By Lemma 4, the left hand side in Lemma 7 is upper bounded by

$$E\Big[\exp\big\{rN^2 E\|(\mathbb{P}_n - E)\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T\|\big\}\Big] \times \exp(4r^2 n^{-1}N^4). \tag{S5}$$

Next we will show that in (S5),

$$E\|(\mathbb{P}_n - E)\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T\| \leq (C_2)^{1/2} n^{-1/2}.$$

For a generic symmetric matrix $\mathbf{A}_{p\times p}$, $\|\mathbf{A}\| \leq \|\mathbf{A}\|_1 \equiv \max_{1\leq j\leq p} \sum_{i=1}^p |a_{i,j}|$ where p is the number of rows. For the kth row, we have

$$\sum_{l=1}^{d_n} E\Big|\frac{1}{n}\sum_{i=1}^n \big\{\Psi_k(X_{i,j})\Psi_l(X_{i,j}) - E\Psi_k(X_{i,j})\Psi_l(X_{i,j})\big\}\Big|,$$

which is upper bounded by

$$\begin{aligned}
&\sum_{l=1}^{d_n}\Big[E\big(\frac{1}{n}\sum_{i=1}^{n}\{\Psi_k(X_{i,j})\Psi_l(X_{i,j})-E\Psi_k(X_{i,j})\Psi_l(X_{i,j})\}\big)^2\Big]^{1/2}\\
\leq\ &\sum_{l=1}^{d_n}\Big[\frac{1}{n^2}\sum_{i=1}^{n}E\Psi_k^2(X_{i,j})\Psi_l^2(X_{i,j})\\
&\quad+\frac{2}{n^2}\sum\sum_{1\leq i<k\leq n}E\{(\Psi_l(X_{i,j})-E\Psi_l(X_{i,j}))(\Psi_l(X_{i,j})-E\Psi_l(X_{i,j}))\}\Big]^{1/2}\\
\leq\ &\Big[d_n\sum_{l=1}^{d_n}n^{-1}E\Psi_k^2(X_{i,j})\Psi_l^2(X_{i,j})\Big]^{1/2}.
\end{aligned}$$

In the first step, the cross terms are 0 because of the independence among the random sample copies. In the second step, we have used the Cauchy Schwartz inequality. Note that $\sum_{l=1}^{d_n}\Psi_l^2(X_{i,j}) \leq \sum_{l=1}^{d_n}\Psi_l(X_{i,j}) < 1$. By Lemma 2, the last line above is further upper bounded by $C_2^{1/2}n^{-1/2}$.

Finally, we have shown that

$$E\|(\mathbb{P}_n - E)\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T\| \leq (C_2)^{1/2}n^{-1/2}.$$

This leads to the conclusion in Lemma 7. The proof of Lemma 7 is now complete.

LEMMA 8. *If $|l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y)| \leq M$ for some constant $M > 0$, then for any constant $r > 0$,*

$$\begin{aligned}
&E\exp\Big(r\big\|\mathbb{P}_n\boldsymbol{\Psi}_j^T(X_{i,j})l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)\big\|\Big)\\
\leq\ &\exp\Big((C_2)^{1/2}Mrn^{-1/2}+4M^2r^2n^{-1}\Big);
\end{aligned}$$

If $|E[l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y)|\mathbf{X}_i]| \leq M_1$ for some constant $M_1 > 0$, then for any $r > 0$,

$$\begin{aligned}
&E\exp\Big(r\big\|\mathbb{P}_n\boldsymbol{\Psi}_j^T(X_{i,j})E[l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i]\big\|\Big)\\
\leq\ &\exp\Big((C_2)^{1/2}M_1rn^{-1/2}+4M_1^2r^2n^{-1}\Big),
\end{aligned}$$

where C_2 is defined in Lemma 2.

Proof of Lemma 8: Apply the result in Lemma 4, we have

$$E\exp\Big(r\big\|\mathbb{P}_n\boldsymbol{\Psi}_j^T(X_{i,j})l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)\big\|\Big)$$

$$\text{(S6)} \qquad \leq \quad \exp\Big(rE\big\|\mathbb{P}_n\boldsymbol{\Psi}_j^T(X_{i,j})l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)\big\|\Big)\exp(r^2c)$$

where $V^+ \leq c$ almost surely. We will carefully analyze V^+ later. For the first term in (S6), we will show that

$$\text{(S7)} \qquad E\|\mathbb{P}_n\boldsymbol{\Psi}_j^T(X_{i,j})l'(\boldsymbol{\Psi}_j^T(X_{i,j})\boldsymbol{\beta}_j^M, Y_i)\| \leq (C_2)^{1/2}Mn^{-1/2}.$$

Note that

$$\begin{aligned}
& E\|\mathbb{P}_n\boldsymbol{\Psi}_j^T(X_{i,j})l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)\| \\
= \quad & E\Big\{\sum_{l=1}^{d_n}\big(\frac{1}{n}\sum_{i=1}^{n}\Psi_l(X_{i,j})l'(\boldsymbol{\Psi}_j^T(X_{i,j})\boldsymbol{\beta}_j^M, Y_i)\big)^2\Big\}^{1/2} \\
\leq \quad & n^{-1}\Big(\sum_{l=1}^{d_n}E\big(\sum_{i=1}^{n}\Psi_l(X_{i,j})l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)\big)^2\Big)^{1/2} \\
= \quad & n^{-1}\Big(\sum_{l=1}^{d_n}\sum_{i=1}^{n}E\big\{\Psi_l^2(X_{i,j})l'(\boldsymbol{\Psi}_j^T(X_{i,j})\boldsymbol{\beta}_j^M, Y_i))^2\big\} \\
& +2\sum_{l=1}^{d_n}\sum_{1\leq i<k\leq n}\sum E\Big\{\Psi_l(X_{i,j})l'(\boldsymbol{\Psi}_j^T(X_{i,j})\boldsymbol{\beta}_j^M, Y_i) \\
& \qquad\qquad \times\Psi_l(X_{k,j})l'(\boldsymbol{\Psi}_j^T(X_{k,j})\boldsymbol{\beta}_j^M, Y_k)\Big\}\Big)^{1/2} \\
= \quad & n^{-1}\Big(\sum_{l=1}^{d_n}\sum_{i=1}^{n}E\Psi_l^2(X_{i,j})(l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i))^2\Big)^{1/2}.
\end{aligned}$$

The second step is by Cauchy Schwartz inequality. The cross terms in the third step are zero by the independence of random samples and the score equations $E\big\{\Psi_l(X_{i,j})l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)\big\} = 0$ for any $l = 1, \cdots, d_n$ and any $i = 1, \cdots, n$ from the definition of $\boldsymbol{\beta}_j^M$. By Lemma 2, it is easy to verify that (S7) is correct.

For the second term in (S6), since Z in Lemma 4 is now defined as

$$Z = \|\mathbb{P}_n\boldsymbol{\Psi}_j^T(X_{i,j})l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)\|,$$

by the triangle inequality and the fact that $\|\boldsymbol{\Psi}_j\| \leq 1$, we have

$$\begin{aligned}
\big|Z - Z_i'\big| \quad \leq \quad & n^{-1}\Big\|\boldsymbol{\Psi}_j^T(X_{i,j})l'(\boldsymbol{\Psi}_j^T(X_{i,j})\boldsymbol{\beta}_j^M, Y_i) \\
& -\boldsymbol{\Psi}_j^T(X_{i,j}')l'(\boldsymbol{\Psi}_j^T(X_{i,j}')\boldsymbol{\beta}_j^M, Y_i')\Big\|
\end{aligned}$$

$$< \quad n^{-1}2M.$$

Therefore, $V^+ < n^{-1}4M^2$. Apply (S6), the proof of the first result in Lemma 8 is now complete.

For the second result in Lemma 8, since $|E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i]| \leq M_1$, and the score equations $E\{\Psi_l(X_{i,j})E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i]\} = 0$ for any $l = 1, \cdots, d_n$ and any $i = 1, \cdots, n$ from the definition of $\boldsymbol{\beta}_j^M$. Following similar arguments as above, we can also prove the second result. Therefore, the proof of Lemma 8 is now complete.

LEMMA 9. *If Condition H2 is satisfied and $d_n = o(n^{(1-2\kappa)/2})$, for any $c_3 > 0$, there exists a constant $c_4 > 0$ such that*

$$P\Big(\big\|\mathbb{P}_n\mathbf{\Psi}_j^T(l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i])\big\| \geq c_3 n^{-\kappa} d_n^{-1/2}\Big) \leq \exp(-c_4 n^{1-2\kappa} d_n^{-2})$$

for sufficiently large n.

Proof of Lemma 9: By Markov inequality,

$$\begin{aligned}
& P\Big(\big\|\mathbb{P}_n\mathbf{\Psi}_j^T(l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i])\big\| \geq c_3 n^{-\kappa} d_n^{-1/2}\Big) \\
= \quad & P\Big(\sum_{l=1}^{d_n}\big(\frac{1}{n}\sum_{i=1}^{n}\Psi_l(X_{i,j})(l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i])\big)\big)^2 \geq c_3^2 n^{-2\kappa} d_n^{-1}\Big) \\
\leq \quad & \exp(-\lambda t) E\exp\Big(\lambda\sum_{l=1}^{d_n}\big(\frac{1}{n}\sum_{i=1}^{n}\Psi_l(X_{i,j})(l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i])\big)^2\Big).
\end{aligned}$$

for any $\lambda > 0$, where $t = c_3^2 n^{-2\kappa} d_n^{-1}$. By Hölder inequality, the second term above can be upper bounded by

$$\begin{aligned}
& \prod_{l=1}^{d_n}\Big\{E\exp\big(\lambda d_n(\frac{1}{n}\sum_{i=1}^{n}\Psi_l(X_{i,j})(l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i]))^2\big)\Big\}^{1/(d_n)} \\
\equiv \quad & \Delta_4.
\end{aligned}$$

For Δ_4, for all $\xi \in \mathbb{R}$,

$$\begin{aligned}
& E\exp\Big(\xi\Psi_l(X_{i,j})\big(l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i]\big)\Big) \\
= \quad & E\Big\{E\big[\exp\big(\xi\Psi_l(X_{i,j})(l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i])\big)|\mathbf{X}_i\big]\Big\} \\
\leq \quad & E\exp\Big(\frac{\xi^2\Psi_l^2(X_{i,j})\sigma^2}{2}\Big) \leq \exp\Big(\frac{\xi^2\sigma^2}{2}\Big),
\end{aligned}$$

because all the B-spline basis are bounded by 1, and conditional on $\mathbf{X}_i$, $l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i) - El'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y)$ is a centered $\sigma-$subgaussian random variable. Therefore, $\Psi_l(X_{i,j})\big(l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i]\big)$ is a centered $\sigma-$subgaussian random variable because of the score equations

$$E\big\{\Psi_l(X_{i,j})l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)\big\} = E\big\{\Psi_l(X_{i,j})E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}]\big\} = 0.$$

Therefore, we have

$$\begin{aligned} & E\exp(\xi\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_l(X_{i,j})(l'(\mathbf{\Psi}_j(X_{i,j})^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y)|\mathbf{X}_i]) \\ = & \prod_{i=1}^{n} E\exp\big(\xi\frac{1}{\sqrt{n}}\Psi_l(X_{i,j})(l'(\mathbf{\Psi}_j(X_{i,j})^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y)|\mathbf{X}_i])\big) \\ \leq & \prod_{i=1}^{n}\exp(\frac{\xi\sigma^2}{2n}) = \exp(\frac{\xi\sigma^2}{2}). \end{aligned}$$

Correspondingly, we have shown that $n^{-1/2}\sum_{i=1}^{n}\Psi_l(X_{i,j})\big(l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i]\big)$ is a centered $\sigma-$subgaussian random variable. By Lemma 5, for $|\frac{2\lambda d_n\sigma^2}{n}| < \frac{1}{4}$, by using the inequality $(1-2x)^{-1/2} \leq \exp(x+2x^2)$ for all $|x| < 1/4$, we further have

$$\Delta_4 \leq \prod_{l=1}^{d_n}\Big(\big(1 - \frac{4\lambda d_n\sigma^2}{n}\big)^{-1/2}\Big)^{1/(d_n)} \leq \exp\Big(\frac{2\lambda d_n\sigma^2}{n} + 2\big(\frac{2\lambda d_n\sigma^2}{n}\big)^2\Big).$$

Finally, we will choose $\lambda = cnd_n^{-1}$ for an appropriate constant c to satisfy the above restrictions. When $d_n = o(n^{(1-2\kappa)/2})$ and n is sufficiently large, the desired tail probability can be achieved. The proof of Lemma 9 is now complete.

LEMMA 10. *If Condition H3 is satisfied and $d_n = o(n^{(1-2\kappa)/3})$, then for any constant $c_3 > 0$, there exists a constant $c_4 > 0$ such that*

$$\begin{aligned} & P\Big(\big\|\mathbb{P}_n\mathbf{\Psi}_j^T(X_{i,j})(l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i])\big\| \geq c_3n^{-\kappa}d_n^{-1/2}\Big) \\ \leq & \ 2\exp(-c_4n^{1/2-\kappa}d_n^{-3/2}) \end{aligned}$$

for sufficiently large n.

