Submitted to Bernoulli

High Dimensional Generalized Linear Models for Temporal Dependent Data

YUEFENG HAN [1], RUEY S. TSAY [2] and WEI BIAO WU [3]

[1]*Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA.*
E-mail: yuefeng.han@rutgers.edu

[2]*Booth School of Business, University of Chicago, Chicago, IL 60637, USA.*
E-mail: ruey.tsay@chicagobooth.edu

[3]*Department of Statistics, University of Chicago, Chicago, IL 60637, USA.*
E-mail: wbwu@galton.uchicago.edu

High dimensional generalized linear models are widely applicable in many scientific fields with data having heavy tails. However, little is known about statistical guarantees on the estimates of such models in a time series setting. In this article, we establish statistical error bounds and support recovery guarantees of the classical ℓ_1 regularized procedure for generalized linear model with temporal dependent data. We also propose a new robust M-estimator for high dimensional time series. Properties of the proposed robust procedure are investigated both theoretically and numerically. As an extension, we introduce a robust estimator for linear regression and show that the proposed robust estimator achieves nearly the optimal rate as that for i.i.d sub-Gaussian data. Simulation results show that the proposed method performs well numerically in the presence of heavy-tailed and serially dependent covariates and/or errors, and it significantly outperforms the classical Lasso method. For applications, we demonstrate, in the supplementary material, the regularized robust procedure via analyzing high-frequency trading data in finance.

Keywords: high dimensional analysis, time series analysis, generalized linear model, robust estimation, support recovery.

1. Introduction

In recent years, information technology has made high dimensional time series data increasingly common. The demand for modelling and forecasting such data arises naturally from communication engineering, environmental studies, market analysis in finance and panel studies in economics, among others. In many applications, we often face the challenge of dealing with a large number of complicated issues such as missing values or heavy tails. The Lasso regularized method, originally introduced by [71] and subsequently investigated by many others, is a popular technique for high dimensional linear regression models with sparse coefficients. As a matter of fact, the ℓ_1-type penalty of the Lasso can also be applied to other models in high dimension, including, for example, logistic regression ([25, 47, 67, 69], among others), multinomial logistic regression [43] or Cox regression [72] by replacing the ℓ_2 loss function by the corresponding negative log-likelihood function.

Generalized linear models (GLM, [51]) are a flexible generalization of the ordinary linear regression by allowing researchers to model the relationship between the predictors and a function of the mean of the response variable, which can follow a continuous or discrete distribution. In a variety of applications, the observed response consists of binary or count data for which GLM is especially useful. See, for example, social network analysis [5, 44, 63, 70, 91], biological neural networks [7, 18, 58], compressed sensing [42, 61, 62], power systems analysis [38], and seismology [57, 80]. This paper deals with the Lasso penalty for GLM applied to high dimensional time series data. Under the independent and identically distributed (i.i.d.) setting, there exists substantial literature on the Lasso methods for high dimensional GLM. For instance, [79] showed non-asymptotic oracle inequalities for the empirical risk minimizer with Lasso penalty for high dimensional GLMs with a Lipschitz loss function. [36] studied Lasso estimator in the Cox proportional hazards regression when the covariates are time dependent, and established oracle inequalities for prediction and estimation errors. A number of papers analyzed penalized methods beyond Lasso. [52] applied group Lasso to high dimensional logistic regression, proposed an efficient algorithm, and showed consistency of the estimator. [56] studied penalized M-estimators with a general class of regularization methods, including an ℓ_2 error bound for the Lasso in GLM. [37] studied weighted absolute penalty and its adaptive, multistage application in GLM. [24] investigated asymptotic equivalence of Lasso and other concave regularized methods in a thresholded parameter space. [41] studied adaptive Lasso and group-Lasso for the functional Poisson regression.

Despite the extensive research in GLMs for i.i.d data, very limited work focused on theoretical properties of the regularized estimates when the observations are dependent. [2] investigated theoretical properties of Lasso estimators with a random design for high dimensional Gaussian processes. [86] analyzed Lasso estimator with a fixed design matrix and Dantzig selector under random design. [31] extended the Lasso estimator to random design and weakly sparse time series with application in now-casting. [90] considered nonparametric sparse additive model for linear autoregression. [27] studied Lasso estimators of high dimensional autoregressive generalized linear models, which was further extended in [49]. See also [26, 30].

The phenomenon of heavy-tails is widely observed in time series data. It is one of the stylized facts in financial econometrics that financial returns and macroeconomic variables have high excess kurtosis. Large scale imaging data in biology, such as neural spike recordings (see, for example, [7, 18, 58]), are often corrupted by non-Gaussian noises. The conventional regression estimator may fare poorly or even be inconsistent when the observations are heavy tailed and/or contaminated by outliers in the predictors and/or the response variable. Therefore, it is important to study effective principles for dealing with heavy-tailed or noisy time series data.

The origin of robust statistics dates back to the fundamental works of John Tukey [76, 77], Peter Huber [39, 40] and Frank Hampel [28, 29]. In general, robustness can be defined in two ways; model misspecification and outliers. For example, Tukey's work [76] is about robustness to a misspecification of the Gaussian model, while Hodges' work

[34] is robustness to contamination of the dataset by extreme outliers or robustness to heavy-tailed distributions in the model that lead to the appearance of some aberrant data. It is well known that if the covariates and/or the errors deviate more wildly from the sub-Gaussian distribution, the linear regression estimator based on the least squares loss no longer converges at the optimal rates. Intuitively, an outlier in the covariates may cause the corresponding M-estimator to behave arbitrarily badly. This motivates the use of generalized M-estimators that downweight high-leverage observations. In the classical theory of robust regression in low dimensions, many weighting functions are introduced, such as Mallows estimator [48], Hill-Ryan estimator [33], and Schweppe estimator [54]. In this paper, we focus on heavy-tailed covariates and heavy-tailed errors for generalized linear model. We also extend the robust M-estimator to high dimensional time series.

Driven by a wide range of contemporary scientific applications, robust regression of high dimensional data is of substantial research interest. Indeed, several papers have shed new light on high dimensional robust M-estimator when the population distribution is heavy tailed or noisy. [12] considered estimation of the mean of heavy-tailed distributions via a robust empirical loss, which is insensitive to extreme values. Cantoni's mean estimator is further extended in [8] to empirical risk minimization. [35, 55] applied "median of mean" estimator to high dimensional sparse regression. [23] introduces a simple principle for robust high dimensional low rank matrix recovery via an appropriate shrinkage on the data. [21] developed estimation bounds for penalized robust regression with the Huber loss function. [45] gave a general framework for robust regularized M-estimators under both convex and non-convex loss functions. However, all prior works focused on the setting where samples are i.i.d. To the best of our knowledge, the existing procedures cannot readily be applied to high dimensional time series data.

The goal of this paper is threefold: (i) To lay a theoretical foundation for high dimensional generalized linear models (GLMs) in situations in which the errors or the covariates can be serially dependent; (ii) To develop novel robust estimation of GLMs for serially dependent data in the case that the dimension can be much larger than the sample size, and to provide a solid theoretical guarantee; (iii) To derive sharp inequalities for tail probabilities for dependent and/or non-sub-Gaussian processes under some mild and easily verifiable conditions. It is worth emphasizing that our model is different from the autoregressive generalized linear model considered in [27] and [49]. It is expected that our framework, inequalities and tools will be useful in other high dimensional problems that involve temporal dependent data.

In this paper, we propose to appropriately shrink the feature variables before calculating the M-estimator to achieve the robustness for high dimensional time series regression. Let X_i be a p-dimensional vector of covariates. If X_i is heavy-tailed, the basic idea is to shrink each feature $X_{ij}(1 \leq j \leq p)$ to a predetermined threshold level τ. We show that the regularized robust regression functions continue to enjoy good behavior. Our first contribution is to provide the asymptotic behavior of the estimated GLM coefficients of the Lasso penalized method for both the original time series data and shrinkage heavy-tailed data. It is shown that an appropriate truncation does not induce significant bias. Under only bounded moment conditions for either noise or covariates, our robust estimator can nearly achieve the error bound for i.i.d. sub-Gaussian data, modulo a price for temporal dependence. The performance of the proposed shrinkage Lasso estimator is thus shown to be much better than that of the vanilla estimator. The allowed dimension p can be as large as $\exp(n^c)$, where n is the sample size and $0 < c < 1$. This means that shrinkage not only overcomes heavy-tailed corruption, but also mitigates the curse of dimensionality. We also establish support recovery guarantees and ℓ_∞ bounds for both methods. Furthermore, unlike the usual robust quasi-likelihood estimators in low dimension, which is non-convex, our method still maintains convexity, thus enjoys certain computational advantages.

In addition, our robust estimator can also be applied to the usual linear regression setting for high dimensional time series. For weakly temporal dependence and heavy-tailed data, our robust method achieves nearly the minimax optimal rate of ℓ_2-norm established by [64] for i.i.d sub-Gaussian data. The difference lies in the scaling requirements on p, n, the sparsity condition, and the additional logarithmic factor of n in the rate, which is induced by the temporal dependence. It is also worth noting that we provide new concentration inequalities, which extend Talagrand's inequality and Bousquet's inequality [6] to high dimensional time series. The extension is of independent interest.

Besides the theoretical properties, we also study the numerical performance of the proposed robust procedure using both simulated and real data. Section 5 considers the simulation studies and shows that our robust procedure performs well numerically in the presence of both symmetric and asymmetric heavy-tailed covariates and/or errors. In particular, the robust procedure significantly outperforms the standard Lasso method, especially in ℓ_1 and ℓ_2 losses of the GLM coefficients. We also illustrate our procedure with an application to high-frequency stock trading for predicting price changes in consecutive transactions via a multinomial logistic regression. Our method leads to marked improvements in prediction compared with the existing methods in financial econometrics.

The rest of the article is organized as follows. Section 2 introduces the standard Lasso procedure and the robust procedure for GLM when the covariates or/and the errors have heavy tails. The framework of high dimensional time series is presented in Section 3.1. Theoretical properties of both robust and non-robust GLM estimators are also investigated in Section 3. After the basic assumptions of Section 3.2, Sections 3.3 and 3.4 study the convergence rates of the standard Lasso procedure and the robust procedure, respectively. Section 4 discusses the conventional linear regression with robust estimator and time series data. Section 5 investigates the numerical performance of the proposed robust procedure and compares it with that of the standard Lasso procedure. Concentration inequalities for high dimensional time series, real data analysis, and all the proofs are given in the supplementary material.

1.1. Notation

Throughout the paper, for a vector $x = (x_1, ..., x_p)'$, define $|x|_2 = (x_1^2 + ... + x_p^2)^{1/2}$, $|x|_\infty = \max\{|x_1|, ..., |x_p|\}$. For a matrix $A = (a_{ij}) \in \mathbb{R}^{p\times m}$, denote entrywise max norm $|A|_\infty = \max_{i,j} |a_{ij}|$ and induced matrix-operator norm $\|A\|_1 = \max_{1\le j\le m} \sum_{i=1}^{p} |a_{ij}|$,

$\|A\|_\infty = \max_{1\le i\le p}\sum_{j=1}^m |a_{ij}|$. If $A \in \mathbb{R}^{p\times p}$ is a square matrix, let $\lambda_1(A) \ge \lambda_2(A) \ge \cdots \ge \lambda_p(A)$ be its eigenvalues in descending order. Also denote $\lambda_{\max}(A) = \lambda_1(A)$ and $\lambda_{\min}(A) = \lambda_p(A)$. For a random variable ξ, let $\|\xi\|_m = (\mathbb{E}|\xi|^m)^{1/m}$ and write $\xi \in \mathcal{L}^m$, $m \ge 1$, if $\|\xi\|_m < \infty$. For simplicity, denote $\|\xi\| = \|\xi\|_2$. For a set S, write $|S|$ as its cardinality. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \le C|b_n|$ holds for all sufficiently large n, write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, and write $a_n \asymp b_n$ if there are positive constants c and C such that $c \le a_n/b_n \le C$ for all sufficiently large n. Denote $a \wedge b = \min\{a,b\}$ and $a \vee b = \max\{a,b\}$. We use $C, C_1, C_2, \cdots$ to denote positive constants whose values may differ from place to place. A constant with a symbolic subscript is used to emphasize the dependence of the value on the subscript. We assume $p = p_n \to \infty$ as $n \to \infty$.

2. The Model

2.1. Generalized linear models and the loss function

Consider n observations $\{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathcal{X} \subset \mathbb{R}^p$ is a p-dimensional vector of covariate variables, and $Y_i \in \mathcal{Y} \subset \mathbb{R}$ is the response variable. We model the dependence of the mean of Y_i on X_i via the linear function $f_{\beta^*}(X_i) = X_i^\top \beta^*$, where β^* is a vector of unknown coefficients and $X_i^\top$ is the transpose of X_i. The goal is to estimate β^*. In a high dimensional model, the number of covariates p can be much larger than the number of observations n. Let $R : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a loss function.

We consider the following estimator of empirical risk minimization with Lasso penalty

$$\hat\beta := \arg\min_\beta \left\{ \frac{1}{n}\sum_{i=1}^n R(f_\beta(X_i), Y_i) + \lambda\,|\beta|_1 \right\}, \tag{1}$$

where $f_\beta(X_i) = X_i^\top \beta$. Assume the response variable Y_i is from an exponential family with the probability density function taking the canonical form

$$h_Y(y;\mu) = \exp\left[y\mu - r(\mu) + b(y)\right]$$

for some known functions $r(\cdot)$ and $b(\cdot)$, and unknown function μ. The function μ is usually called the canonical or natural parameter. The mean response is $r'(\mu)$, the first derivative of $r(\mu)$ with respect to μ. The generalized linear model assumes the form:

$$\mathbb{E}(Y|X) = r'(\mu(X)) = r'(X^\top \beta^*).$$

The canonical link function is thus defined as $g := (r')^{-1}$. Let $z = \mu(X)$. The maximum (marginal) log-likelihood loss function is then

$$R(z, y) = -yz + r(z), \quad y \in \mathcal{Y}, \quad z \in \mathbb{R}. \tag{2}$$

Define the objective empirical risk function as

$$\mathcal{R}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} R(f_\beta(X_i), Y_i).$$

The gradient and Hessian of $\mathcal{R}_n(\beta)$ are respectively

$$\nabla \mathcal{R}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left(r'(f_\beta(X_i)) - Y_i\right) X_i, \tag{3}$$

$$\boldsymbol{H}_n(\beta) = \nabla^2 \mathcal{R}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} r''(f_\beta(X_i)) X_i X_i^\top. \tag{4}$$

We also define the symmetric Bregman divergence as

$$D_{\mathcal{R}}(\beta, \beta^*) = (\beta - \beta^*)^\top (\nabla \mathcal{R}_n(\beta) - \nabla \mathcal{R}_n(\beta^*)), \tag{5}$$

which can be viewed as symmetric, partial Kullback-Leibler distance between the log-likelihood at β and β^*.

In this paper, we focus on $\hat{\beta}$ of the empirical risk minimization of (1) and (2). Extension to quasi-likelihood loss will be discussed in the supplementary material.

2.2. Robust Lasso estimator

Inspired by the theory on robust estimation for linear regression (see e.g. [23]), we study regularized versions of high dimensional robust GLM estimators and establish their statistical guarantees. In order to deal with heavy-tailed data, we propose a robust estimator to be used in (1) by the simple and classical principle of truncation [23], or more generally shrinkage. Our approach is simple: we truncate or shrink appropriately the heavy-tailed covariates or/and the response variable. Intuitively, shrinkage reduces sensitivity of the estimator to data corruption caused by the heavy-tailed distributions. However, shrinkage leads to bias. We shall find an appropriate shrinkage level to balance the induced bias and the statistical error rate. The resulting estimator is then defined as follows:

$$\hat{\beta} := \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} R_\tau(f_\beta(X_i), Y_i) + \lambda \left|\beta\right|_1 \right\}, \tag{6}$$

where τ is a predetermined threshold level,

$$R_\tau(f_\beta(X_i), Y_i) := R(f_\beta(\widetilde{X}_i), Y_i),$$

and $\widetilde{X}_i$ is a truncated version of the covariates X_i if they are heavy-tailed and equals the original covariates (truncation threshold goes to infinity) if they are light-tailed. When the covariates X_i are heavy-tailed, we choose

$$\widetilde{X}_{ij} = \operatorname{sgn}(X_{ij})(|X_{ij}| \wedge \tau), \quad 1 \le j \le p,$$

where $a \wedge b = \min(a, b)$. In practice, we may need to first normalize X_i for each covariate X_{ij}, for $1 \leq j \leq p$, before applying truncation. As discussed in [31], such normalization will not lose signal information when the time series is stationary.

In general, under the GLM setting, the response (e.g., binary data) is considered not to be corrupted by heavy tailed noise. Thus, in Section 3, we focus on the case when the data generating distribution of the response is indeed the model distribution so that we only need to trim the covariates. If the response is also corrupted by random noises, we shall also truncate the response properly. For the linear regression model with least squares loss in Section 4, it is common in the literature to consider the case that the response might have heavy tails. Then we need to further truncate the response to achieve robustness.

With the aforementioned data robustifications, the proposed methodology yields an estimator that, under a bounded moment condition on the covariates or/and the response, has the similar statistical rate as that of the estimator available in the literature for sub-Gaussian distributions. Our study gives a formal theoretical consideration of both the original estimator in (1) and the robust one in (6). As shown in Sections 3 and 4, compared with [23], our rates are nearly optimal up to an additional multiplicative $\log n$ term.

The first and most important advantage of our robust method is to maintain the convexity of (negative) log-likelihood loss. There are several alternatives to robust estimation in the context of low dimensional generalized linear model. For example, [11] proposed a class of robust quasi-likelihood loss function. [3] constructed robust estimator for the logistic regression by bounded deviance, which was further extended to other generalized linear models. [87] introduced a class of robust estimators for generalized linear models motivated by the Bregman divergence. However, most of these robust estimators in the low dimensional case are non-convex.

Our robust method is also much easier to implement than many existing ones, as it only needs to truncate or shrink the data before applying the standard method to the transformed data. The tuning parameter τ plays a key role by adapting to covariates and/or errors with different shapes and tails. In practice, the optimal values of tuning parameters τ and λ can be chosen by a two-dimensional grid search using an information-based criterion or time series cross-validation, e.g., the Akaike information criterion or Bayesian information criterion. Specifically, we may partition a rectangle in the scale of $(\log(\tau), \log(\lambda))$ to form our search grid. Then the optimal values are achieved by the combination of the two parameters that minimizes the cross-validated measurement, the Akaike information criterion or Bayesian information criterion. In addition, the robust cross-validation proposed in [23] can also be used to tune (λ, τ) by replacing the K-fold cross-validation with time series rolling forecast.