Proof of Lemma 10: let $t = c_3n^{-\kappa}d_n^{-1/2}$, by Markov inequality, for any $\lambda > 0$,

$$P\Big(\big\|\mathbb{P}_n\mathbf{\Psi}_j^T(X_{i,j})(l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i])\big\| \geq c_3n^{-\kappa}d_n^{-1/2}\Big)$$

$$\leq \exp(-\lambda t) E \exp\Big(\lambda \big\| \mathbb{P}_n \mathbf{\Psi}_j^T(X_{i,j})(l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i])\big\|\Big)$$

$$\leq \exp(-\lambda t) E \exp\Big(\lambda \sum_{l=1}^{d_n} \big|\frac{1}{n}\sum_{i=1}^{n} \Psi_l(X_{i,j})(l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i])\big|\Big).$$

By Hölder inequality, the second term in the last line is further upper bounded by

$$\prod_{l=1}^{d_n} \Big\{ E \exp\Big(\frac{\lambda d_n}{\sqrt{n}} \big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Psi_l(X_{i,j})(l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i])\big|\Big)\Big\}^{1/(d_n)}.$$

Note that if $E[\exp(\xi l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y))|\mathbf{X}] \leq \infty$ for all $|\xi| \leq c_0$, since $|\Psi_l(X_{i,j})| \leq 1$ and $E[l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)|\mathbf{X}]$ is bounded, then

$$E \exp\Big(\xi \Psi_l(X_{i,j})\big(l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y) - E[l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)|\mathbf{X}_i]\big)\Big) \leq \infty$$

for all $|\xi| \leq c_0$. By Lemma 6, there are non-negative numbers (ν, b) such that

$$E \exp\Big(\zeta \Psi_l(X_{i,j})\big(l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y) - E[l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)|\mathbf{X}_i]\big)\Big) \leq \exp(\frac{\nu^2\zeta^2}{2})$$

for all $|\zeta| < 1/b$. Therefore,

$$E \exp\Big(\zeta \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Psi_l(X_{i,j})\big(l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y) - E[l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)|\mathbf{X}_i]\big)\Big)$$

$$\leq \prod_{i=1}^{n} \exp(\frac{\nu^2\zeta^2}{2n}) = \exp(\frac{\nu^2\zeta^2}{2})$$

for all $|\zeta| < 1/b$. Note that $E\exp(\xi|X|) \leq E\exp(\xi X) + E\exp(-\xi X)$ for a generic random variable X and any $\xi > 0$. Therefore, if we let $|\frac{\lambda d_n}{\sqrt{n}}| < \frac{1}{b}$, then

$$\prod_{l=1}^{d_n} \Big\{ E \exp\Big(\frac{\lambda d_n}{\sqrt{n}} \big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Psi_l(X_{i,j})(l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i])\big|\Big)\Big\}^{1/(d_n)}$$

$$\leq \prod_{l=1}^{d_n} \Big\{ 2\exp\big(\frac{\nu^2\lambda^2 d_n^2}{2n}\big)\Big\}^{1/(d_n)} = 2\exp(\frac{\nu^2\lambda^2 d_n^2}{2n}).$$

In summary, there exists a constant $c_4 > 0$ such that

$$P\Big(\big\|\mathbb{P}_n \mathbf{\Psi}_j^T(X_{i,j})(l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i) - E[l'(\mathbf{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i])\big\| \geq c_3 n^{-\kappa} d_n^{-1/2}\Big)$$

$$\begin{aligned} &\leq 2\exp\Big(-\lambda c_3 n^{-\kappa} d_n^{-1/2}\Big)\exp\Big(\frac{\nu^2\lambda^2 d_n^2}{2n}\Big) \\ &\leq 2\exp(-c_4 n^{1/2-\kappa} d_n^{-3/2}) \end{aligned}$$

for $d_n = o(n^{(1-2\kappa)/3})$ and sufficiently large n. The proof of Lemma 10 is now complete.

LEMMA 11 (Symmetrization, Lemma 2.3.1, van der Vaart and Wellner 1996). *let $Z_1, \cdots, Z_n$ be independent random variables with values in $\mathcal{Z}$ and $\mathcal{F}$ is a class of real valued functions on $\mathcal{Z}$. Then*

$$E\{\sup_{f\in\mathcal{F}} |(\mathbb{P}_n - E)f(Z)|\} \leq 2E\{\sup_{f\in\mathcal{F}} |\mathbb{P}_n \epsilon f(Z)|\}$$

where $\epsilon_1, \cdots, \epsilon_n$ be a Rademacher sequence (i.e., i.i.d. sequence taking values ± 1 with probability 1/2) independent of $Z_1, \cdots, Z_n$.

LEMMA 12 (Contraction theorem, Ledoux and Talagrand 1991). *Let $z_1, \cdots, z_n$ be nonrandom elements of some space $\mathcal{Z}$, and let $\mathcal{F}$ be a class of real valued functions on $\mathcal{Z}$. Let $\epsilon_1, \cdots, \epsilon_n$ be a Rademacher sequence. Consider Lipschitz functions $\gamma_i : \mathbb{R} \to \mathbb{R}$, that is, $|\gamma_i(s) - \gamma_i(\widetilde{s})| \leq |s - \widetilde{s}|$ $\forall$ $s, \widetilde{s} \in \mathbb{R}$. Then for any function $\widetilde{f} : \mathcal{Z} \to \mathbb{R}$, we have*

$$E\{\sup_{f\in\mathcal{F}} |\mathbb{P}_n \epsilon(\gamma(f) - \gamma(\widetilde{f})|\} \leq 2E\{\sup_{f\in\mathcal{F}} \epsilon(f - \widetilde{f})|\}$$

LEMMA 13. *Let*

$$\begin{aligned} \Delta(\boldsymbol{\beta}) \equiv\ & \mathbb{P}_n(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)^T \boldsymbol{\Psi}_j \int_0^1 [l'(\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^M + t(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)), Y_i) - l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i)]dt \\ & -\mathbb{P}_n(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)^T \boldsymbol{\Psi}_j \int_0^1 E[l'(\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^M + t(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)), Y_i) - l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i]dt. \end{aligned}$$

If $|l'(\omega, Y)| \leq M$ for some constant $M > 0$, then for any $r > 0$,

$$E\exp\Big(r \sup_{\boldsymbol{\beta}\in B(N)} |\Delta(\boldsymbol{\beta})|\Big) \leq \exp(C_3 r N n^{-1/2} + C_4 r^2 N^2 n^{-1})$$

for some positive constants C_3 and C_4.

Proof of Lemma 13: First, we will apply symmetrization inequality. In Lemma 13, let

$$Z_i = (\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)^T \boldsymbol{\Psi}_j \Big(\int_0^1 [l'(\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^M + t(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)), Y_i) - l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y_i)]dt$$

$$-\int_0^1 E[l'(\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^M + t(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)), Y_i) - l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i]dt\Big),$$

then we have for an independent sequence of Rademacher variables $\{\epsilon_i\}_{i=1}^n$ such that

$$E\Big\{\sup_{\boldsymbol{\beta}\in B(N)}|\Delta(\boldsymbol{\beta})|\Big\} \le 2E\{\sup_{\boldsymbol{\beta}\in B(N)}|\mathbb{P}_n\epsilon_i Z_i|\Big\}.$$

Then we will apply Lemma 12. Note that

$$|Z_i| \le 4M|(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)^T\boldsymbol{\Psi}_j(X_{ij})|.$$

Therefore, we have

$$E\{\sup_{\boldsymbol{\beta}\in B(N)}|\Delta(\boldsymbol{\beta})|\} \le 16ME\Big\{\sup_{\boldsymbol{\beta}\in B(N)}\frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i\boldsymbol{\Psi}_j^T(X_{i,j})(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)\Big|\Big\}.$$

Since by Cauchy Schwartz inequality,

$$\sup_{\boldsymbol{\beta}\in B(N)}\frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i\boldsymbol{\Psi}_j^T(X_{i,j})(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)\Big| \le \sup_{\boldsymbol{\beta}\in B(N)}\|\frac{1}{n}\sum_{i=1}^n \epsilon_i\boldsymbol{\Psi}_j^T(X_{i,j})\|\|\boldsymbol{\beta}-\boldsymbol{\beta}_j^M\|,$$

then we have

$$\begin{aligned}
E\Big\{\sup_{\boldsymbol{\beta}\in B(N)}|\Delta(\boldsymbol{\beta})|\Big\} &\le 16MNE\Big\|\frac{1}{n}\sum_{i=1}^n \epsilon_i\boldsymbol{\Psi}_j^T(X_{i,j})\Big\| \\
&\le 16MN\Big\{E\sum_{l=1}^{d_n}\big(\frac{1}{n}\sum_{i=1}^n \epsilon_i\Psi_l\big)^2\Big\}^{1/2} \\
&= 16MN\Big\{\sum_{l=1}^{d_n}\frac{1}{n^2}\sum_{i=1}^n E\epsilon_i^2 E\Psi_l^2(X_{i,j})\Big\}^{1/2} \\
&\le C_3Nn^{-1/2}.
\end{aligned}$$

The second step is by Cauchy-Schwartz inequality. The third step is by the independence of ϵ_i and $X_{i,j}$. Next, we will apply Lemma 4, in which we let $Z = \sup_{\boldsymbol{\beta}\in B(N)}|\Delta(\boldsymbol{\beta})|$. Since for generic functions $f(x)$ and $g(x)$, we have $\sup_x f(x) - \sup_x g(x) \le \sup_x |f(x) - g(x)|$. Therefore, in Lemma 4,

$$|Z - Z_i'| \le \frac{4M}{n}\sup_{\boldsymbol{\beta}\in B(N)}\Big|(\boldsymbol{\beta} - \boldsymbol{\beta}_j^M)^T(\Psi_j(X_{i,j}) + \Psi_j(X_{i,j}'))\Big| \le 8MNn^{-1}.$$

Hence, in Lemma 4, $V^+ \le C_4 N^2 n^{-1}$. Applying Lemma 4, we have

$$E\exp\Big(r\sup_{\boldsymbol{\beta}\in B(N)}|\Delta(\boldsymbol{\beta})|\Big) \le \exp\Big(C_3 r N n^{-1/2} + C_4 r^2 N^2 n^{-1}\Big).$$

The proof of Lemma 13 is now complete.

To prove Theorem 3.2.1, we will construct a linear combination of $\boldsymbol{\beta}_j^M$ and $\widehat{\boldsymbol{\beta}}_j^M$. Given a positive value N, Let $\boldsymbol{\beta}_j^s = s\widehat{\boldsymbol{\beta}}_j^M + (1-s)\boldsymbol{\beta}_j^M$ with $s = (1 + \|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|/N)^{-1}$. This combination has been used in van de Geer (2002) and Fan & Song (2010). It is easy to show that $\|\boldsymbol{\beta}_j^s - \boldsymbol{\beta}_j^M\| \le N$. Next we define $W(\boldsymbol{\beta}) := \sum_{i=1}^n W_i(\boldsymbol{\beta})$ where

$$\begin{aligned} W_i(\boldsymbol{\beta}) &= \frac{1}{n}\frac{b_1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)^T E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M) \\ &\quad -\frac{1}{n}\big[l(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}, Y_i) - l(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)\big] \end{aligned}$$

for an arbitrary $\boldsymbol{\beta}$ in $B(N) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_j^M\| \le N\}$, where b_1 is defined in Lemma 1 and further determined in the proof of Lemma 14. Construction of $W(\boldsymbol{\beta})$ is newly proposed in the current paper to facilitate the proof.