3. Asymptotic Properties

We now consider the properties of the standard Lasso method (1) and the robust Lasso method (6). We first show the statistical error rates and the support recovery guarantees of the estimated GLM coefficients for the original time series data and then demonstrate that the convergence rates of robust estimator for shrinkage heavy-tailed data are significantly improved.

3.1. High dimensional time series

Let $\varepsilon_i, i \in \mathbb{Z}$, be i.i.d. random vectors and $\mathcal{F}_i = \sigma(\cdots, \varepsilon_{i-1}, \varepsilon_i)$. We assume that the covariate process $(X_i, i = 1, ..., n)$ is high dimensional and stationary in the form

$$X_i = (g_1(\mathcal{F}_i), ..., g_p(\mathcal{F}_i))^\top, \tag{7}$$

and the response Y_i assumes the form

$$Y_i = g_y(\mathcal{F}_i), \tag{8}$$

where $g_1(\cdot), ..., g_p(\cdot)$ and $g_y(\cdot)$ are measurable functions in $\mathbb{R}$ such that X_i and Y_i are well-defined. In the scalar case with $p = 1$, (7) and (8) include a very general class of stationary processes; c.f. [60, 65, 73, 74, 81, 82].

Following [82], at lag $i \geq 0$, we define the functional dependence measure

$$\begin{aligned}\delta_{i,q,j} &= \|X_{ij} - X_{ij,\{0\}}\|_q = \|g_j(\mathcal{F}_i) - g_j(\mathcal{F}_{i,\{0\}})\|_q,\\ \delta_{i,q,y} &= \|Y_i - Y_{i,\{0\}}\|_q = \|g_y(\mathcal{F}_i) - g_y(\mathcal{F}_{i,\{0\}})\|_q,\end{aligned}$$

where the coupled process $X_{ij,\{0\}} = g_j(\mathcal{F}_{i,\{0\}})$ and $Y_{i,\{0\}} = g_y(\mathcal{F}_{i,\{0\}})$ with $\mathcal{F}_{i,\{0\}} = \sigma(..., \varepsilon_{-1}, \varepsilon'_0, \varepsilon_1, ..., \varepsilon_{i-1}, \varepsilon_i)$ and $\varepsilon'_0, \varepsilon_l, l \in \mathbb{Z}$, being i.i.d. random elements. The dependence measure $\delta_{i,q,j}$ quantifies the q-th moment of the difference between the original process X_{ij} and the coupled process $X_{ij,\{0\}}$ with ε_0 replaced by ε'_0 and all the other innovations kept the same. We assume short-range dependence so that

$$\Delta_{m,q,j} := \sum_{i=m}^{\infty} \delta_{i,q,j} < \infty,$$

$$\Delta_{m,q,y} := \sum_{i=m}^{\infty} \delta_{i,q,y} < \infty.$$

Then for fixed m, $\Delta_{m,q,j}$ and $\Delta_{m,q,y}$ measure the cumulative effect of ε_0 on $(X_{ij})_{i\geq m}$ and $(Y_i)_{i\geq m}$.

We introduce the following dependence adjusted norms

$$\|X_{\cdot j}\|_{q,\alpha} = \sup_{m\geq 0} (m+1)^\alpha \Delta_{m,q,j}, \quad \alpha \geq 0, \tag{9}$$

$$\|X_{\cdot j}\|_{q,\mathrm{GMC}} = \sup_{m\geq 0} \rho_j^{-m} \Delta_{m,q,j}, \quad \text{for some } 0 < \rho_j < 1, \tag{10}$$

$$\|X_{\cdot j}\|_{\psi_\nu} = \sup_{q\geq 2} q^{-\nu} \Delta_{0,q,j}. \tag{11}$$

Similarly, we can define $\|Y_\cdot\|_{q,\alpha}$, $\|Y_\cdot\|_{q,\mathrm{GMC}}$ and $\|Y_\cdot\|_{\psi_\nu}$. Both $\|X_{\cdot j}\|_{q,\alpha}$ and $\|X_{\cdot j}\|_{q,\mathrm{GMC}}$ are called the q-th dependence adjusted norms. By (9), if $\|X_{\cdot j}\|_{q,\alpha} < \infty$, then $\Delta_{m,q,j} =$

$\sum_{i=m}^{\infty} \delta_{i,q,j} = O(m^{-\alpha})$ so that a larger α indicates weaker temporal dependence. In other words, $\|X_{\cdot j}\|_{q,\alpha}$ assumes polynomial decay of the dependence measure with a larger value of α leading to weaker temporal dependence. In contrast, $\|X_{\cdot j}\|_{q,\mathrm{GMC}}$ allows exponential decay of the functional dependence measure $\delta_{i,q,j}$. Property (10) is also called geometric moment contraction (GMC(q)); c.f. [68, 85, 88]. In addition, $\|X_{\cdot j}\|_{\psi_\nu}$ represents the $\psi_{1/\nu}$ Orlicz norm in the dependence case. In the special case that X_{ij} are i.i.d. with mean 0, we have $\delta_{i,q,j} = 0$ for all $i \geq 1$, and $\delta_{0,q,j} = \|X_{0j} - X_{0j,\{0\}}\|_q$. In this case, the q-th dependence adjusted norms $\|X_{\cdot j}\|_{q,\alpha}$ and $\|X_{\cdot j}\|_{q,\mathrm{GMC}}$ are equivalent to the $\mathcal{L}^q$ norm $\|X_{ij}\|_q$ in the sense that $\|X_{ij}\|_q \leq \delta_{0,q,j} \leq \|X_{ij}\|_q + \|X_{ij,\{0\}}\|_q = 2\|X_{ij}\|_q$. Moreover, if $\nu = 1/2$ (resp. $\nu = 1$), $\|X_{\cdot j}\|_{\psi_\nu}$ is a sub-Gaussian (resp. sub-exponential) norm of the random variable X_{ij}. Hence, the dependence adjusted norms $\|\cdot\|_{q,\alpha}$, $\|\cdot\|_{q,\mathrm{GMC}}$ and $\|\cdot\|_{\psi_\nu}$ can be naturally interpreted as the polynomial moment and the exponential moment accounting for dependence, respectively.

***Remark* 3.1.** *If $\|X_{ij}\|_{q^*} < \infty$ for some $q^* > 0$ and $\|X_{\cdot j}\|_{q_0,\mathrm{GMC}} < \infty$ holds for the process (X_{ij}) for some $0 < q_0 \leq q^*$ and $\rho_j \in (0,1)$, then $\|X_{\cdot j}\|_{q,\mathrm{GMC}} < \infty$ also satisfies with the same ρ_j for all $q \in (0, q^*]$. The above property of geometric moment contraction follows by Lemma 2 in [84].*

We provide an example of high dimensional time series below, for which we can calculate the bounds of the dependence adjust norms defined in (9), (10) and (11).

***Example* 3.1.** *Let ε_{ij}, $i, j \in \mathbb{Z}$, be i.i.d. random variables with mean 0 and $\|\varepsilon_{ij}\|_q < \infty$ for some $q \geq 2$. Define the p-dimensional linear process*

$$X_i = \sum_{k=0}^{\infty} A_k \varepsilon_{i-k}, \tag{12}$$

where $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ip})^\top$, A_k are $p \times p$ real coefficient matrices such that $\sum_{k=0}^{\infty} \mathrm{tr}(A_k A_k^\top) < \infty$. Then by Kolmogorov's three series theorem, the linear process (12) is well defined. Let $A_{k,j\cdot}$ be the jth row of A_k. By Rosenthal's inequality [66], we can obtain

$$\delta_{i,q,j} = \|A_{i,j\cdot}\varepsilon_0\|_q \leq (q-1)^{1/2} |A_{i,j\cdot}|_2 \|\varepsilon_{00}\|_q.$$

(i). If there exist $\theta > 1$ and $K > 0$ such that $\max_{1 \leq j \leq p} |A_{i,j\cdot}|_2 \leq K(i+1)^{-\theta}$ for all $i \geq 0$, then with $\alpha = \theta - 1$, we have

$$\|X_{\cdot j}\|_{q,\alpha} = \sup_{m \geq 0} (m+1)^\alpha \sum_{i=m}^{\infty} \delta_{i,q,j} \leq C_{\theta,q} K \|\varepsilon_{00}\|_q,$$

where the constant $C_{\theta,q}$ only depends on θ and q. If $\alpha' > \theta - 1$, then $\|X_{\cdot j}\|_{q,\alpha'}$ may be ∞.
(ii). If there exist $\rho_j \in (0,1)$ and $K > 0$ such that $\max_{1 \leq j \leq p} |A_{i,j\cdot}|_2 \leq K\rho_j^i$ holds for all $i \geq 0$, then since $\sum_{i=m}^{\infty} \rho_j^i = \rho_j^m/(1-\rho_j)$, we have

$$\|X_{\cdot j}\|_{q,\mathrm{GMC}} = \sup_{m \geq 0} \rho_j^{-m} \sum_{i=m}^{\infty} \delta_{i,q,j} \leq \frac{K(q-1)^{1/2}\|\varepsilon_{00}\|_q}{1-\rho_j}.$$