LEMMA 14. *With the conditions in Lemma 7 and 8, for any positive constant b_6, choose N as $bd_n^{1/2}n^{-\kappa}$ for an appropriate constant b and for sufficiently large n, if Condition H1 is satisfied and $d_n = o(n^{1-2\kappa})$, then*

$$P(W(\boldsymbol{\beta}_j^s) \ge b_6 n^{-2\kappa}) \le \exp(-b_7 n^{1-2\kappa} d_n^{-1})$$

for some constant $b_7 > 0$. If Condition H2 is satisfied and $d_n = o(n^{(1-2\kappa)/2})$, then

$$P(W(\boldsymbol{\beta}_j^s) \ge b_6 n^{-2\kappa}) \le \exp(-b_7 n^{1-2\kappa} d_n^{-1}) + \exp(-b_8 n^{1-2\kappa} d_n^{-2}).$$

If Condition H3 is satisfied and $d_n = o(n^{(1-2\kappa)/3})$, then

$$P(W(\boldsymbol{\beta}_j^s) \ge b_6 n^{-2\kappa}) \le \exp(-b_7 n^{1-2\kappa} d_n^{-1}) + 2\exp(-b_8 n^{1/2-\kappa} d_n^{-3/2}).$$

Proof of Lemma 14: Consider the definition of $W_i(\boldsymbol{\beta})$, apply the general decomposition in (S1) and change variable, we have

$$\begin{aligned} & l(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^s, Y) - l(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y) \\ = \;& l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y)\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^s - \boldsymbol{\beta}_j^M) \end{aligned}$$

$$
\begin{aligned}
&+(\boldsymbol{\beta}_j^s-\boldsymbol{\beta}_j^M)^T\boldsymbol{\Psi}_j\int_0^1\big[l'(\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^M+t(\boldsymbol{\beta}_j^s-\boldsymbol{\beta}_j^M)),Y)-l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M,Y)\big]dt\\
\equiv\;& \Delta l_1(\boldsymbol{\beta}_j^s)+\Delta l_2(\boldsymbol{\beta}_j^s).
\end{aligned}
$$

Therefore,

$$
\text{(S8)}\quad W_i(\boldsymbol{\beta}_j^s)=\frac{1}{n}\frac{b_1}{2}(\boldsymbol{\beta}_j^s-\boldsymbol{\beta}_j^M)^TE\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^s-\boldsymbol{\beta}_j^M)-\frac{1}{n}\big[\Delta l_1(\boldsymbol{\beta}_j^s)+\Delta l_2(\boldsymbol{\beta}_j^s)\big].
$$

Note that for any $\|\boldsymbol{\beta}-\boldsymbol{\beta}_j^M\|\le N$,

$$
\|\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^M+t(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M))\|_\infty\le\|\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M\|_\infty+t\|\boldsymbol{\Psi}_j\|\times\|\boldsymbol{\beta}-\boldsymbol{\beta}_j^M\|.
$$

Due to the convexity of $l()$, $El(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j,Y)$ has a unique minimum over $\boldsymbol{\beta}_j\in\mathcal{B}$ at an interior point $\boldsymbol{\beta}_j^M$, where each coordinate of $\boldsymbol{\beta}_j^M$ has absolute value bounded by a constant B. See Fan & Song (2010). Note that $\|\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M\|_\infty\le(\sum_{i=1}^{d_n}\Psi_i)B=B$. Therefore,

$$
\|\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^M+t(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M))\|_\infty\le B+tN.
$$

With the above argument, we can apply the definition of conditional strictly convex loss and Lemma 1 and achieve an upper bound for $W(\boldsymbol{\beta}_j^s)$.

We will first consider the situation where $l''(\omega,Y)$ exists. The formula (S1) can be further replaced by a Taylor expansion and $W(\boldsymbol{\beta}_j^s)$ is upper bounded by

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta}\in B(N)}\Big\{\frac{b_1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)^TE\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)\\
&\qquad\qquad-\mathbb{P}_n\int_0^1 l''\big(\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^M+t(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)),Y_i\big)\big(\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)\big)^2(1-t)dt\Big\}\\
&+\sup_{\boldsymbol{\beta}\in B(N)}\mathbb{P}_n\boldsymbol{\Psi}_j^Tl'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M,Y_i)(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)\\
\le\;&\frac{b_1}{2}\sup_{\boldsymbol{\beta}\in B(N)}\big|(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)^T(E-\mathbb{P}_n)\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)\big|\\
&+\sup_{\boldsymbol{\beta}\in B(N)}\|\boldsymbol{\beta}-\boldsymbol{\beta}_j^M\|\cdot\|\mathbb{P}_n\boldsymbol{\Psi}_j^T(X_{i,j})l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M,Y_i)\|\\
\equiv\;&\widetilde{W}\equiv\Delta l_3+\Delta l_4,
\end{aligned}
$$

where in the second step we have applied Lemma 1.

If Condition H1 is satisfied, by Markov inequality,

$$
\text{(S9)}\qquad P(W(\boldsymbol{\beta}_j^s)\ge t)\le P(\widetilde{W}\ge t)\le\exp(-rt)E[\exp(r\widetilde{W})]
$$

for any positive r. We will choose an appropriate r such that the right hand side leads to our desired tail probability.

By Cauchy Schwartz inequality,

$$E[\exp(r\widetilde{W})] \leq \Big[E\exp(2r\Delta l_3)\Big]^{1/2}\Big[E\exp(2r\Delta l_4)\Big]^{1/2}. \tag{S10}$$

By Lemma 7 and 8, we have

$$\begin{aligned} P(W(\boldsymbol{\beta}_j^s)\geq t) \leq\ & \exp(-rt)\exp\big(0.5(C_2)^{1/2}b_1rN^2n^{-1/2}+2b_1^2r^2N^4n^{-1} \\ & +(C_2)^{1/2}MrNn^{-1/2}+8M^2r^2N^2n^{-1}\big). \end{aligned} \tag{S11}$$

When $t=b_6n^{-2\kappa}$, choose $r=b_{16}ntN^{-2}$ and $N=b_{17}d_n^{1/2}n^{-\kappa}$ for constants $b_{16}>0$ and $b_{17}>0$, where b_{16} and b_{17} satisfy

$$\begin{aligned} & (C_2)^{1/2}Mb_{17}d_n^{1/2}n^{\kappa-1/2}+8M^2b_{16}b_6+0.5(C_2)^{1/2}b_1b_{17}^2d_nn^{-1/2} \\ & +2b_1^2b_{16}b_{17}^2d_nn^{-2\kappa}b_6<b_6(1-\frac{\epsilon}{2}) \end{aligned} \tag{S12}$$

for some $0<\epsilon<1$, then the right hand side of expression (S11) is less than $\exp(-b_7n^{1-2\kappa}d_n^{-1})$ for some positive constant b_7. Note that since $d_n=o(n^{1-2\kappa})$, $d_n=o(n^{1/3})$ and $d_n=O(n^{2\kappa})$, we can choose sufficiently small b_{16} and b_{17} such that (S12) is satisfied.

If Condition H2 is satisfied, by triangle inequality,

$$\begin{aligned} \Delta l_4 \leq\ & N\big\|\mathbb{P}_n\boldsymbol{\Psi}_j^T(X_{i,j})E[l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M,Y_i)|\mathbf{X}_i]\big\| \\ & +N\big\|\mathbb{P}_n\boldsymbol{\Psi}_j^T(X_{i,j})\big(l'\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M,Y_i)-E[l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M,Y_i)|\mathbf{X}_i]\big)\big\| \\ \equiv\ & \Delta l_5+\Delta l_6. \end{aligned}$$

Therefore,

$$P\Big(W(\boldsymbol{\beta}_j^s)\geq t\Big)\leq P\Big(\Delta l_3+\Delta l_5\geq\frac{t}{2}\Big)+P\Big(\Delta l_6\geq\frac{t}{2}\Big).$$

Following similar arguments above, let $t=b_6n^{-2\kappa}$ and $N=b_{17}d_n^{1/2}n^{-\kappa}$, for sufficiently large n,

$$P\Big(\Delta l_3+\Delta l_5\geq\frac{t}{2}\Big)\leq\exp(-c_4n^{1-2\kappa}d_n^{-1}),$$

for some positive constant c_4. By Lemma 9,

$$P(\Delta l_6\geq\frac{t}{2})\leq\exp(-c_5n^{1-2\kappa}d_n^{-2}),$$

for some positive constant c_5 and sufficiently large n.

If Condition H3 is satisfied, by Lemma 10,

$$P(\Delta l_6 \geq \frac{t}{2}) \leq 2\exp(-c_6 n^{1/2-\kappa} d_n^{-3/2}),$$

for some positive constant c_6 and sufficiently large n.

Next we consider the situation where $l''(\omega, Y)$ does not exist but $|l'(\omega, Y)|$ is bounded. By formula (S1), $W(\boldsymbol{\beta}_j^s)$ can be upper bounded by

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in B(N)} \Big\{ \frac{b_1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)^T E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M) \\
&\quad -\mathbb{P}_n(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)^T\boldsymbol{\Psi}_j \int_0^1 [l'\big(\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^M + t(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)), Y_i\big) - l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)]dt \Big\} \\
&+ \sup_{\boldsymbol{\beta}\in B(N)} \mathbb{P}_n\boldsymbol{\Psi}_j^T l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M) \\
\leq\ &\sup_{\boldsymbol{\beta}\in B(N)} \Big\{ \frac{b_1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)^T E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M) \\
&\quad -\mathbb{P}_n(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)^T\boldsymbol{\Psi}_j \int_0^1 E[l'\big(\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^M + t(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)), Y_i\big) - l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i]dt \Big\} \\
&+ \sup_{\boldsymbol{\beta}\in B(N)} \Big| \mathbb{P}_n(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)^T\boldsymbol{\Psi}_j \int_0^1 [l'\big(\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^M + t(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)), Y_i\big) - l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)]dt \\
&\quad -\mathbb{P}_n(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)^T\boldsymbol{\Psi}_j \int_0^1 E[l'\big(\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^M + t(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)), Y_i\big) - l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)|\mathbf{X}_i]dt \Big| \\
&+ \sup_{\boldsymbol{\beta}\in B(N)} \mathbb{P}_n\boldsymbol{\Psi}_j^T l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y_i)(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M) \\
\equiv\ &\Delta l_7 + \Delta l_8 + \Delta l_4.
\end{aligned}$$

By Lemma 1, the first term Δl_7 is further upper bounded by

$$\frac{b_1}{2} \sup_{\boldsymbol{\beta}\in B(N)} |(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)^T(E-\mathbb{P}_n)\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}_j^M)| = \Delta l_3.$$

Similar to (S9) and (S10), by Markov inequality and Hölder inequality, we have

$$P(W(\boldsymbol{\beta}_j^s) \geq t) \leq \exp(-rt)[E\exp(2r(\Delta l_3 + \Delta l_4)]^{1/2}[E\exp(2r\Delta l_8)]^{1/2}.$$

For Δl_8, by Lemma 13, with similar argument to (S12), we can also achieve the desired tail probability. Therefore, the proof of Lemma 14 is now complete.

Proof of Theorem 3.2.1: By the definition of $\boldsymbol{\beta}_j^s$ and the convexity of $l()$, we have

$$\mathbb{P}_n l(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^s, Y) \leq s\mathbb{P}_n l(\boldsymbol{\Psi}_j^T\widehat{\boldsymbol{\beta}}_j^M, Y) + (1-s)\mathbb{P}_n l(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y) \leq \mathbb{P}_n l(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y),$$

because $\widehat{\boldsymbol{\beta}}_j^M$ is the minimizer due to the empirical expectation. Therefore,

$$\mathbb{P}_n[l(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^s, Y) - l(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y)] \leq 0. \tag{S13}$$

By the definition of $\boldsymbol{\beta}_j^s$, we have

$$P\Big(\|\boldsymbol{\beta}_j^s - \boldsymbol{\beta}_j^M\| \geq \frac{N}{2}\Big) = P(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \geq N). \tag{S14}$$

Due to Lemma 3,

$$D_1\|\boldsymbol{\beta}_j^s - \boldsymbol{\beta}_j^M\|^2 d_n^{-1} \leq (\boldsymbol{\beta}_j^s - \boldsymbol{\beta}_j^M)^T E\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^T(\boldsymbol{\beta}_j^s - \boldsymbol{\beta}_j^M).$$

By the definition of $W(\boldsymbol{\beta}_j^s)$ in Lemma 14 and (S13), we have

$$\frac{b_1}{2}D_1\|\boldsymbol{\beta}_j^s - \boldsymbol{\beta}_j^M\|^2 d_n^{-1} \leq W(\boldsymbol{\beta}_j^s).$$

Therefore, for any constant $c_3 > 0$ and let $N = c_3^{1/2}d_n^{1/2}n^{-\kappa}$,

$$P\Big(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|^2 \geq c_3 d_n n^{-2\kappa}\Big) \leq P\Big(W(\boldsymbol{\beta}_j^s) \geq \frac{b_1 c_3 D_1 n^{-2\kappa}}{8}\Big).$$

Let us consider the situation where $l''(\omega, Y)$ exists and $|l'(\omega, Y)|$ is bounded for any bounded ω here. Discussions for other situations are similar as in Lemma 14. Since $\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M$ is bounded as shown in Lemma 14, let us assume $|l'(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, Y)| \leq M$ for some positive constant M. Apply Lemma 14, in (S12), replace b_{17} by $c_3^{1/2}$ and b_6 by $b_1D_1c_3/8$. For any $0 < \epsilon < 1$, choose sufficient small b_{16} such that

$$(8M^2 + 2b_1^2c_3d_nn^{-2\kappa})b_{16} \leq \epsilon/2.$$

For any constant $c_3 > 0$ and some $0 < \epsilon < 1$, choose sufficiently large sample size n and corresponding d_n such that

$$(C_2)^{1/2}Md_n^{1/2}n^{\kappa-1/2} + 0.5(C_2)^{1/2}b_1c_3^{1/2}d_nn^{-1/2} < (1-\epsilon)\frac{b_1D_1c_3^{1/2}}{8},$$

where b_1, C_2 and D_1 are defined in Lemmas 1, 2, 3 respectively. Correspondingly, (S12) will be valid. Therefore, the conclusion in Theorem 3.2.1 will be correct. The proof of Theorem 3.2.1 is now complete.

LEMMA 15. *For Types 1-3, if Condition H3 is satisfied, for any $b_3 > 0$, there exists a constant $b_4 > 0$ such that*

$$P(|\widehat{\beta}_0^M - \beta_0^M| \geq b_3 n^{-\kappa}) \leq 2\exp(-b_4 n^{1-2\kappa})$$

for sufficiently large n.