(iii). Suppose $\max_{|\theta|_2=1} q^{-\nu}\|\varepsilon_i^\top\theta\|_q \le C_\nu < \infty$ *for all* $q \ge 2$*. If* $\max_{1\le j\le p}\sum_{i=0}^\infty |A_{i,j\cdot}|_2 \le K$*, for some* $K > 0$*, then*

$$\|X_{\cdot j}\|_{\psi_\nu} = \sup_{q\ge 2} q^{-\nu}\sum_{i=0}^{\infty}\delta_{i,q,j} \le C_\nu K.$$

When $\nu = 1/2$*, it is similar to the sub-Gaussian Orlicz norm in the i.i.d. case; see for example [20, 21].*

To account for the cross-sectional dependence of the p-dimensional stationary process (X_i), we define the $\mathcal{L}^\infty$ functional dependence measure and its corresponding dependence adjusted norm (cf. [14, 15, 89])

$$\omega_{i,q} = \|\max_{1\le j\le p}|X_{ij} - X_{ij,\{0\}}|\|_q = \||X_i - X_{i,\{0\}}|_\infty\|_q,$$

$$\||X_\cdot|_\infty\|_{q,\alpha} = \sup_{m\ge 0}(m+1)^\alpha\Omega_{m,q}, \quad \alpha \ge 0, \quad \text{and} \quad \Omega_{m,q} = \sum_{i=m}^{\infty}\omega_{i,q}.$$

Additionally, we define

$$\||X_\cdot|_\infty\|_{\psi_\nu} = \sup_{q\ge 2} q^{-\nu}\Omega_{0,q}.$$

In this paper, we use the above dependence adjusted norms to study the limiting properties of Lasso estimators in the presence of serial dependence. The framework of functional dependence measure allows many linear and nonlinear time series models; see [68, 82, 85] for examples.

3.2. Assumptions

To prove theoretical properties of our method, we make the following assumptions.

***Assumption* 1** (Convex Loss)**.** *Throughout this paper, the map*

$$z \mapsto R(z, y)$$

is convex for all $y \in \mathcal{Y}$*.*

This assumption is important from a computational perspective; see e.g. [45, 78]. It also plays a crucial role in our theory, as it allows us to prove that the estimator $\hat\beta$ is in a neighborhood of β^*.

***Assumption* 2.** *Consider the function* r *in the maximum log-likelihood loss function (2). It holds that* $|r'(x)| \le M_1 < \infty$, $|r''(x)| \le M_2 < \infty$ *for any* $x \in \mathbb{R}$*. Moreover,* $|r''(x_1) - r''(x_2)| \le M_3|x_1 - x_2|$ *for any* $x_1, x_2 \in \mathbb{R}$ *with* $M_3 < \infty$*.*

Assumption 2 describes some smoothness conditions. Similar assumptions were imposed in [1, 20, 46]. Assumptions 2 requires that the second order derivative is Lipschitz. It is used to control the symmetric Bregman divergence and Hessian of the log-likelihood in a neighborhood of β^*. In fact, except for Poisson regression, many examples of generalized linear models and loss functions satisfy this condition, such as logistic regression, multinomial logistic regression, huber loss, etc.

***Assumption* 3.** *Let $S = \{j : \beta_j^* \neq 0\}$. Assume $|S| \leq s$.*

***Assumption* 4.** *Assume $\mathbb{E}X_i = 0$ and there exists a constant $\kappa_{\mathrm{H}} > 0$ such that $\lambda_{\min}(\mathbb{E}\boldsymbol{H}_n(\beta^*)) = \lambda_{\min}(\mathbb{E}r''(X_i^\top \beta^*)X_i X_i^\top) \geq \kappa_{\mathrm{H}}$.*

Assumption 4 is similar to the compatibility condition in [9, 79]. It is well known that the restricted strong convexity (RSC) of the loss function underpins the statistical guarantee of the M-estimator [56]. In high dimensional sparse linear regression, RSC is implied by the restricted eigenvalue condition [4]. However, in generalized linear models, the Hessian matrix $\boldsymbol{H}_n(\beta)$ depends on β, which creates some technical difficulty of verifying RSC for the loss function. To address the issue, we need to establish local RSC (LRSC) of $\mathcal{R}_n(\beta)$ within a neighborhood of β^*, which has been shown to be sufficient for statistical guarantee of regularized M-estimators in the high dimensional regime [22]. The LRSC is also closely related to the quadratic margin condition in [9, 79]. It can be verified by our Assumptions 2 and 4 in every case of this paper.

Formally, we say a loss function $\mathcal{R}_n(\beta)$ satisfies $\mathrm{LRSC}(\mathcal{C}(S), \mathcal{N}, \kappa_{\mathcal{R}}, \varphi_{\mathcal{R}})$ if any $\Delta \in \mathcal{C}(S)$ and $\beta \in \mathcal{N}$,

$$\mathcal{R}_n(\beta + \Delta) - \mathcal{R}_n(\beta) - (\nabla \mathcal{R}_n(\beta))^\top \Delta \geq \kappa_{\mathcal{R}} |\Delta|_2^2 - \varphi_{\mathcal{R}}, \tag{13}$$

where $\mathcal{N}$ is a neighborhood of β^*, $\kappa_{\mathcal{R}}$ is a positive constant, $\varphi_{\mathcal{R}}$ is a small tolerance term, $\mathcal{C}(S)$ is the restricted cone for $S \subset \{1, 2, ..., p\}$ and $|S| = s$,

$$\mathcal{C}(S) := \{\Delta \in \mathbb{R}^p : |\Delta_{S^c}|_1 \leq 3|\Delta_S|_1\}. \tag{14}$$

Note that, according to [56], when $\lambda \geq 2|\nabla \mathcal{R}_n(\beta^*)|_\infty$, $\hat{\beta} - \beta^*$ falls in the restricted cone $\mathcal{C}(S)$.

3.3. Rate of convergence for the standard Lasso procedure

In this section, we present ℓ_1, ℓ_2 and ℓ_∞ bounds of the standard Lasso estimator $\hat{\beta}$ and the model selection consistency. We first establish the LRSC of Lasso method for the original time series data under the finite polynomial moment case.

Lemma 3.1. *Suppose Assumptions 3 and 4 hold. Assume $|\beta^*|_1 \leq L < \infty$, $|r''(x)| \leq M_2 < \infty$ and $|r''(x_1) - r''(x_2)| \leq M_3|x_1 - x_2|$ with $M_3 < \infty$. Furthermore, assume that*

$\max_{|\nu|_2=1} \mathbb{E}|X_i^\top \nu|^\gamma \le c_0 < \infty$ and $\||X_{\cdot}|_\infty\|_{\gamma,\alpha_X} < \infty$, where $\gamma > 32/7$, $\alpha_X > 21/2 - 8/\gamma$. Let $\alpha_2 = \alpha_X/7 - 1$. Suppose

$$\lambda \asymp a_{p,4}\sqrt{\frac{\log p}{n}} + a_{p,5} n^{16/(7\gamma)-1}(\log p)^{3/2},$$

where

$$a_{p,4} = \||X_{\cdot}|_\infty\|_{\gamma,\alpha_X}^{1/7} \max_j \|X_{\cdot j}\|_{\gamma,\alpha_X}^2 + \max_j \|X_{\cdot j}\|_{\gamma,\alpha_2}^2, \tag{15}$$

$$a_{p,5} = \||X_{\cdot}|_\infty\|_{\gamma,\alpha_X}^{1/7} \||X_{\cdot}|_\infty\|_{\gamma,\alpha_X}^2 + \||X_{\cdot}|_\infty\|_{\gamma,\alpha_2}^2. \tag{16}$$

Let $\mathcal{N} = \{\beta \in \mathbb{R}^p : |\beta - \beta^*|_2^2 \le C_1 s\lambda^2, \beta - \beta^* \in \mathcal{C}(S)\}$, where $\mathcal{C}(S)$ is defined in (14). Then, as long as $n^{2/\gamma}\sqrt{s}\lambda + s\lambda \le C_L$, $\mathcal{R}_n(\beta)$ satisfies LRSC($\mathcal{C}(S), \mathcal{N}, \kappa_{\mathrm{H}}/2, 0$) in an event with probability at least $1 - C_2(\log p)^{-7/(16\gamma)} - C_3 n^{-1} - p^{-C_4}$, where $C_1, ..., C_4 > 0$ are constants and $C_L > 0$ depends on L.

Based on the above lemma, we can derive the statistical rate of the Lasso estimators. For any Lasso solution $\hat{\beta}$ of Equation (1), the following theorem provides the rates of convergence of $|\hat{\beta} - \beta^*|_1$ and $|\hat{\beta} - \beta^*|_2$ by the dependence adjusted norms.