Proof of Lemma 15: The proof of Lemma 15 has some similarity to the proof of Theorem 3.2.1, but the details and conclusion have some differences. Let $\beta_0^s = \widehat{\beta}_0^M + (1-s)\beta_0^M$ where $s = (1+|\widehat{\beta}_0^M - \beta_0^M|/N)^{-1}$ where N is some constant. It is easy to show that $|\beta_0^s - \beta_0^M| \leq N$. Because $l()$ is a convex function,

$$\mathbb{P}_n l(\beta_0^s, Y) \leq s\mathbb{P}_n l(\widehat{\beta}_0^M, Y) + (1-s)\mathbb{P}_n l(\beta_0^M, Y) \leq \mathbb{P}_n l(\beta_0^M, Y)$$

by the definition of $\widehat{\beta}_0^M$. Therefore, $\mathbb{P}_n[l(\beta_0^s, Y) - l(\beta_0^M, Y)] \leq 0$. Let $W(\beta) = \frac{b_1}{2}(\beta - \beta_0^M)^2 - \mathbb{P}_n[l(\beta_0^s, Y_i) - l(\beta_0^M, Y_i)]$, define $\widetilde{B}(N) = \{\beta : |\beta - \beta_0^M| \leq N\}$. We first consider Type 1 and 2 where $l''(\omega, Y)$ exists. The general decomposition (S1) can be further replaced by a Taylor expansion such that

$$\begin{aligned}
& W(\beta_0^s) \\
= \; & \frac{b_1}{2}(\beta_0^s - \beta_0^M)^2 - \mathbb{P}_n \int_0^1 l''(\beta_0^M + t(\beta_0^s - \beta_0^M), Y_i)(\beta_0^s - \beta_0^M)^2(1-t)dt \\
& -\mathbb{P}_n l'(\beta_0^M, Y_i)(\beta_0^s - \beta_0^M) \\
\leq \; & \sup_{\beta \in \widetilde{B}(N)} \left\{ \frac{b_1}{2}(\beta - \beta_0^M)^2 - \mathbb{P}_n \int_0^1 l''(\beta_0^M + t(\beta - \beta_0^M), Y_i)(\beta - \beta_0^M)^2(1-t)dt \right\} \\
& + \sup_{\beta \in \widetilde{B}(N)} |\beta - \beta_0^M| \times \left| \mathbb{P}_n l'(\beta_0^M, Y_i) \right|
\end{aligned}$$

By the definition of conditional strictly convex loss, the first term of the last line above is upper bounded by

$$\sup_{\beta \in \widetilde{B}(N)} \left\{ \frac{b_1}{2}(\beta - \beta_0^M)^2 - \mathbb{P}_n b_1 (\beta - \beta_0^M)^2 \int_0^1 (1-t)dt \right\} = 0. \tag{S15}$$

Therefore, let $N = b_3 n^{-\kappa}$, we have

$$\begin{aligned}
P\left(|\beta_0^s - \beta_0^M|^2 \geq \frac{b_3^2}{4} n^{-2\kappa}\right) &\leq P\left(W(\beta_0^s) \geq b_1 \frac{b_3^2}{8} n^{-2\kappa}\right) \\
&\leq P\left(N|\mathbb{P}_n l'(\beta_0^M, Y_i)| \geq b_1 \frac{b_3^2}{8} n^{-2\kappa}\right)
\end{aligned}$$

$$= \quad P\Big(|\mathbb{P}_n l'(\beta_0^M, Y_i) - El'(\beta_0^M, Y)| \geq b_1 \frac{b_3}{8} n^{-\kappa}\Big)$$

For the last expression of the above line, we have used $El'(\beta_0^M, Y) = 0$ by the definition of β_0^M.

When $E[\exp(\lambda l'(\beta_0^M, Y))|\mathbf{X}] < \infty$ for all $|\lambda| < c_0$, we have $E\exp(\lambda l'(\beta_0^M, Y)) < \infty$ for all $|\lambda| < c_0$. Therefore, by Lemma 6, $l'(\beta_0^M, Y)$ is sub-exponential random variable. By Theorem 5.1 in Buldygin & Kozachenko (2000), there exists some positive constant b_4 such that

$$P\Big(|\mathbb{P}_n l'(\beta_0^M, Y_i) - El'(\beta_0^M, Y)| \geq b_1 \frac{b_3}{8} n^{-\kappa}\Big) \leq 2\exp(-b_4 n^{1-2\kappa})$$

for sufficiently large n. Note that

$$P(|\widehat{\beta}_0 - \beta_0^M| \geq b_3 n^{-\kappa}) = P(|\beta_0^s - \beta_0^M| \geq \frac{b_3}{2} n^{-\kappa})$$

by the definition of β_0^s, we can achieve the desired tail probability.

For Type 3 where $l''(\omega, Y)$ does not exist, by the general decomposition (S1),

$$\begin{aligned}
& W(\beta_0^s) \\
\leq \quad & \sup_{\beta \in \widetilde{B}(N)} \Big\{ \frac{b_1}{2}(\beta - \beta_0^M)^2 - \mathbb{P}_n(\beta - \beta_0^M) \int_0^1 [l'(\beta_0^M + t(\beta - \beta_0^M), Y_i) - l'(\beta_0^M, Y_i)]dt \Big\} \\
& + \sup_{\beta \in \widetilde{B}(N)} |\beta - \beta_0^M| \times |\mathbb{P}_n l'(\beta_0^M, Y_i)| \\
\leq \quad & \sup_{\beta \in \widetilde{B}(N)} \Big\{ \frac{b_1}{2}(\beta - \beta_0^M)^2 - \mathbb{P}_n(\beta - \beta_0^M) \int_0^1 E[l'(\beta_0^M + t(\beta - \beta_0^M), Y_i) - l'(\beta_0^M, Y_i)|\mathbf{X}_i]dt \Big\} \\
& + \sup_{\beta \in \widetilde{B}(N)} \Big| \mathbb{P}_n(\beta - \beta_0^M) \int_0^1 [l'(\beta_0^M + t(\beta - \beta_0^M), Y_i) - l'(\beta_0^M, Y_i)]dt \\
& \qquad\qquad - \mathbb{P}_n(\beta - \beta_0^M) \int_0^1 E[l'(\beta_0^M + t(\beta - \beta_0^M), Y_i) - l'(\beta_0^M, Y_i)|\mathbf{X}_i]dt \Big| \\
& + N|\mathbb{P}_n l'(\beta_0^M, Y_i)|
\end{aligned}$$

Similar to (S15), the first term is upper bounded by 0 due to the definition of conditional strictly convex loss and Lemma 1. For the second term and the third term, similar to the arguments in Lemmas 10, 13 and 14, we can also achieve the desired tail probability for sufficiently large n. The proof of Lemma 15 is now complete.

Proof of Theorem 3.3.1: We first consider Part (ii) Type 3 with $l'(\omega, Y) = \mathbf{I}(Y - \omega < 0) - \alpha$ and use the property that $l'()$ is bounded and $l'(\omega, Y)$ is nondecreasing in ω. By the definition of $G_{n,j}$ and $G_j^\star$,

$$\begin{aligned} & G_{n,j} - G_j^\star \\ = \; & [\mathbb{P}_n l(\widehat{\beta}_0^M, Y) - \mathbb{P}_n l(\beta_0^M, Y)] - [\mathbb{P}_n l(\boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M, Y) - \mathbb{P}_n l(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)] \\ & - \big\{[\mathbb{P}_n l(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y) - \mathbb{P}_n l(\beta_0^M, Y)] - [El(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y) - El(\beta_0^M, Y)]\big\} \\ \equiv \; & S_1 - S_2 - (S_3 - S_4). \end{aligned}$$

We will prove that $|S_1| + |S_2| + |S_3 - S_4|$ is upper bounded by $c_3 d_n n^{-2\kappa}$ with probability converging to 1. First consider $|S_3 - S_4|$. By the general decomposition (S1),

$$|S_3 - S_4| = \left| (\mathbb{P}_n - E)(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M - \beta_0^M) \int_0^1 l'(\beta_0^M + t(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M - \beta_0^M), Y) dt \right|.$$

Since $|l'()|$ is bounded, and $(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M - \beta_0^M)$ is also bounded, by Hoeffding's inequality, for any constant $c_5 > 0$, there exists $c_6 > 0$ such that

$$P(|S_3 - S_4| \geq c_5 n^{-\kappa}) \leq 2\exp(-c_6 n^{1-2\kappa}),$$

which means that $|S_3 - S_4|$ is upper bounded by $c_5 n^{-\kappa}$ except on an event with probability at most $2\exp(-c_6 n^{1-2\kappa})$. Next for S_1, by the general decomposition (S1), we have

$$\begin{aligned} S_1 = \; & \mathbb{P}_n l'(\beta_0^M, Y)(\widehat{\beta}_0^M - \beta_0^M) \\ & + \mathbb{P}_n(\widehat{\beta}_0^M - \beta_0^M) \int_0^1 [l'(\beta_0^M + t(\widehat{\beta}_0^M - \beta_0^M), Y) - l'(\beta_0^M, Y)] dt. \end{aligned} \tag{S16}$$

In the first term of (S16), since $|l'()|$ is bounded, by Hoeffding's inequality, for any $c_7 > 0$, there exists $c_8 > 0$ such that

$$\begin{aligned} P(|\mathbb{P}_n l'(\beta_0^M, Y)| \geq c_7 n^{-\kappa}) &= P(|\mathbb{P}_n l'(\beta_0^M, Y) - El'(\beta_0^M, Y)| \geq c_7 n^{-\kappa}) \\ &\leq 2\exp(-c_8 n^{1-2\kappa}). \end{aligned}$$

where the first equality is by $El'(\beta_0^M, Y) = 0$ due to the definition of β_0^M. Therefore, by Lemma 15, we have for any $c_9 > 0$, there exists $c_{10} > 0$ such that

$$P(|\mathbb{P}_n l'(\beta_0^M, Y)(\widehat{\beta}_0^M - \beta_0^M)| \geq c_9 n^{-2\kappa}) \leq 2\exp(-c_8 n^{1-2\kappa}) + 2\exp(-b_4 n^{1-2\kappa}).$$

For the second term of (S16), since $l'(\omega, Y)$ is non-decreasing in ω, the second term is non-negative. Furthermore, on the event $\{|\widehat{\beta}_0^M - \beta_0^M| \leq b_3 n^{-\kappa}\}$,

$$\begin{aligned} &\mathbb{P}_n(\widehat{\beta}_0^M - \beta_0^M) \int_0^1 [l'(\beta_0^M + t(\widehat{\beta}_0^M - \beta_0^M), Y) - l'(\beta_0^M, Y)]dt \\ \leq\ & \mathbb{P}_n b_3 n^{-\kappa} \int_0^1 [l'(\beta_0^M + t b_3 n^{-\kappa}, Y) - l'(\beta_0^M, Y)]dt. \end{aligned} \tag{S17}$$

In the last line, since $l'()$ is bounded, by the Hoeffding's inequality,

$$\begin{aligned} & P(|(\mathbb{P}_n - E) \int_0^1 [l'(\beta_0^M + t b_3 n^{-\kappa}, Y) - l'(\beta_0^M, Y)]dt| > c_{11} n^{-\kappa}) \\ \leq\ & 2\exp(-c_{12} n^{1-2\kappa}). \end{aligned}$$

Therefore, expression (S17) can be further upper bounded by

$$b_3 n^{-\kappa} E \int_0^1 [l'(\beta_0^M + t b_3 n^{-\kappa}, Y) - l'(\beta_0^M, Y)]dt + b_3 c_{11} n^{-2\kappa}$$

except on an event with probability at most $2\exp(-c_{12} n^{1-2\kappa})$. By Lemma 1 the last line is further upper bounded by $(b_2 b_3^2/2 + b_3 c_{11}) n^{-2\kappa}$. In summary,

$$P(|S_1| \geq c_{13} n^{-2\kappa}) \leq 4\exp(-c_{14} n^{1-2\kappa}) + 2\exp(-b_4 n^{1-2\kappa}).$$

For S_2, by the general decomposition (S1), S_2 can be expressed as

$$\begin{aligned} &\mathbb{P}_n l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y) \boldsymbol{\Psi}_j^T (\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M) + \\ &\mathbb{P}_n \boldsymbol{\Psi}_j^T (\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M) \int_0^1 [l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M + t \boldsymbol{\Psi}_j^T (\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M), Y) - l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)]dt. \end{aligned}$$

Note that $\boldsymbol{\Psi}_j^T (\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M) \leq \|\boldsymbol{\Psi}_j\| \|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \leq \|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\|$. Since $l'()$ is bounded, by Hoeffding's inequality and the definition of $\boldsymbol{\beta}_j^M$,

$$\begin{aligned} & P(|\mathbb{P}_n l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)| \geq c_{15} n^{-\kappa}) \\ =\ & P(|\mathbb{P}_n l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y) - E l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)| \geq c_{15} n^{-\kappa}) \\ \leq\ & 2\exp(-c_{16} n^{1-2\kappa}). \end{aligned}$$

Therefore, by Theorem 3.2.1, on the event that $\{\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \leq c_3^{1/2} d_n^{1/2} n^{-\kappa}\}$,

$$|\mathbb{P}_n l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y) \boldsymbol{\Psi}_j^T (\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M)| \leq c_{14} c_3^{1/2} d_n^{1/2} n^{-2\kappa}$$

except on an event with probability at most $\exp(-c_4 n^{1-2\kappa} d_n^{-1}) + 2\exp(-c_{16} n^{1-2\kappa})$, where we have used (S14) and the result in Theorem 3.2.1.

For the second term in S_2, on the event $\{\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \le c_3^{1/2} d_n^{1/2} n^{-\kappa}\}$, we have

$$\begin{aligned}
&\mathbb{P}_n(\boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M) \int_0^1 \big[l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M + t(\boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M), Y) \\
&\qquad\qquad -l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)\big] dt \\
\le\ & \mathbb{P}_n c_3^{1/2} d_n^{1/2} n^{-\kappa} \int_0^1 [l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M + t c_3^{1/2} d_n^{1/2} n^{-\kappa}, Y) - l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)] dt
\end{aligned}$$

since $l'(\omega, Y)$ is a non-decreasing function in ω. The last line is further upper bounded by

$$\begin{aligned}
&E c_3^{1/2} d_n^{1/2} n^{-\kappa} \int_0^1 [l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M + t c_3^{1/2} d_n^{1/2} n^{-\kappa}, Y) \\
&\qquad\qquad -l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)] dt \\
&+ c_3^{1/2} c_{17} d_n^{1/2} n^{-2\kappa}
\end{aligned} \tag{S18}$$

except on an event with probability at most $2\exp(-c_{18} n^{1-2\kappa})$ by Hoeffding's inequality. By Lemma 1, (S18) is further upper bounded by $(b_2 c_3 + c_3^{1/2} c_{17}) n^{-2\kappa} d_n$. To summarize,

$$P(|S_2| \ge c_{19} d_n n^{-2\kappa}) \le 4\exp(-c_{20} n^{1-2\kappa}) + \exp(-c_4 n^{1-2\kappa} d_n^{-1}).$$

By condition E, $d_n^{-1} = O(n^{-\kappa})$. For simplicity of discussion, without loss of generality, suppose $d_n^{-1} \le n^{-\kappa}$. Hence we have $d_n n^{-2\kappa} \ge n^{-2\kappa}$, and $d_n n^{-2\kappa} \ge n^{-\kappa}$. Finally,

$$P(|G_{n,j} - G_j^\star| > c_{21} d_n n^{-2\kappa}) \le \exp(-c_4 n^{1-2\kappa} d_n^{-1}) + 12\exp(-c_{22} n^{1-2\kappa}).$$

By Lemma 3.1.2, we have $\min_{j \in M_\star} G_j^\star \ge \frac{b^\star}{2} c_1 \xi d_n n^{-2\kappa}$. On the event

$$A_n := \Big\{ \max_{j \in M_\star} |G_{n,j} - G_j^\star| \le \frac{b^\star}{4} c_1 \xi d_n n^{-2\kappa} \Big\},$$

we have $G_{n,j} \ge \frac{b^\star}{4} c_1 \xi d_n n^{-2\kappa}, \forall j \in M_\star$. Hence by the choice of ν_n, we have $M_\star \subset \widehat{M}_{\nu_n}$. The result now follows from $P(M_\star \subset \widehat{M}_{\nu_n}) \ge P(A_n) = 1 - P(A_n^c)$. The proof for Theorem 3.3.1 (ii) is now complete.

For Part (i) Types 1 & 2, we use the property that $l''(\omega, Y)$ is continuous in ω, $l''(\omega, Y) > 0$ and $l''(\omega, Y)$ is bounded when ω is bounded, and the

analysis is different. We want to show that there exists some constant $\zeta > 0$ such that $P(G_{n,j} > \zeta d_n n^{-2\kappa}) \to 1$ for $j \in M_\star$. Note that by the definition of $G_{n,j}$ and Taylor expansion,

$$\begin{aligned} G_{n,j} &= \mathbb{P}_n l'(\boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M, Y)(\widehat{\beta}_0^M - \boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M) \\ &\quad + \mathbb{P}_n \int_0^1 l''(\boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M + t(\widehat{\beta}_0^M - \boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M), Y)(\widehat{\beta}_0^M - \boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M)^2 (1-t) dt. \end{aligned}$$

The first term is 0 because of the definition of $\widehat{\boldsymbol{\beta}}_j^M$. Therefore, we will find a lower bound of the second term. Since $l''(\omega, Y)$ is continuous in ω and $l'' > 0$, for a bounded ω, there exists positive constants r_1 and r_2 such that $r_1 \le l''(\omega, Y) \le r_2$. On the event that $\{|\widehat{\beta}_0^M - \beta_0^M| \le r_3 n^{-\kappa}\}$ and $\{\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \le r_4 d_n^{1/2} n^{-\kappa}\}$, there exists positive constants r_5 and r_6 such that

$$0 < r_5 \le \frac{l''(\boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M + t(\widehat{\beta}_0^M - \boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M), Y)}{l''(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M + t(\beta_0^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M), Y)} \le r_6$$

for all $1 \le j \le p_n$. The restrictions of r_3 and r_4 will be determined later. Correspondingly,

$$G_{n,j} \ge r_5 \mathbb{P}_n \int_0^1 l''(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M + t(\beta_0^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M), Y)(\widehat{\beta}_0^M - \boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M)^2 (1-t) dt.$$

Use binomial expansion for $(\widehat{\beta}_0^M - \boldsymbol{\Psi}_j^T \widehat{\boldsymbol{\beta}}_j^M)^2 = [\beta_0^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M + \widehat{\beta}_0^M - \beta_0^M - \boldsymbol{\Psi}_j^T (\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M)]^2$. Let

$$\begin{aligned} P_1 &\equiv \mathbb{P}_n \int_0^1 l''(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M + t(\beta_0^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M), Y)(\beta_0^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M)^2 (1-t) dt, \\ P_2 &\equiv \mathbb{P}_n \int_0^1 \Big[l''(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M + t(\beta_0^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M), Y)(\beta_0^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M) \\ &\qquad \times (\widehat{\beta}_0^M - \beta_0^M - \boldsymbol{\Psi}_j^T (\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M))(1-t) \Big] dt, \\ P_3 &\equiv \mathbb{P}_n \int_0^1 \Big[l''(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M + t(\beta_0^M - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M), Y) \\ &\qquad \times (\widehat{\beta}_0^M - \beta_0^M - \boldsymbol{\Psi}_j^T (\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M))^2 (1-t) \Big] dt, \end{aligned}$$

then on the event $\{|\widehat{\beta}_0^M - \beta_0^M| \le r_3 n^{-\kappa}\}$ and $\{\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \le r_4 d_n^{1/2} n^{-\kappa}\}$, we have

$$G_{n,j} \ge r_5 (P_1 + 2P_2 + P_3). \tag{S19}$$

By Hoeffding's inequality,

(S20)
$$P_1 \geq E\int_0^1 l''(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M + t(\beta_0^M - \boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M), Y)(\beta_0^M - \boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M)^2(1-t)dt - r_7 n^{-\kappa}$$

except on an event with probability at most $2\exp(-c_{23}n^{1-2\kappa})$. Note that by the definition of $G_j^\star$ and apply Taylor expansion,

$$G_j^\star = E\int_0^1 l''(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M + t(\beta_0^M - \boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M), Y)(\beta_0^M - \boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M)^2(1-t)dt,$$

which is exactly the first term in (S20). By conditions E, $d_n^{-1} = O(n^{-\kappa})$. To simplify the notation, without loss of generality, we can use $d_n \geq n^\kappa$. Therefore, (S20) further implies

$$P_1 \geq G_j^\star - r_7 d_n n^{-2\kappa}. \tag{S21}$$

For P_2, on the event $\{|\widehat{\beta}_0^M - \beta_0^M| \leq r_3 n^{-\kappa}\}$ and $\{\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \leq r_4 d_n^{1/2} n^{-\kappa}\}$, there exists some constant r_8 such that $|\widehat{\beta}_0^M - \beta_0^M - \boldsymbol{\Psi}_j^T(\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M)| \leq r_8^{1/2} d_n^{1/2} n^{-\kappa}$. Therefore, $|P_2|$ is upper bounded by

(S22)
$$\Big\{E\int_0^1 l''(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M + t(\beta_0^M - \boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M), Y)|\beta_0^M - \boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M| r_8^{1/2} d_n^{1/2} n^{-\kappa}(1-t)dt + r_9 n^{-\kappa}\Big\}$$

except on an event with probability at most $2\exp(-c_{24}n^{1-2\kappa})$, where we apply the Hoeffding's inequality. Apply Cauchy Schwartz inequality to (S22), we further have that (S22) is upper bounded by

(S23)
$$\begin{aligned}&\Big\{[E\int_0^1 l''(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M + t(\beta_0^M - \boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M), Y)(\beta_0^M - \boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M)^2(1-t)dt]^{1/2}\\ &\quad\times[E\int_0^1 l''(\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M + t(\beta_0^M - \boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M), Y) r_8 d_n n^{-2\kappa}(1-t)dt]^{1/2}\\ &\qquad + r_9 n^{-\kappa}\Big\}\\ &\quad\leq \Big((G_j^\star)^{1/2} r_{10}^{1/2} d_n^{1/2} n^{-\kappa} + r_9 d_n n^{-2\kappa}\Big).\end{aligned}$$

for some constant r_{10}, where we use the fact that $l''(\omega, Y)$ is bounded when ω is bounded. For $j \in M_\star$, $G_j^\star \geq c d_n n^{-2\kappa}$ by Lemma 3.1.2 for some positive constant c to simplify the notation. On the event $\{|\widehat{\beta}_0^M - \beta_0^M| \leq r_3 n^{-\kappa}\}$ and

$\{\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \le r_4 d_n^{1/2} n^{-\kappa}\}$ where r_3 and r_4 are appropriately chosen and r_7, r_9 and r_{10} are sufficiently small. Based on (S21) and (S23), for $j \in M_\star$,

$$G_{n,j} \ge r_5(c - 2c^{1/2} r_{10}^{1/2} - r_7 - 2r_9) d_n n^{-2\kappa} > 0$$

except on an event with probability at most $P(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \ge r_4 d_n^{1/2} n^{-\kappa}) + 6\exp(-c_{25} n^{1-2\kappa})$. If we use threshold $\nu_n = \nu d_n n^{-2\kappa}$ where $\nu \le r_5(c - 2c^{1/2} r_{10}^{1/2} - r_7 - 2r_9)$, then

$$\begin{aligned} P(M_\star \subset \widehat{M}_{\nu_n}) &= 1 - P(M_\star \not\subset \widehat{M}_{\nu_n}) \\ &\ge 1 - \sum_{j \in M_\star} P(G_{n,j} < \nu_n) \\ &\ge 1 - s_0 \Big\{ P(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \ge r_4 d_n^{1/2} n^{-\kappa}) + 6\exp(-c_{25} n^{1-2\kappa}) \Big\}, \end{aligned}$$

where in the last expression, we can further apply the results in Theorem 3.2.1. For the consideration of model selection size control in Theorem 3.4.1, with sufficiently small r_7, r_9 and r_{10}, we further require that

$$(\frac{\nu}{r_6} - r_7 - 2r_9)^{1/2} - r_{10}^{1/2} \ge h^{1/2} > 0 \tag{S24}$$

for some positive constant h. The proof of Theorem 3.3.1 (i) is now complete.

Proof of Theorem 3.4.1: Let $M(\omega, \mathbf{X}) = E[l'(\omega, Y)|\mathbf{X}]$. By definition of conditional strictly convex loss, we have

$$\{M(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, \mathbf{X}) - M(E\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, \mathbf{X})\}[\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M - E\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M] \ge b_1 (\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M - E\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M)^2. \tag{S25}$$

By the score equation $EM(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, \mathbf{X})\boldsymbol{\Psi}_j^T = 0$, we have

$$EM(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, \mathbf{X})\boldsymbol{\Psi}_j^T [\boldsymbol{\beta}_j^M - \mathbf{1} E\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M] = 0.$$

where $\mathbf{1} = (1, \cdots, 1)^T$ with the same dimension as $\boldsymbol{\beta}_j^M$. By taking the expectation on both sides of (S25), we can obtain

$$\begin{aligned} & -EM(E\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, \mathbf{X})\boldsymbol{\Psi}_j^T [\boldsymbol{\beta}_j^M - \mathbf{1} E\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M] \\ \ge\ & b_1 (\boldsymbol{\beta}_j^M - \mathbf{1} E\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M)^T E(\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T)(\boldsymbol{\beta}_j^M - \mathbf{1} E\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M), \end{aligned}$$

where we have used the partition of unity. Applying Lemma 3 to the right side of the last line implies that

$$b_1 D_1 d_n^{-1} \|\boldsymbol{\beta}_j^M - \mathbf{1} E\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M\|^2 \le -EM(E\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, \mathbf{X})\boldsymbol{\Psi}_j^T [\boldsymbol{\beta}_j^M - \mathbf{1} E\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M]. \tag{S26}$$

By the definition of M function and the definition of conditional derivative separable loss,

$$\begin{aligned} & -EM(E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M, \mathbf{X})\boldsymbol{\Psi}_j^T[[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M] \\ = \; & E[H(\mathbf{X})K(E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M) - G(E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M)]\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M] \\ = \; & K(E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M)E[H(\mathbf{X})\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M]], \end{aligned} \tag{S27}$$

where in the second step we have used the fact that $E\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M] = 0$.