Theorem 3.1. *Suppose Assumptions 1, 2, 3 and 4 hold. Assume $|\beta^*|_1 \le L < \infty$. Also assume $\max_{|\nu|_2=1} \mathbb{E}|X_i^\top \nu|^\gamma \le c_0 < \infty$, $\||X_{\cdot}|_\infty\|_{\gamma,\alpha_X} < \infty$ and $\|Y_{\cdot}\|_{q,\alpha_Y} < \infty$, where $\gamma > 32/7$, $q > 2$, $\alpha_X > 21/2 - 8/\gamma$, $\alpha_Y > 1/2 - 1/\gamma - 1/q > 0$. Let $\alpha_1 = \min\{\alpha_X, \alpha_Y\}$ and $\alpha_2 = \alpha_X/7 - 1$. Define*

$$a_{p,1} = \max_j \|X_{\cdot j}\|_{\gamma,\alpha_1} \|Y_{\cdot}\|_{q,\alpha_1} + \||X_{\cdot}|_\infty\|_{\gamma,\alpha_X}^{1/7} \max_j \|X_{\cdot j}\|_{\gamma,\alpha_X} + \max_j \|X_{\cdot j}\|_{\gamma,\alpha_2},$$

$$a_{p,2} = \||X_{\cdot}|_\infty\|_{\gamma,\alpha_1} \|Y_{\cdot}\|_{q,\alpha_1},$$

$$a_{p,3} = \||X_{\cdot}|_\infty\|_{\gamma,\alpha_X}^{1/7} \||X_{\cdot}|_\infty\|_{\gamma,\alpha_X} + \||X_{\cdot}|_\infty\|_{\gamma,\alpha_2}.$$

Suppose that

$$\lambda \asymp (a_{p,1} + a_{p,4})\sqrt{\frac{\log p}{n}} + \frac{a_{p,2}(\log p)^{3/2}}{n^{1-1/q-1/\gamma}} + \frac{a_{p,3}(\log p)^{3/2}}{n^{1-8/(7\gamma)}} + \frac{a_{p,5}(\log p)^{3/2}}{n^{1-16/(7\gamma)}}, \tag{17}$$

where $a_{p,4}$ and $a_{p,5}$ are defined in (15) and (16). Then, as long as $n^{2/\gamma}\sqrt{s}\lambda + s\lambda \le C_L$, in an event with probability at least $1 - C_1(\log p)^{-q\gamma/(q+\gamma)} - C_2(\log p)^{-7\gamma/16} - n^{-1} - p^{-C_3}$,

$$|\hat{\beta} - \beta^*|_2^2 \le C_4 s\lambda^2, \tag{18}$$

$$|\hat{\beta} - \beta^*|_1 \le C_5 s\lambda, \tag{19}$$

where $C_1, ..., C_5 > 0$ are constants, and $C_L > 0$ only depends on L.

Note that quantities $a_{p,1}, ..., a_{p,5}$ characterize the dependence adjusted norms and may depend on the dimension p. The temporal dependence contributes some additional factors in the statistical error rates and in the sample size requirement. The first term in the order of λ in (17) resembles the well known sub-Gaussian rate $\sqrt{(\log p)/n}$ in the i.i.d. case. The quantities $\alpha_1, \alpha_2, \alpha_X$ measure the temporal dependence strength, and the moment condition γ quantifies the heaviness of the tail. Thus the remaining terms in (17), which are introduced by the heavy-tailedness, will induce a polynomial term of p if $\||X_.|_\infty\|_{\gamma,\alpha_X} \asymp p^{1/\gamma}$.

Here we assume that the short-range dependence condition holds, that is,

$$\||X_.|_\infty\|_{\gamma,\alpha_X} < \infty \quad \text{and} \quad \|Y_.\|_{q,\alpha_Y} < \infty.$$

If it fails, the processes (X_i) and (Y_i) may exhibit some long-range dependence, and the asymptotic behavior can be quite different.

In Theorem 3.1, both α_X and α_Y control the decaying speed of the functional dependence measures. To gain more insight into the effects of temporal dependence, we provide the statistical error rates in the following proposition under the i.i.d. setting. The quantity λ in (17) can be substantially simplified with $(\log p)^{3/2}$ there being replaced by $\log p$, and $a_{p,1}, ..., a_{p,5}$ replaced by the moments that do not involve temporal dependencies.

Proposition 3.1. *Suppose Assumptions 1, 2, 3 and 4 hold. Assume that the observations (X_i, Y_i) are i.i.d. and $\max_{|\nu|_2=1} \mathbb{E}|X_i^\top \nu|^\gamma \le c_0 < \infty$, $\|Y_i\|_q < \infty$, where $\gamma \ge 4$, $q > 2$. Suppose that*

$$\lambda \asymp (\max_j \|X_{ij}\|_\gamma \|Y_i\|_q + \max_j \|X_{ij}\|_\gamma^2)\sqrt{\frac{\log p}{n}} + \frac{(\log p)\|\max_j |X_{ij}|\|_\gamma \|Y_i\|_q}{n^{1-1/q-1/\gamma}} + \frac{(\log p)\|\max_j |X_{ij}|\|_\gamma^2}{n^{1-2/\gamma}}.$$

Then, as long as $n^{2/\gamma}\sqrt{s}\lambda + s\lambda \le C_0$, in an event with probability at least $1 - C_1(\log p)^{-\gamma/2} - C_2(\log p)^{-q\gamma/(q+\gamma)} - n^{-1} - p^{-C_3}$, the statistical error rates in (18) and (19) continue to hold, where $C_0, C_1, C_2, C_3 > 0$ are constants.

***Remark* 3.2** (Effects of heavy-tailedness)**.** *In both Theorem 3.1 and Proposition 3.1, the tuning parameter λ includes a polynomial term of dimension p and sample size n, indicating how the dimension breaks down if the moment condition weakens or the dependence becomes stronger. In addition, error bounds (18) and (19) with the new rates for λ show that the convergence rate of the estimated coefficient $\hat{\beta}$ can be much slower than that in the i.i.d. sub-Gaussian setting.*

When the process X_i has an exponential type tail bound, we verify the LRSC in the following Lemma 3.2. Differently from Theorem 3.1, sharper statistical rates of $\hat{\beta}$ are established in Theorem 3.2.

Lemma 3.2. *Suppose Assumptions 3 and 4 hold. Assume $|\beta^*|_1 \le L < \infty$, $|r''(x)| \le M_2 < \infty$ and $|r''(x_1) - r''(x_2)| \le M_3|x_1 - x_2|$ with $M_3 < \infty$. Furthermore, assume*

$\max_{1\le j\le p} \|X_{.j}\|_{\psi_\iota} < \infty$, *and* $\sup_{\gamma\ge 1}\max_{|\theta|_2=1}\gamma^{-\iota}\|X_i^\top\theta\|_\gamma \le c_0 < \infty$, *where* $\iota > 0$. *Let* $\mathcal{N} = \{\beta\in\mathbb{R}^p : |\beta-\beta^*|_2^2 \le C_1 s\lambda^2, \beta-\beta^*\in\mathcal{C}(S)\}$, *where* $\mathcal{C}(S)$ *is defined in* (14).
(i). Suppose

$$\lambda \asymp \max_j \|X_{.j}\|_{\psi_\iota}^3 \frac{(\log p)^{1/2+3\iota}}{\sqrt{n}}.$$

Then, as long as $(\log n)^\iota\sqrt{s}\lambda + s\lambda \le C_L$, $\mathcal{R}_n(\beta)$ *satisfies LRSC*$(\mathcal{C}(S),\mathcal{N},\kappa_{\mathrm{H}}/2,0)$ *in an event with probability at least* $1-n^{-C_2}-p^{-C_3}$, *where* $C_1, C_2, C_3 > 0$ *are constants and* $C_L > 0$ *depends on* L.

(ii). Assume the process X_i *also satisfies the geometric moment contraction, so that* $\|X_{.j}\|_{6,\mathrm{GMC}} < \infty$ *for some constant* $0<\rho_j<1$, *and* $\rho = \min_j\{\rho_j\}\in(0,1)$. *Suppose*

$$\lambda \asymp \max_j \|X_{.j}\|_{6,\mathrm{GMC}}^3\sqrt{\frac{\log p}{n}} + \frac{(\log p+\log n)^{1+2\iota}}{n}.$$

Then, so long as $(\log n)^\iota\sqrt{s}\lambda + s\lambda \le C_L$, $\mathcal{R}_n(\beta)$ *satisfies LRSC*$(\mathcal{C}(S),\mathcal{N},\kappa_{\mathrm{H}}/2,0)$ *in an event with probability at least* $1-n^{-C_5}-p^{-C_6}$, *where* $C_1, C_5, C_6 > 0$ *are constants and* $C_L > 0$ *depends on* L.

Theorem 3.2. *Suppose Assumptions 1, 2, 3 and 4 hold. Assume* $|\beta^*|_1 \le L < \infty$. *Also assume* $\sup_{\gamma\ge 1}\max_{|\theta|_2=1}\gamma^{-\iota}\|X_i^\top\theta\|_\gamma \le c_0 < \infty$, $\max_{1\le j\le p}\|X_{.j}\|_{\psi_\iota} < \infty, \|Y_.\|_{\psi_\nu} < \infty$, *where* $\iota,\nu > 0$.

(i). Suppose

$$\lambda \asymp \max_j \|X_{.j}\|_{\psi_\iota}\|Y_.\|_{\psi_\nu}\frac{(\log p)^{1/2+\nu+\iota}}{\sqrt{n}} + \max_j \|X_{.j}\|_{\psi_\iota}^3\frac{(\log p)^{1/2+3\iota}}{\sqrt{n}}. \tag{20}$$

Then, so long as $(\log n)^\iota\sqrt{s}\lambda + s\lambda \le C_L$, *in an event with probability at least* $1-n^{-C_1}-p^{-C_2}$, *we have*

$$|\hat\beta-\beta^*|_2^2 \le C_3 s\lambda^2, \tag{21}$$

$$|\hat\beta-\beta^*|_1 \le C_4 s\lambda, \tag{22}$$

where $C_1, C_2, C_3, C_4 > 0$ *are constants, and* $C_L > 0$ *only depends on* L.