Denote $\boldsymbol{\beta}_j^\star$ as the coefficient vector for $\boldsymbol{\Psi}_j^T$ in the joint regression model. Let $\boldsymbol{\beta}^\star = ((\boldsymbol{\beta}_1^\star)^T, \cdots, (\boldsymbol{\beta}_{p_n}^\star)^T)^T$. Let $\boldsymbol{\Psi} = (\Psi_1(X_1), \cdots, \Psi_{d_n}(X_1), \cdots, \Psi_1(X_{p_n}), \cdots, \Psi_{d_n}(X_{p_n}))^T$ be the basis functions in the joint regression model. Define $\mathcal{S}_{1,j}$ as the set that $\boldsymbol{\Psi}^T\boldsymbol{\beta}^\star$ and $\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M]$ have the same sign, and $\mathcal{S}_{2,j}$ as the set that $\boldsymbol{\Psi}^T\boldsymbol{\beta}^\star$ and $\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M]$ have the opposite sign. By the definition of conditional strictly convex loss,

$$\begin{aligned} & (M(\beta_0^\star + \boldsymbol{\Psi}^T\boldsymbol{\beta}^\star, \mathbf{X}) - M(\beta_0^\star, \mathbf{X}))\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M]\mathbf{I}_{\mathcal{S}_{1,j}} \\ \leq \; & b_2\boldsymbol{\Psi}^T\boldsymbol{\beta}^\star\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M]\mathbf{I}_{\mathcal{S}_{1,j}}; \\ & (M(\beta_0^\star + \boldsymbol{\Psi}^T\boldsymbol{\beta}^\star, \mathbf{X}) - M(\beta_0^\star, \mathbf{X}))\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M]\mathbf{I}_{\mathcal{S}_{2,j}} \\ \leq \; & b_1\boldsymbol{\Psi}^T\boldsymbol{\beta}^\star\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M]\mathbf{I}_{\mathcal{S}_{2,j}} \end{aligned}$$

where $\mathbf{I}$ is the indicator function. Taking the summation of the above two expression, we have

$$\begin{aligned} & (M(\beta_0^\star + \boldsymbol{\Psi}^T\boldsymbol{\beta}^\star, \mathbf{X}) - M(\beta_0^\star, \mathbf{X}))\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M] \\ \leq \; & b_1\boldsymbol{\Psi}^T\boldsymbol{\beta}^\star\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M] \\ & +(b_2 - b_1)\boldsymbol{\Psi}^T\boldsymbol{\beta}^\star\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M]\mathbf{I}_{\mathcal{S}_{1,j}}. \end{aligned} \tag{S28}$$

By the score equation of the joint model $EM(\beta_0^\star + \boldsymbol{\Psi}^T\boldsymbol{\beta}^\star, \mathbf{X})\boldsymbol{\Psi}_j^T = \mathbf{0}$, we have

$$EM(\beta_0^\star + \boldsymbol{\Psi}^T\boldsymbol{\beta}^\star, \mathbf{X})\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M] = 0.$$

Taking expectation on both sides of (S28) leads to

$$\begin{aligned} & -E\big[M(\beta_0^\star, \mathbf{X})\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M]\big] \\ \leq \; & b_1E\big[\boldsymbol{\Psi}^T\boldsymbol{\beta}^\star\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M]\big] \\ & +(b_2 - b_1)E\big[\boldsymbol{\Psi}^T\boldsymbol{\beta}^\star\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M]\mathbf{I}_{\mathcal{S}_{1,j}}\big]. \end{aligned} \tag{S29}$$

By the definition of M function,

$$-E\big[M(\beta_0^\star, \mathbf{X})\boldsymbol{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\boldsymbol{\Psi}_j^T\boldsymbol{\beta}_j^M]\big] \tag{S30}$$

$$
\begin{aligned}
&= E[H(\mathbf{X})K(\beta_0^\star) - G(\beta_0^\star)]\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M] \\
&= K(\beta_0^\star)E[H(\mathbf{X})\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]],
\end{aligned}
$$

For a generic vector $\mathbf{b} = (b_1, \cdots, b_p)^T$, we define the notation $|\mathbf{b}| = (|b_1|, \cdots, |b_p|)^T$. Combining (S29) and (S30), we have

$$
\begin{aligned}
&K(\beta_0^\star)E[H(\mathbf{X})\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]] \\
\leq\ & b_1|(\boldsymbol{\beta}^\star)^T E\mathbf{\Psi}\mathbf{\Psi}_j^T| \cdot |[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]| \\
&+(b_2 - b_1)|(\boldsymbol{\beta}^\star)^T E\mathbf{\Psi}\mathbf{\Psi}_j^T| \cdot |[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]| \\
=\ & b_2|(\boldsymbol{\beta}^\star)^T E\mathbf{\Psi}\mathbf{\Psi}_j^T| \cdot |[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]|.
\end{aligned} \tag{S31}
$$

Similar to the argument in (S28), we can also show that

$$
\begin{aligned}
&[M(\beta_0^\star, \mathbf{X}) - M(\beta_0^\star + \mathbf{\Psi}^T\boldsymbol{\beta}^\star, \mathbf{X})]\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M] \\
\leq\ & -b_1\mathbf{\Psi}^T\boldsymbol{\beta}^\star\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]\mathbf{I}_{\mathcal{S}_{1,j}} \\
&-b_2\mathbf{\Psi}^T\boldsymbol{\beta}^\star\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]\mathbf{I}_{\mathcal{S}_{2,j}} \\
=\ & -b_1\mathbf{\Psi}^T\boldsymbol{\beta}^\star\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M] \\
&-(b_2 - b_1)\mathbf{\Psi}^T\boldsymbol{\beta}^\star\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]\mathbf{I}_{\mathcal{S}_{2,j}}.
\end{aligned} \tag{S32}
$$

Applying expectation on both sides of (S32), by the definition of M function, we have

$$
\begin{aligned}
&-K(\beta_0^\star)E[H(\mathbf{X})\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]] \\
\leq\ & -b_1E[\mathbf{\Psi}^T\boldsymbol{\beta}^\star\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]] \\
&-(b_2 - b_1)E[\mathbf{\Psi}^T\boldsymbol{\beta}^\star\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]\mathbf{I}_{\mathcal{S}_{2,j}}] \\
\leq\ & b_2|(\boldsymbol{\beta}^\star)^T E\mathbf{\Psi}\mathbf{\Psi}_j^T| \cdot |[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]|,
\end{aligned} \tag{S33}
$$

where the last step is similar to the argument in (S31). Combining (S26), (S27), (S31) and (S33), since $K(\beta_0^\star) \neq 0$ from the condition of Theorem 3.4.1, we have

(S34)

$$
b_1D_1d_n^{-1}\|[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]\|^2 \leq \Big|\frac{K(E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M)}{K(\beta_0^\star)}\Big| b_2|(\boldsymbol{\beta}^\star)^T E\mathbf{\Psi}\mathbf{\Psi}_j^T| \cdot |[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]|.
$$

Define $\boldsymbol{\beta}^M = ((\boldsymbol{\beta}_1^M - \mathbf{1}E\mathbf{\Psi}_1^T\boldsymbol{\beta}_1^M)^T, \cdots, (\boldsymbol{\beta}_{p_n}^M - \mathbf{1}E\mathbf{\Psi}_{p_n}^T\boldsymbol{\beta}_{p_n}^M)^T)^T$. Let

$$
C = \max_{1 \leq j \leq p_n} \Big|\frac{K(E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M)}{K(\beta_0^\star)}\Big|.
$$

Taking summation for (S34) yields

$$
\begin{aligned}
\text{(S35)} \qquad & b_1 D_1 d_n^{-1} \sum_{j=1}^{p_n} \|[\boldsymbol{\beta}_j^M - \mathbf{1} E \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M]\|^2 \\
\leq\ & C b_2 |(\boldsymbol{\beta}^\star)^T E \boldsymbol{\Psi}\boldsymbol{\Psi}^T| \cdot |\boldsymbol{\beta}^M| \\
\leq\ & C b_2 \|(\boldsymbol{\beta}^\star)^T E \boldsymbol{\Psi}\boldsymbol{\Psi}^T\| \cdot \|\boldsymbol{\beta}^M\|,
\end{aligned}
$$

where in the last step we apply the Cauchy Schwartz inequality. Note that $\|\boldsymbol{\beta}^M\| = (\sum_{j=1}^{p_n} \|[\boldsymbol{\beta}_j^M - \mathbf{1} E \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M]\|^2)^{1/2}$, therefore, (S35) implies

$$
\begin{aligned}
& \sum_{j=1}^{p_n} \|[\boldsymbol{\beta}_j^M - \mathbf{1} E \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M]\|^2 \\
\leq\ & D_7 d_n^2 \|(\boldsymbol{\beta}^\star)^T E \boldsymbol{\Psi}\boldsymbol{\Psi}^T\|^2 \\
=\ & D_7 d_n^2 (\boldsymbol{\beta}^\star)^T (E \boldsymbol{\Psi}\boldsymbol{\Psi}^T)(E \boldsymbol{\Psi}\boldsymbol{\Psi}^T) \boldsymbol{\beta}^\star \\
\text{(S36)} \qquad \leq\ & D_7 d_n^2 \lambda_{\max}(E \boldsymbol{\Psi}\boldsymbol{\Psi}^T) E(\boldsymbol{\Psi}^T \boldsymbol{\beta}^\star)^2.
\end{aligned}
$$

Next we will show that $\sum_{j=1}^{p_n} E(f_{nj}^M - E f_{nj}^M)^2 = O(d_n \lambda_{\max}(\boldsymbol{\Sigma}))$ where $\boldsymbol{\Sigma} = E\boldsymbol{\Psi}\boldsymbol{\Psi}^T$. Note that

$$
\begin{aligned}
\sum_{j=1}^{p_n} E(f_{nj}^M - E f_{nj}^M)^2 &= \sum_{j=1}^{p_n} E\big[([\boldsymbol{\beta}_j^M - \mathbf{1} E \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M])^T \boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T [\boldsymbol{\beta}_j^M - \mathbf{1} E \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M]\big] \\
&\leq \sum_{j=1}^{p_n} \|[\boldsymbol{\beta}_j^M - \mathbf{1} E \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M]\|^2 \|E \boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^T\| \\
&\leq D_7 d_n^2 \lambda_{\max}(E \boldsymbol{\Psi}\boldsymbol{\Psi}^T) E(\boldsymbol{\Psi}^T \boldsymbol{\beta}^\star)^2 D_2 d_n^{-1} \\
\text{(S37)} \qquad &= O(d_n \lambda_{\max}(\boldsymbol{\Sigma})).
\end{aligned}
$$

In the third step we use the result in Lemma 3 and the result in (S36), and in the fourth step we use the condition that $E(\boldsymbol{\Psi}^T \boldsymbol{\beta}^\star)^2 = O(1)$. The proof of Theorem 3.4.1 (i) is now complete.

Note that

$$
\begin{aligned}
G_j^\star &\equiv E[l(\beta_0^M, Y) - l(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)] \\
&\leq E[l(E \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y) - l(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)].
\end{aligned}
$$

The second step is because of the definition of β_0^M. Apply the general decomposition (S1) to the right hand side of the last line, $G_j^\star$ is upper bounded by

$$
E l'(\boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M, Y)(-\boldsymbol{\Psi}_j^T [\boldsymbol{\beta}_j^M - \mathbf{1} E \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j^M])
$$

$$+E(-\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M])\int_0^1 [l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M + t(-\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M]), Y) - l'(\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M, Y)]dt$$

$$\leq \frac{b_2}{2}E(\mathbf{\Psi}_j^T[\boldsymbol{\beta}_j^M - \mathbf{1}E\mathbf{\Psi}_j^T\boldsymbol{\beta}_j^M])^2.$$

Combine with (S37), $\sum_{j=1}^{p_n} G_j^\star \leq O\big(d_n\lambda_{\max}(\mathbf{\Sigma})\big)$. This implies that the number of $\{j : G_j^\star > \epsilon d_n n^{-2\kappa}\}$ can not exceed $O(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma}))$ for any $\epsilon > 0$.

For Type 3, on the event $B_n = \{\max_{1\leq j\leq p_n}\big|G_{n,j} - G_j^\star\big| \leq \epsilon d_n n^{-2\kappa}\}$, the number of $\{j : G_{n,j} > 2\epsilon d_n n^{-2\kappa}\}$ can not exceed the number of $\{j : G_j^\star > \epsilon d_n n^{-2\kappa}\}$, which is bounded by $O\big(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\big)$. By taking $\epsilon = \nu/2$, we have

$$P\big(|\widehat{M}_{\nu_n}| \leq O(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma}))\big) \geq P(B_n) = 1 - P(B_n^c).$$

The conclusion follows from Theorem 3.3.1 (ii).

For Types 1 and 2, for any $j = 1, \cdots, p_n$, similar to (S19), on the event $\{|\widehat{\beta}_0^M - \beta_0^M| \leq r_3 n^{-\kappa}\}$ and $\{\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \leq r_4 d_n^{1/2} n^{-\kappa}\}$, $G_{n,j} \leq r_6\big\{P_1 + 2P_2 + P_3\big\}$. Similar to (S21), we have $P_1 \leq G_j^\star + r_7 n^{-\kappa}$ except on an event with probability at most $2\exp(-c_{23}n^{1-2\kappa})$. By (S23), $P_2 \leq \big\{(G_j^\star)^{1/2} r_{10}^{1/2} d_n^{1/2} n^{-\kappa} + r_9 n^{-\kappa}\big\}$ except on an event with probability at most $2\exp(-c_{24}n^{1-2\kappa})$. Correspondingly, we have $P_3 \leq r_{10} d_n n^{-2\kappa}$. Overall,

$$G_{n,j} \leq r_6\Big\{\big((G_j^\star)^{1/2} + r_{10}^{1/2} d_n^{1/2} n^{-\kappa}\big)^2 + (r_7 + 2r_9) d_n n^{-2\kappa}\Big\} \tag{S38}$$

except on an event with probability at most $P(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \geq r_4 d_n^{1/2} n^{-\kappa}) + 6\exp(-c_{25}n^{1-2\kappa})$. Based on (S38) and (S24), if $G_j^\star \leq h d_n n^{-2\kappa}$, then $G_{n,j} \leq \nu_n$ except on an event with probability at most $P(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \geq r_4 d_n^{1/2} n^{-\kappa}) + 6\exp(-c_{25}n^{1-2\kappa})$. Therefore, the number of $\{j : G_{n,j} > \nu_n\}$ can not exceed the number of $\{j : G_j^\star > h d_n n^{-2\kappa}\}$ except on an event with probability at most $p_n[P(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \geq r_4 d_n^{1/2} n^{-\kappa}) + 6\exp(-c_{25}n^{1-2\kappa})]$. Since the number of $\{j : G_j^\star > h d_n n^{-2\kappa}\}$ can not exceed $O(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma}))$, we have

$$P\Big(|\widehat{M}_{\nu_n}| \leq O(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma}))\Big) \geq 1 - p_n[P(\|\widehat{\boldsymbol{\beta}}_j^M - \boldsymbol{\beta}_j^M\| \geq r_4 d_n^{1/2} n^{-\kappa}) + 6\exp(-c_{25}n^{1-2\kappa})].$$

In the last expression, we can further apply the results in Theorem 3.2.1. The proof of Theorem 3.4.1 (ii) is now complete.

2. Discussion on Condition (D). We will discuss the connections of our condition (D) with the condition (C) in Fan, Feng & Song (2011) and condition (C2) in He, Wang & Hong (2013).

When l is the squared error loss, minimization in (10) leads to $f_j^M(X_j) = E(Y|X_j)$. When Y is independent of X_j, $f_j^M(X_j) - Ef_j^M(X_j) = 0$. Note that in Fan, Feng & Song (2011) the response variable Y has been subtracted by the mean such that $EY = 0$. Therefore, our condition (D) is exactly the condition (C) in their paper.

When l is the quantile regression loss, minimization in (10) leads to $f_j^M(X_j) = Q_\alpha(Y|X_j)$. When Y is independent of X_j, $f_j^M(X_j) - Ef_j^M(X_j) = 0$. The condition (C2) in He, Wang & Hong (2013) is that $\min_{j\in M_\star} E(Q_\alpha(Y|X_j) - Q_\alpha(Y))^2 \geq c_1 n^{-2\kappa}$. Ignoring the difference in the signal strength at this moment. By simple algebra, we have

$$\begin{aligned} E\big[Q_\alpha(Y|X_j) - Q_\alpha(Y)\big]^2 &= E\big[Q_\alpha(Y|X_j) - EQ_\alpha(Y|X_j)\big]^2 \\ &\quad + \big[EQ_\alpha(Y|X_j) - Q_\alpha(Y)\big]^2. \end{aligned} \tag{S39}$$

Therefore, similar to our condition (D),

$$\min_{j\in M_\star} E\big[Q_\alpha(Y|X_j) - EQ_\alpha(Y|X_j)\big]^2 \geq c_1 n^{-2\kappa}$$

will automatically lead to their condition (C2). However, on the other hand, even for $j \notin M_\star$, it can still be satisfied that $E(Q_\alpha(Y|X_j) - Q_\alpha(Y))^2 \geq c_1 n^{-2\kappa}$ because of the constant term $(EQ_\alpha(Y|X_j) - Q_\alpha(Y))^2$ in (S39). To illustrate this issue, let us consider a simple example. Assume $Y = X_1 + X_2$ and $\mathbf{X} = (X_1, \cdots, X_p)^T \sim N_p(0, \mathbf{\Sigma})$. The diagonal elements of $\mathbf{\Sigma}$ are 1, $\mathrm{cor}(X_1, X_j) = 0$ for $j = 2, \cdots, p$ and $\mathrm{cor}(X_2, X_j) = \rho$ for $j > 2$. It is not difficult to show that $Q_\alpha(Y) = \sqrt{2}\Phi^{-1}(\alpha)$, $Q_\alpha(Y|X_j) = X_j + \Phi^{-1}(\alpha)$ for $j = 1, 2$ and $Q_\alpha(Y|X_j) = \rho X_j + \sqrt{2-\rho^2}\Phi^{-1}(\alpha)$ for $j > 2$. Let $\alpha = 0.95$, $\rho = \sqrt{3}/2$ and use a threshold as 0.8. Based on our criterion,

$$E\big[Q_\alpha(Y|X_j) - EQ_\alpha(Y|X_j)\big]^2 = \begin{cases} 1 & j = 1, 2 \\ 0.75 & j > 2 \end{cases}$$

and only X_1, X_2 will be selected, while based on the criterion in He, Wang & Hong (2013),

$$E\big[Q_\alpha(Y|X_j) - Q_\alpha(Y)\big]^2 \simeq \begin{cases} 1.464 & j = 1, 2 \\ 0.987 & j > 2 \end{cases}$$

and all $\{X_j\}_{j=1}^p$ will be selected.

3. Numerial Results. In this section, we will provide further simulation results for Minimum Model Size under Models 1, 2, 7 & 8, as well as results for iterative screening procedures. A figure about selected genes from the data analysis will also be provided.

Iterative Variable Selection: In the tables S3-S5, we compare iterative version of screening methods and focus on Models 1-6. The GI-Goffins and I-Goffins are constructed based on iterative combination of our Goffins and penGAM in Meier, van de Geer & Bühlmann (2009). ISIS is an iterative version of SIS and the computation is directly based on the R package "SIS". I-EL is an iterative version of EL and penalized empirical likelihood regression described in Chang, Tang & Wu (2013). The methods considered here are based on the availability of iterative screening methods in the existing literature. The simulation round is 200 for GI-Goffins, I-Goffins and ISIS, and is set as 50 for I-EL due to its intensive computation. For Models 1, 2, 5, 6, both GI-Goffins and I-Goffins can select all the four important variables (X_1, X_2, X_3, X_4) even when the dimensionality p increases from 1000 to 5000. GI-Goffins performs better than I-Goffins in terms of selecting fewer false positives. On the other hand, both ISIS and I-EL will miss some true variables while ISIS tends to select too many false positives. For Models 3 and 4, GI-Goffins and I-Goffins can select all the true variables in most of the simulation rounds, but we can miss some important variables occasionally. In contrast, ISIS and I-EL fail to select all true positives but select significantly more false positives than our methods. In terms of the computation time, ISIS is the fastest algorithm because it avoids the nonparametric modeling of the variables. I-EL is very computationally intensive. It requires roughly 4 to 10 times of computation time compared with our methods depending on the various simulation examples. In tables S3-S5, we also present the frequencies of selecting the four important variables (X_1, X_2, X_3, X_4) over the simulation rounds. Ideally, the frequency should be 1 or close to 1, meaning that the variable is selected with high probability. Such detailed investigation further supports the superior performance of our methods compared with other competitors.

REFERENCES.

Boucheron, S., Lugosi, G. and Massart, P. (2003). Concentration inequalities using the entropy method. *Annals of Probability*, **31**, 1583-1614.

Buldygin, V and Kozachenko, Y. (2000). Metric characterization of random variables and random processes. *Translations of Mathematical Monographs*, **188**, American Mathematical Society.

de Boor, C. (1978). A practical guide to splines. *Spinger*, New York.

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Annals of Statistics*, **26**, 755-770.

TABLE S1
The median (IQR) and 5%, 25%, 75%, 95% quantiles of minimum model size.

Model	p	Methods	Median (IQR)	5%	25%	75%	95%
Model 1	1000	Goffins	4(5)	4	4	9	74.3
		Kfilter	116.5(200.75)	9	39	239.75	609.20
		QaSIS	12(40)	4	5	45	302.45
		SIS	488 (468.25)	56.95	242	710.25	931.20
		SIRS	485(513.75)	60.70	241	754.75	939
		DC	42(74)	6	20	94	279.2
		EL	946.5(140)	705.75	856	996	1000
	2000	Goffins	5(9)	4	4	13	106.85
		Kfilter	237(438.25)	7.95	69.75	508	1116.65
		QaSIS	19.5(112.5)	4	5	117.5	643.65
Size=4		SIS	1009(970.5)	101	517	1487.5	1886.5
		SIRS	963(965)	89.75	513.25	1478.25	1908.15
		DC	81.5(178.75)	9	31.75	210.5	528.65
		EL	1894(256.5)	1411.7	1735.5	1992	2000
	5000	Goffins	6(20.25)	4	4	24.25	336.10
		Kfilter	583.5(1126.25)	19	185.75	1312	3049.80
		QaSIS	49.5(231.75)	4	8	239.75	1184.35
		SIS	2493(2560.25)	249	1279	3839.25	4800.15
		SIRS	2549(2555.5)	203.95	1281.5	3837	4757.6
		DC	219.5(403.5)	16	77.75	481.25	1257.25
		EL	4816(635.5)	3591.4	4348.5	4984	5000
Model 2	1000	Goffins	56(158.5)	4	14	172.5	556.05
		Kfilter	265.5(378)	23.95	124	502	838.15
		QaSIS	315(471)	28	127	598	889.2
		SIS	500.5(508.5)	54.95	255	763.5	960.05
		SIRS	518.5(471.5)	67.9	283.5	755	943.3
		DC	173.5(264)	17.9	78	342	704.65
		EL	1000(2)	932.75	998	1000	1000
	2000	Goffins	111(289.75)	6	25.75	315.50	1066.2
		Kfilter	534(787.25)	52	217.75	1005	1596.45
		QaSIS	619.5(807.5)	57.8	295.25	1102.75	1833.75
Size=4		SIS	951(1099)	97.8	443.75	1542.75	1911.05
		SIRS	973.5(915.5)	122.75	493.25	1408.75	1895.25
		DC	331.5(500)	41	142	642	1371
		EL	2000(3)	1872	1997	2000	2000
	5000	Goffins	277.5(687.5)	7.95	65	752.50	2324.75
		Kfilter	1448(2069.75)	106.90	564.75	2634.50	4168.20
		QaSIS	1540.5(2237.25)	110.70	661.50	2898.75	4441
		SIS	2550.5(2624.25)	230.45	1205.75	3830.00	4847.75
		SIRS	2493(2472.25)	264.55	1376.00	3848.25	4736.35
		DC	893(1381.5)	82.90	401.50	1783.00	3258.45
		EL	5000(9)	4503.9	4991.0	5000	5000

TABLE S2
The median (IQR) and 5%, 25%, 75%, 95% quantiles of minimum model size.

Model	p	Methods	Median (IQR)	5%	25%	75%	95%
Model 7	1000	Goffins	13(8)	8	10	18	44.15
		Kfilter	21(26)	9	13	39	123.15
		QaSIS	18(13)	9	13	26	54.05
		SIS	846(298.75)	270	624.75	923.50	987
		SIRS	325.5(373.25)	50.9	180.75	554	862.2
		DC	813.5(332.75)	253.85	588.5	921.25	986.05
		EL	839(215)	388.80	726.25	941.25	990
	2000	Goffins	17(15)	9	12	27	65
		Kfilter	30(54)	9	16	70	281.25
		QaSIS	22(23)	10	15	38	94.1
Size=7		SIS	1682.5(551.75)	592.85	1310.5	1862.25	1975.05
		SIRS	692.5(749.5)	116.65	391.75	1141.25	1661.30
		DC	1628.5(642)	522.85	1208.25	1850.25	1969.10
		EL	1690(410.25)	807.40	1465.00	1875.25	1985.05
	5000	Goffins	26(30)	10	17	47	163
		Kfilter	62(134.25)	10	23	157.25	560.05
		QaSIS	39(53.25)	12	21	74.25	226.15
		SIS	4095.5(1533)	1112.05	3152	4685	4929.25
		SIRS	1712(1891.75)	270.8	869.5	2761.25	4327
		DC	4015(1702)	1098.8	2965.75	4667.75	4949.05
		EL	4281.5(1033.75)	2317.15	3688.25	4722	4959.05
Model 8	1000	Goffins	22.5(25)	10	15	40	95
		Kfilter	44(73)	12	21	94	212.1
		QaSIS	31(37)	13	20	57	130.15
		SIS	831.5(330.25)	285.90	610.00	940.25	986.00
		SIRS	411.5(420.75)	67.00	243.75	664.50	908.05
		DC	790(357)	231.35	566.50	923.50	992.00
		EL	860(217.25)	413.55	722.00	939.25	991.05
	2000	Goffins	33(39)	12	19	58	173.05
		Kfilter	66(117)	12.95	27.00	144.00	452.50
		QaSIS	51.5(64)	15	28	92	241.35
Size=7		SIS	1646.5(681)	471.80	1175.50	1856.50	1976.15
		SIRS	881.5(858)	116.95	454.75	1312.75	1836.30
		DC	1594(686.25)	456.70	1132.00	1818.25	1974.00
		EL	1700(421.75)	860.05	1461.75	1883.50	1979.00
	5000	Goffins	62(99)	17	36	135	622.8
		Kfilter	158(339.25)	14	51	390.25	1219.60
		QaSIS	110.5(178.5)	21	53.75	232.25	912.30
		SIS	4174(1432.5)	1427.65	3201.00	4633.50	4937.05
		SIRS	2151(2249)	290.90	1176.50	3425.50	4444.35
		DC	4028(1644.5)	1206.1	2966.0	4610.5	4935.2
		EL	4229(1138.5)	2112.80	3572.75	4711.25	4947.05

TABLE S3

Average values of the number of true positives (TP), false positives (FP), computation time (in seconds), and frequencies of selecting X_1, X_2, X_3, X_4. Standard deviations are given in parentheses. $p = 1000$.