(ii). Assume the processes X_i *and* Y_i *also satisfy geometric moment contraction, so that* $\|X_{.j}\|_{6,\mathrm{GMC}} < \infty$ *for some constant* $0<\rho_j<1$, $\|Y_.\|_{4,\mathrm{GMC}} < \infty$ *for some constant* $0<\rho_y<1$, *and* $\rho = \min\{\rho_j,\rho_y\}\in(0,1)$. *Suppose*

$$\begin{aligned}\lambda \asymp{}& (\max_j \|X_{.j}\|_{4,\mathrm{GMC}}\|Y_.\|_{4,\mathrm{GMC}} + \max_j \|X_{.j}\|_{6,\mathrm{GMC}}^3)\sqrt{\frac{\log p}{n}} \\ &+ \frac{(\log p+\log n)^{1+\iota+\nu}}{n} + \frac{(\log p+\log n)^{1+2\iota}}{n}.\end{aligned} \tag{23}$$

Then, as long as $(\log n)^\iota\sqrt{s}\lambda + s\lambda \le C_L$, in an event with probability at least $1 - n^{-C_1} - p^{-C_2}$,

$$|\hat\beta - \beta^*|_2^2 \le C_3 s\lambda^2, \tag{24}$$

$$|\hat\beta - \beta^*|_1 \le C_4 s\lambda, \tag{25}$$

where $C_1, C_2, C_3, C_4 > 0$ are constants, and $C_L > 0$ only depends on L.

Theorem 3.2 describes how the rate of convergence depends on the sample size n, the dimension p, and the dependence adjusted norms which are characterized by ι and ν. It suggests that, under the short-range dependence with exponential moment conditions, we can take $\lambda \asymp (\log p)^{c_1}/n^{c_2}$ for some positive constants c_1, c_2 and $c_1 < c_2$. Based on the scaling condition $(\log n)^\iota\sqrt{s}\lambda + s\lambda \le C_L$, p is allowed to be of ultra high dimension in that $\exp(n^c)$ for some $0 < c < 1$.

***Remark* 3.3.** *In the setting of Theorem 3.2(ii), besides exponential moment, we also assume X_i and Y_i are weakly dependent and satisfy the geometric moment contraction, which is defined in (10). In other words, if X_i and Y_i have exponentially decay speed of the functional dependence measures, then, a sharper convergence rate of $\hat\beta$ can be achieved. In the order of the tuning parameter λ in (23), $\max_j \|X_{\cdot j}\|_{4,\mathrm{GMC}}\|Y_\cdot\|_{4,\mathrm{GMC}} + \max_j \|X_{\cdot j}\|_{6,\mathrm{GMC}}^3$ contributes some additional dependence adjusted norm terms. Meanwhile, the term 1 in the power of the terms $(\log p + \log n)^{1+\iota+\nu} + (\log p + \log n)^{1+2\iota}$ is introduced by the geometric moment contraction. The following Proposition 3.2 shows the results in the case that the observations (X_i, Y_i) are i.i.d. and have exponential tail bounds.*

Proposition 3.2. *Suppose Assumptions 1, 2, 3 and 4 hold. Assume that the observations (X_i, Y_i) are i.i.d. and $\sup_{\gamma\ge 1}\max_{|\theta|_2=1}\gamma^{-\iota}\|X_i^\top\theta\|_\gamma \le c_0 < \infty$, $\sup_{q\ge 1} q^{-\nu}\|Y_i\|_q \le c_0 < \infty$, where $\iota, \nu > 0$. Suppose that*

$$\lambda \asymp \sqrt{\frac{\log p}{n}} + \frac{(\log p + \log n)^{\iota+\nu}}{n} + \frac{(\log p + \log n)^{2\iota}}{n}.$$

Then, as long as $(\log n)^\iota\sqrt{s}\lambda + s\lambda \le C_0$, in an event with probability at least $1 - n^{-C_1} - p^{-C_2}$, (24) and (25) continue to hold, where $C_0, C_1, C_2 > 0$ are constants.

Again, both Theorem 3.2 and Proposition 3.2 show that the range of dimension p is narrower than the range $\log(p) = o(n)$ in the i.i.d. sub-Gaussian setting, and the convergence rates of the estimated coefficient $\hat\beta$ are also slower.

Theorems 3.1 and 3.2 also lead to the following result on the ℓ_∞ bound of $\hat\beta$ and support recovery.

Corollary 3.1. *Let $H(\beta^*) = \mathbb{E}H_n(\beta^*)$. Suppose the following incoherence condition holds*

$$\left\| H(\beta^*)_{S^cS} \left(H(\beta^*)_{SS}\right)^{-1} \right\|_\infty \leq \eta < 1,$$

where $\|\cdot\|_\infty$ denotes the ℓ_∞ matrix operator norm. Further, suppose the conditions of Theorem 3.1 (resp. Theorem 3.2(i) or 3.2(ii)) are satisfied. Then, as long as $n^{2/\gamma}s\lambda + s^{3/2}\lambda \leq C_L$ (resp. $(\log n)^\iota s\lambda + s^{3/2}\lambda \leq C_L$), in an event with probability at least $1 - C_1(\log p)^{-\chi} - C_2(\log p)^{-7\gamma/16} - n^{-1} - p^{-C_3}$ (resp. $1 - n^{-C_1} - p^{-C_2}$), we have $\hat{\beta}_{S^c} = 0$ and

$$|\hat{\beta} - \beta^*|_\infty \leq C_0 \| \left(H(\beta^*)_{SS}\right)^{-1} \|_\infty \lambda,$$

where $C_0, ..., C_3 > 0$, $C_L > 0$ depends on L, λ is defined in (17) (resp. (20) or (23)).

Similarly to Corollary 3 in [46], the required scaling condition for support recovery would be stronger than that for Theorem 3.1 or Theorem 3.2. It is also worth noting that the incoherence condition in Corollary 3.1 can be removed by using various non-convex regularizers, as shown in [46].

***Remark* 3.4.** *For i.i.d. data, the asymptotic behavior of Lasso estimator for GLM was studied in [79]. Many other papers investigated high dimensional robust M-estimator, such as [21] and [45]. A key technique used in these articles is Massarts inequality [50], Bousquet's inequality [6] or other similar inequalities for empirical process. These inequalities cannot be directly used for serially dependent data. It is noteworthy that the proofs of Theorem 3.1 and 3.2 require new concentration inequalities for high dimensional time series. We establish Bousquet-type inequality for time dependent data, which is shown in the supplementary material.*

***Remark* 3.5.** *The setting in our Theorems 3.1 and 3.2 is very general as it allows dependent and/or non sub-Gaussian processes and it also allows heteroscedasticity in that the error process and the covariate process can be dependent. Comparing with the conditions in the i.i.d setting (Propositions 3.1 and 3.2), our Theorems 3.1 and 3.2 require an additional condition $|\beta^*| \leq L < \infty$ to derive functional dependence measures for temporal dependent processes.*

3.4. Rate of convergence for the robust procedure

To tackle the problem of heavy-tailed data, we propose to use the robust regularization method in Section 2.2, and analyze the robust estimator in this section. Under similar assumptions, we establish LRSC of the proposed robust procedure.

Lemma 3.3. *Assume $|\beta^*|_1 \leq L < \infty$, $|r''(x)| \leq M_2 < \infty$ and $|r''(x_1) - r''(x_2)| \leq M_3|x_1 - x_2|$ with $M_3 < \infty$. Also assume $\mathbb{E}X_{ij}^6 \leq C < \infty$, for any $1 \leq j \leq p$,*

$\max_{|\nu|_2=1}\mathbb{E}|X_i^\top\nu|^2 \le c_0 < \infty$, $\|X_{.j}\|_{6,\mathrm{GMC}} < \infty$ for some constant $0 < \rho_j < 1$, and $\rho = \min_j \rho_j \in (0,1)$. Suppose $\lambda \asymp (\log n)\sqrt{(\log p)/n}$ and $\tau \asymp n^{1/4}(\log p)^{-1/4}(\log n)^{-1/2}$. Let $\mathcal{N} = \{\beta \in \mathbb{R}^p : |\beta-\beta^*|_2^2 \le C_1 s\lambda^2, \beta-\beta^* \in \mathcal{C}(S)\}$, where $\mathcal{C}(S)$ is defined in (14). Then, as long as $s^2(\log n)(\log p/n)^{1/2} \le C_2$, $\mathcal{R}_n(\beta)$ satisfies LRSC($\mathcal{C}(S), \mathcal{N}, \kappa_{\mathrm{H}}/2, 0$) in an event with probability at least $1-p^{-C_3}$, where C_1 and C_2 are positive constants and $C_3 > 0$ depends on L, ρ, and $\max_j \|X_{.j}\|_{6,\mathrm{GMC}}$.

The next theorem shows the rate of convergence of $|\hat\beta-\beta^*|_1$ and $|\hat\beta-\beta^*|_2$ by the dependence adjusted norms. In Theorem 3.3, the response is generated from a particular distribution in the exponential family. Therefore, there does not exist any model misspecification issue, and the response has sub-exponential tails.

Theorem 3.3. *Suppose Assumptions 1, 2, 3 and 4 hold. Assume $|\beta^*|_1 \le L < \infty$. Also assume $\mathbb{E}X_{ij}^6 \le C < \infty$, for any $1 \le j \le p$, $\max_{|\nu|_2=1}\mathbb{E}|X_i^\top\nu|^2 \le c_0 < \infty$, and $\mathbb{E}Y_i^4 \le C < \infty$. Let $\|X_{.j}\|_{6,\mathrm{GMC}} < \infty$ for some constant $0 < \rho_j < 1$, $\|Y_.\|_{4,\mathrm{GMC}} < \infty$ for some constant $0 < \rho_y < 1$, $\|Y_.\|_{\psi_\nu} < \infty$ for $\nu > 0$, and $\rho = \min\{\rho_j, \rho_y\} \in (0,1)$. Choose $\tau \asymp n^{1/4}(\log p)^{-1/4}(\log n)^{-1/2}$, and $\lambda = C_1(\log n)\sqrt{(\log p)/n}$. Then, as long as $(\log n)^{2\nu+1}(\log p/n)^{1/2} + s^2(\log n)(\log p/n)^{1/2} \le C_2$, in an event with probability at least $1-n^{-C_3}-p^{-C_4}$,*

$$|\hat\beta-\beta^*|_2^2 \le C_5 s\left(\frac{(\log n)^2\log p}{n}\right), \tag{26}$$

$$|\hat\beta-\beta^*|_1 \le C_6 s\left(\frac{(\log n)^2\log p}{n}\right)^{1/2}, \tag{27}$$

where $C_1, ..., C_6 > 0$ are constants depending on L, ρ, $\max_{1\le j\le p}\|X_{.j}\|_{6,\mathrm{GMC}}$ and $\|Y_.\|_{4,\mathrm{GMC}}$.

Corollary 3.2. *If $\max_i |Y_i| < C < \infty$, such as categorical variables, then, as long as $s^2(\log n)(\log p/n)^{1/2} \le C_2$, in an event with probability at least $1-p^{-C_4}$, the upper bounds (26) and (27) hold.*

Corollary 3.3. *Let $H(\beta^*) = \mathbb{E}H_n(\beta^*)$. Suppose the following incoherence condition holds*

$$\left\|H(\beta^*)_{S^cS}\left(H(\beta^*)_{SS}\right)^{-1}\right\|_\infty \le \eta < 1.$$

Also, suppose the conditions of Theorem 3.3 are satisfied. Choose $\lambda = C_1(\log n)\sqrt{(\log p)/n}$ and $\tau \asymp n^{1/4}(\log p)^{-1/4}(\log n)^{-1/2}$. Then, as long as $(\log n)^{2\nu+1}(\log p/n)^{1/2} + s^3(\log n)(\log p/n)^{1/2} \le C_2$, in an event with probability at least $1-n^{-C_3}-p^{-C_4}$, we have $\hat\beta_{S^c} = 0$ and

$$|\hat\beta-\beta^*|_\infty \le C_0\|\left(H(\beta^*)_{SS}\right)^{-1}\|_\infty(\log n)\sqrt{(\log p)/n},$$

where $C_0, ..., C_4 > 0$ and depend on L, ρ, $\max_{1\le j\le p}\|X_{.j}\|_{6,\mathrm{GMC}}$ and $\|Y_.\|_{4,\mathrm{GMC}}$.

Theorem 3.3 indicates that the robust estimator admits nearly the same rate as that of the i.i.d. sub-Gaussian setting. It is much sharper than the deviation bounds in Theorem 3.1 under finite dependence adjusted norms, which will include a polynomial term of p if $\||X_\cdot|_\infty\|_{\gamma,\alpha_X} \asymp p^{1/\gamma}$. In comparison with the robust procedure in the i.i.d. case, we have an additional $\log n$ factor when the result is generalized to the time series setting. Under the scaling condition $(\log n)^{2\nu+1}(\log p/n)^{1/2} + s^2(\log n)(\log p/n)^{1/2} \leq C_2$, our robust regularization method in (6) can handle the ultra high dimension case with $\log p = o(s^{-4}(\log n)^{-2}n + (\log n)^{-4\nu-2}n)$, which is much wider than the range of Theorem 3.1, which only allows a polynomial increase with n. In the robust procedure, the order of τ is not related to the finite moment condition. Only the constant term of τ should be adapted to the degree of heavy-tailedness. Comparing Corollary 3.3 with Corollary 3.1, the support recovery guarantee and the ℓ_∞ bound are also improved by the robust procedure.

***Remark* 3.6.** *The requirement of geometric moment contraction (exponentially decaying dependence measure) is needed. In the high dimensional setting, concentration inequalities are the cornerstone for theoretical analysis. [13] showed that exponential decay bounds for large deviation type tail probabilities do not hold generally with polynomial decaying functional dependence measures as defined in* (9)*,* i.e. $\sum_{i=m}^{\infty} \delta_{i,q,j} = O(m^{-\alpha})$*, even if the process is uniformly bounded. In other words, if the temporal dependence is not very weak, any robust regularized M-estimators for observations with finite moments are not able to achieve statistical error rates close to the optimal rates under the i.i.d. sub-Gaussian case, and can only handle the dimensionality $p = o(n^c)$ for some $c > 0$. This is significantly different from the low dimensional M-estimates, where polynomial decaying dependence measures and even long-range dependent process can be applied; see for example [83].*

4. Linear Regression

Robust estimation of linear regression can be regarded as a generalized linear model with quadratic loss. In this special case, although the first derivative of the loss function is not bounded, we still have an explicit concentration result for our robust estimator. Consider the usual linear regression setup for the response variable Y_i and the covariate vector X_i,

$$Y_i = X_i^\top \beta^* + e_i, \tag{28}$$

where $\beta^* \in \mathbb{R}^p$ is the unknown parameter vector to be estimated and e_i is the error term. Let the loss function $R(f_\beta(X_i), Y_i) = (Y_i - f_\beta(X_i))^2/2$. Differently from Theorem 3.3 in Section 3, if the error has heavy tails, the response Y_i also has heavy tails. This motivates us to truncate both the heavy tailed covariates X_i and the response Y_i under the ℓ_2 loss. Then, similarly to (6), we propose to use the following M-estimator of β^*

with the generalized ℓ_2 loss to robustify the estimation:

$$\hat{\beta} := \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (\widetilde{Y}_i - f_\beta(\widetilde{X}_i))^2 + \lambda \left|\beta\right|_1 \right\}, \tag{29}$$

where

$$f_\beta(\widetilde{X}_i) = \widetilde{X}_i^\top \beta. \tag{30}$$

Here, we choose $\widetilde{Y}_i = \widetilde{Y}_i(\tau_1) = \operatorname{sgn}(Y_i)(|Y_i| \wedge \tau_1)$ and $\widetilde{X}_{ij} = \operatorname{sgn}(X_{ij})(|X_{ij}| \wedge \tau_2)$ for all $1 \le j \le p$, where both τ_1 and τ_2 are predetermined thresholds.

To be solvable in the high dimensional regression setting, β^* is usually assumed to be sparse or weakly sparse, *i.e.*, many elements of β^* are 0 or small. In particular, we impose the following weak sparsity condition and a condition on the covariance matrix of the covariates.

***Assumption* 5.** *(Weak Sparsity Condition). There exits some $0 \le \vartheta < 1$, with a uniform radius K_ϑ, such that*

$$\sum_{j=1}^{p} |\beta_j^*|^\vartheta \le K_\vartheta. \tag{31}$$

Note that K_ϑ might depend on n and p. In the special case $\vartheta = 0$, this quantity corresponds to an exact sparsity constraint–that is, β^* has at most K_0 nonzero entries.

***Assumption* 6.** *Suppose $\mathbb{E}X_i = 0$ and there exists a constant $\kappa_0 > 0$ such that $\lambda_{\min}(\mathbb{E}X_i X_i^\top) \ge \kappa_0$.*

To derive the statistical error rate of $\hat{\beta}$, we establish, in the following lemma, the restricted strong convexity of the robust procedure.

Lemma 4.1. *Assume $\mathbb{E}X_{ij}^4 \le C < \infty$, for any $1 \le j \le p$. Let $\|X_{\cdot j}\|_{4,\mathrm{GMC}} < \infty$ for some constant $0 < \rho_j < 1$. Choose $\tau_2 \asymp n^{1/4}(\log p)^{-1/4}(\log n)^{-1/2}$. Then, for some constants $C_1, C_2 > 0$, which only depend on $\min_j \rho_j$ and $\max_j \|X_{\cdot j}\|_{4,\mathrm{GMC}}$, we have*

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} \beta^\top \widetilde{X}_i \widetilde{X}_i^\top \beta \ge \beta^\top (\mathbb{E}X_i X_i^\top)\beta - C_1(\log n)\sqrt{\frac{\log p}{n}} |\beta|_1^2, \quad \forall \beta \in \mathbb{R}^p \right) \le p^{-C_2}. \tag{32}$$

Finally, in the following Theorem 4.1, we present the statistical error rate of $\hat{\beta}$ as defined in (29). We show that $|\hat{\beta} - \beta^*|_2$ and $|\hat{\beta} - \beta^*|_1$ are upper bounded by nearly optimal rates under light tails and i.i.d. data so long as the tuning parameters τ_1, τ_2 and λ are properly chosen. Corollary 4.1 concerns support recovery and ℓ_∞ bounds.