Model	Method	TP	FP	Time	X_1	X_2	X_3	X_4
Model 1	GI-Goffins	4(0)	0.66(0.82)	71.93(8.75)	1	1	1	1
	I-Goffins	4(0)	2.38(2.36)	45.96(19.84)	1	1	1	1
	ISIS	3.04(0.20)	62.96 (0.20)	9.18(4.72)	1	0.04	1	1
	I-EL	2.69(0.63)	4.47(2.15)	381.57(104.52)	1	0.01	0.81	0.87
Model 2	GI-Goffins	4(0)	0.70(0.79)	63.06(8.70)	1	1	1	1
	I-Goffins	4(0)	7.18(5.71)	57.77(26.96)	1	1	1	1
	ISIS	3.10(0.29)	62.91(0.29)	10.32(5.36)	1	0.095	1	1
	I-EL	2.97(0.20)	5.10(1.91)	261.98(63.09)	1	0.005	0.965	1
Model 3	GI-Goffins	3.99(0.10)	0.80(0.83)	275.08(63.15)	1	0.99	1	1
	I-Goffins	3.96(0.20)	3.14(2.96)	231.96(146.93)	1	0.96	1	1
	ISIS	2.60(0.54)	13.41(0.54)	5.89(1.65)	1	0.545	1	0.05
	I-EL	2.53(0.52)	9.57(4.46)	1518.46(822.39)	1	0.52	1	0.01
Model 4	GI-Goffins	3.93(0.27)	0.75(0.85)	257.07(70.45)	1	0.95	1	0.98
	I-Goffins	3.93(0.26)	3.00(2.93)	219.32(157.03)	1	0.935	1	0.99
	ISIS	3.05(0.34)	12.95(0.35)	6.88(3.85)	1	0.965	1	0.085
	I-EL	2.94(0.24)	10.05(6.32)	3023.88(2139.71)	1	0.94	1	0
Model 5	GI-Goffins	4(0)	0.7(0.76)	114.59(16.31)	1	1	1	1
	I-Goffins	4(0)	2.49(2.53)	69.48(31.39)	1	1	1	1
	ISIS	3.09(0.28)	29.92(0.28)	11.86(4.74)	1	1	1	0.085
	I-EL	3.00(0)	2.74(1.16)	1203.57(316.56)	1	1	1	0
Model 6	GI-Goffins	4(0)	0.67(0.82)	118.44(16.14)	1	1	1	1
	I-Goffins	4(0)	4.32(3.97)	84.12(38.08)	1	1	1	1
	ISIS	3.05(0.22)	29.95(0.22)	10.76(4.12)	1	1	1	0.05
	I-EL	3.00(0)	3.95(1.67)	1296.98(256.86)	1	1	1	0

Ledoux, M. and Talagrand, M. (1991). Probability in Banach Spaces: Isoperimetry and Processes. Springer, Berlin.

Stone, C. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, **13**, 689-705.

Stone, C. (1986). The dimensionality reduction principle for generalized additive models. *Annals of Statistics*, **14**, 590-606.

van de Geer, S. (2002). M-estimation using penalties or sieves. *Journal of Statistical Planning Inference*, **108**, 55-69.

van der Vaart, A. and Wellner, J. (1996). Weak convergence and empirical processes. Springer, New York.

Zhou, S., Shen, X. and Wolfe, D. A. (1998). Local asympotitics for regression splines and confidence regions. *Annals of Statistics*, **26**, 1760-1782.

DEPARTMENT OF STATISTICAL SCIENCE
FOX BUSINESS SCHOOL
TEMPLE UNIVERSITY
PHILADELPHIA, PA 19122
USA
E-MAIL: hanxu3@temple.edu

TABLE S4

Average values of the number of true positives (TP), false positives (FP), computation time (in seconds), and frequencies of selecting X_1, X_2, X_3, X_4. Standard deviations are given in parentheses. $p = 2000$.

Model	Method	TP	FP	Time	X_1	X_2	X_3	X_4
Model 1	GI-Goffins	4(0)	0.70(0.76)	124.50(17.37)	1	1	1	1
	I-Goffins	4(0)	2.89(2.61)	82.19(32.80)	1	1	1	1
	ISIS	3.04(0.17)	62.96(0.17)	13.81(7.80)	1	0.04	1	1
	I-EL	2.58(0.71)	4.14(2.07)	582.51(154.25)	1	0	0.77	0.81
Model 2	GI-Goffins	3.995(0.07)	0.745(0.81)	132.53(17.03)	1	0.995	1	1
	I-Goffins	3.995(0.07)	7.995(6.36)	131.67(60.25)	1	0.995	1	1
	ISIS	3.015(0.12)	62.985(0.12)	16.93(9.71)	1	0.015	1	1
	I-EL	2.92(0.37)	4.78(2.19)	448.01(158.45)	0.99	0.00	0.96	0.97
Model 3	GI-Goffins	3.98(0.16)	0.74(0.79)	480.07(113.32)	1	0.975	1	1
	I-Goffins	3.95(0.23)	3.41(3.25)	436.35(307.93)	1	0.95	1	0.995
	ISIS	2.49(0.54)	13.51(0.54)	10.20(3.58)	1	0.455	1	0.035
	I-EL	2.48(0.50)	12.64(3.15)	3146.64(1791.67)	1	0.46	1	0.02
Model 4	GI-Goffins	3.875(0.42)	0.63(0.77)	465.09(143.47)	1	0.925	1	0.95
	I-Goffins	3.87(0.38)	3.36(3.19)	421.03(294.90)	1	0.9	1	0.97
	ISIS	2.9(0.35)	13.10(0.34)	12.25(7.97)	1	0.88	1	0.02
	I-EL	2.88(0.39)	11.56(5.90)	7135.29(3914.02)	1	0.86	1	0.02
Model 5	GI-Goffins	4(0)	0.71(0.80)	204.37(26.91)	1	1	1	1
	I-Goffins	4(0)	2.82(2.79)	131.85(58.57)	1	1	1	1
	ISIS	3.02(0.12)	29.99(0.12)	18.32(7.69)	1	1	1	0.015
	I-EL	3.0(0)	3.2 (1.55)	2184.23(530.58)	1	1	1	0
Model 6	GI-Goffins	4(0)	0.79(0.80)	199.67(27.18)	1	1	1	1
	I-Goffins	4(0)	5.1(4.61)	155.99(80.37)	1	1	1	1
	ISIS	3.04(0.18)	29.97(0.18)	15.57(5.89)	1	1	1	0.035
	I-EL	3.0(0)	4.3(1.64)	3012.11(475.31)	1	1	1	0

TABLE S5

Average values of the number of true positives (TP), false positives (FP), computation time (in seconds), and frequencies of selecting X_1, X_2, X_3, X_4. Standard deviations are given in parentheses. $p = 5000$.

Model	Method	TP	FP	Time	X_1	X_2	X_3	X_4
Model 1	GI-Goffins	4(0)	0.69(0.77)	305.95(42.11)	1	1	1	1
	I-Goffins	4(0)	2.92(2.85)	212.67(88.62)	1	1	1	1
	ISIS	3.02(0.14)	62.98(0.14)	31.09(17.43)	1	0.02	1	1
	I-EL	2.56(0.67)	4.54(2.14)	1396.24(393.38)	1	0	0.72	0.84
Model 2	GI-Goffins	4(0)	0.74(0.76)	304.65(40.83)	1	1	1	1
	I-Goffins	3.99(0.10)	9.72(6.92)	321.62(137.74)	1	0.99	1	1
	ISIS	3.03(0.17)	62.97(0.17)	37.31(22.30)	1	0.03	1	1
	I-EL	2.86(0.45)	4.16(2.21)	1133.36(409.21)	1	0	0.9	0.96
Model 3	GI-Goffins	3.955(0.21)	0.59(0.74)	1200.29(272.24)	1	0.955	1	1
	I-Goffins	3.86(0.38)	3.42(3.03)	1180.01(830.38)	1	0.885	1	0.975
	ISIS	2.29(0.47)	13.71(0.47)	26.69(11.34)	1	0.26	1	0.03
	I-EL	2.5(0.51)	15.3(2.17)	11733.59(5496.55)	1	0.5	1	0
Model 4	GI-Goffins	3.81(0.52)	0.69(0.81)	1223.72(365.48)	0.99	0.895	1	0.925
	I-Goffins	3.725(0.58)	3.44(2.98)	1100.01(701.61)	0.98	0.825	1	0.92
	ISIS	2.845(0.38)	13.14(0.41)	32.48(22.40)	1	0.835	1	0.01
	I-EL	2.82(0.39)	13.2(4.41)	19738.35(12055.85)	0.98	0.84	1	0
Model 5	GI-Goffins	4(0)	0.61(0.74)	494.42(60.24)	1	1	1	1
	I-Goffins	4(0)	2.63(2.78)	326.54(158.28)	1	1	1	1
	ISIS	3.02(0.12)	29.99(0.12)	40.24(18.42)	1	1	1	0.015
	I-EL	3(0)	3.38(1.63)	5233.61(1584.50)	1	1	1	0
Model 6	GI-Goffins	4(0)	0.75(0.78)	499.04(58.87)	1	1	1	1
	I-Goffins	4(0)	5.46(5.17)	386.97(201.62)	1	1	1	1
	ISIS	3.02(0.14)	29.98(0.14)	34.46(16.03)	1	1	1	0.02
	I-EL	3(0)	4.9(2.32)	6324.55(1728.95)	1	1	1	0

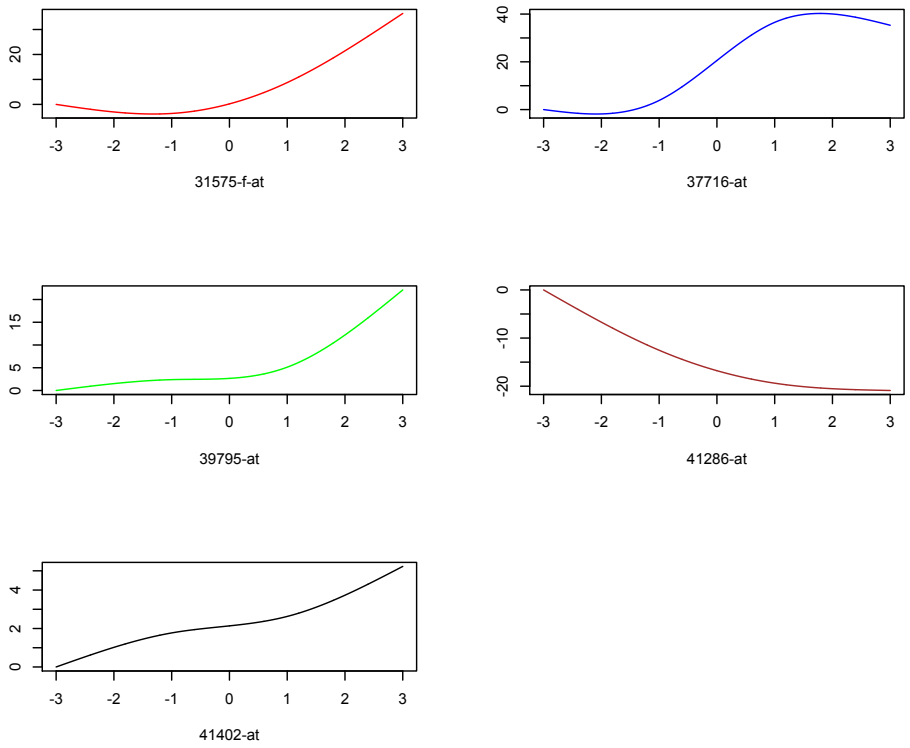

FIG S1. *Fitted nonparametric functions for the five selected genes by GI-Goffins.*