Theorem 4.1. *Suppose Assumptions 5 and 6 hold. Assume $|\beta^*|_1 \leq L < \infty$. Also assume $\max_j \mathbb{E}X_{ij}^4 \leq C < \infty$ and $\mathbb{E}Y_i^4 \leq C < \infty$. Let $\|X_{\cdot j}\|_{4,\mathrm{GMC}} < \infty$ for some constant $0 < \rho_j < 1$, $\|Y_{\cdot}\|_{4,\mathrm{GMC}} < \infty$ for some constant $0 < \rho_y < 1$, and $\rho = \min\{\rho_j, \rho_y\} \in (0,1)$. Choose $\tau_1 \asymp \tau_2 \asymp n^{1/4}(\log p)^{-1/4}(\log n)^{-1/2}$ and $\lambda = C_1(\log n)\sqrt{(\log p)/n}$. Then, as long as $K_\vartheta[(\log n)^2(\log p)/n]^{(1-\vartheta)/2} \leq C_2$, we have, in an event with probability at least $1 - p^{-C_3}$,*

$$|\hat{\beta} - \beta^*|_2^2 \leq C_4 K_\vartheta \left(\frac{(\log n)^2 \log p}{n}\right)^{1-\vartheta/2}, \tag{33}$$

$$|\hat{\beta} - \beta^*|_1 \leq C_5 K_\vartheta \left(\frac{(\log n)^2 \log p}{n}\right)^{(1-\vartheta)/2}, \tag{34}$$

where $C_1, ..., C_5 > 0$ depend on L, ρ, $\max_{1\leq j\leq p}\|X_{\cdot j}\|_{4,\mathrm{GMC}}$ and $\|Y_{\cdot}\|_{4,\mathrm{GMC}}$.

Corollary 4.1. *Consider the exact sparsity case with $K_0 = s$. Let $\Sigma = \mathbb{E}(X_i X_i^\top)$. Suppose the following incoherence condition holds*

$$\left\|\Sigma_{S^cS}\left(\Sigma_{SS}\right)^{-1}\right\|_\infty \leq \eta < 1.$$

Suppose also the conditions of Theorem 4.1 are satisfied. Choose $\lambda = C_1(\log n)\sqrt{(\log p)/n}$ and $\tau_1 \asymp \tau_2 \asymp n^{1/4}(\log p)^{-1/4}(\log n)^{-1/2}$. Then, as long as $s[(\log n)^2(\log p)/n]^{1/2} \leq C_2$, in an event with probability at least $1 - p^{-C_3}$, we have $\hat{\beta}_{S^c} = 0$ and

$$|\hat{\beta} - \beta^*|_\infty \leq C_0 \|\left(\Sigma_{SS}\right)^{-1}\|_\infty (\log n)\sqrt{(\log p)/n},$$

where $C_0, ..., C_3 > 0$ depend on L, ρ, $\max_{1\leq j\leq p}\|X_{\cdot j}\|_{4,\mathrm{GMC}}$ and $\|Y_{\cdot}\|_{4,\mathrm{GMC}}$.

Theorem 4.1 reveals that $|\hat{\beta}-\beta^*|_2$ has a convergence rate $K_\vartheta^{1/2}[(\log n)^2(\log p)/n]^{1/2-\vartheta/4}$. In comparison with the optimal rate for weak sparsity models under light tails and i.i.d data setting [64], our result concerns high dimensional time series and relaxes the Gaussian/sub-Gaussian assumption to the existence of finite moments at the minor cost of a logarithmic factor of n in the convergence rate. In a special case that $\vartheta = 0$, $|\hat{\beta} - \beta^*|_2$ converges at the rate $K_0^{1/2}(\log n)((\log p)/n)^{1/2}$, where K_0 is the number of non-zero elements of β^*. It suggests that our robust method does not lose much information for heavy-tailed data with exponentially decaying functional dependence measures. In addition, to achieve the desired statistical rate, we need the scaling condition $K_\vartheta(\log n)(\log p/n)^{(1-\vartheta)/2} \leq c_2$, which also includes an additional $\log n$ factor.

***Remark* 4.1.** *Theorem 2 in [23] studied robust linear regression in the i.i.d. case and suggested $\lambda \asymp \sqrt{\log p/n}$ to achieve the minimax optimal rate for $\hat{\beta}$. In comparison, there is an additional multiplicative $\log n$ term in the order of λ and also the convergence rates for the dependent case, which is induced by the additional order $(\log n)^2$ in the Bernstein-type inequality under geometric moment contraction (see the supplementary materials), which serves as the main technical tool. To the best of our knowledge,* *the sharpest available Bernstein-type inequalities for weakly dependent random variables without any structural assumptions (such as linear process) involve additional* $\log n$ *terms [53, 88].*

5. Simulation Study

In this section, we expound upon some concrete instances of our theoretical results and provide some simulation results. We assess the finite sample performance of the robust procedure and compare it with the standard Lasso procedure in logistic regression and linear regression. The implementation of our robust procedure is simple: truncate or shrink the data appropriately, then apply the standard procedure to the transformed data. The simulation results are based on 5000 independent Monte Carlo replications. We select the optimal values of tuning parameters λ and τ by a two dimensional grid search using Bayesian information criterion. It is worth noting that the robust cross-validation proposed in [23] can also be used by replacing the K-fold cross-validation with time series rolling forecast.

We first specify the parameters of the logistic regression. We generate data from independent AR(1) processes, say

$$Z_{ij} = \phi_j Z_{i-1,j} + \xi_{ij}, \quad 1 \leq j \leq p, \tag{35}$$

where $\phi_j \sim U[0.2, 0.6]$ or $\phi_j \sim U[-0.6, -0.2]$, and the innovations a_{ij} is given below. To generate cross-dependence, let

$$\Sigma = (\sigma_{jk}) = (\rho^{|j-k|}), \quad 0 < \rho < 1,$$

and $\Sigma^{1/2}$ be the square-root matrix of Σ. We use $X_i = \Sigma^{1/2} Z_i$, $1 \leq i \leq n$, where $Z_i = (Z_{i1}, ..., Z_{ip})'$. We choose the true regression coefficient vector as

$$\beta^* = (3, ..., 3, 0, ..., 0)',$$

where the first 20 elements are all 3 and the rest are all 0. Let $\theta_i = 1/(1+\exp(-X_i'\beta^*))$ be the probability of success of the Bernoulli distribution of Y_i. Thus, Y_i is a random draw from Bernoulli(θ_i). We run the simulations for sample sizes $n = 200, 300, 400, 500, 800, 1000, 2000, 3000$, and choose the number of parameters p to be 400, dependence parameter $\rho = 0.5$. For each case, additional 500 observations are generated and used for out-of-sample predictions. To entertain various shapes of covariate distributions, we consider the following two scenarios for ξ_{ij} of (35):

1. $\xi_{ij} = 0.1 t_5$, i.e. the Student-t distribution with 5 degrees of freedom divided by 10;
2. $\xi_{ij} = 0.2 \log \text{Normal}(0, 0.5^2)$, i.e. a log-normal distribution with parameters 0 and 0.25^2 divided by 5.

They represent heavy-tailed symmetric and asymmetric distributions, respectively. To meet the model assumptions, the covariates are standardized to have *mean 0*. The constants used are chosen to ensure appropriate signal-to-noise ratio and θ_i not trivially equals to either 0 or 1 for better presentation. The numerical performance of the robust procedure and standard Lasso procedure under the two scenarios is evaluated by the following five measurements.

1. ℓ_2 error, which is defined as $|\hat{\beta} - \beta^*|_2$;
2. ℓ_1 error, which is defined as $|\hat{\beta} - \beta^*|_1$;
3. the number of false positive results, FP, which is the number of noise covariates that are selected;
4. the number of false negative results, FN, which is the number of signal covariates that are not selected;
5. one-step-ahead forecast errors of a total of 500 out-of-sample observations, FE, which is the misclassification rate.

For the robust Lasso, we choose the optimal tuning parameters λ and τ on the basis of 100 independent validation data sets. For each case, we run a two-dimensional grid search to find the best (λ, τ) pair that minimizes the misclassification rate of the 100 validation data sets. Then the optimal pair is used in the simulation. Similar methods are applied in choosing the tuning optimal parameters in other models. The means, over 5000 repetitions, of the five performance measures are summarized in Table 1.

The results of Table 1 show that our robust Lasso method has certain advantages over the standard Lasso method when the covariates are heavy-tailed. The results are in agreement with the theorems. As the sample size increases, the performance measures improve. In both symmetric and asymmetric covariates cases, our robust method has smaller ℓ_1 and ℓ_2 errors. The advantage of the proposed robust method is more pronounced when the sample size is large. In addition, FP increases slightly with the sample size n, but FN approaches zero as n increases.

We also investigate the empirical properties of the proposed method in linear regression. We again generate data from independent AR(1) model (35). Analogously to the logistic regression, we set $X_i = \Sigma^{1/2} Z_i$. For response, we generate time series process Y_i from the model,

$$Y_i = X_i' \beta^* + e_i, \tag{36}$$

where e_i is given below. The following two scenarios for ξ_{ij} are considered:

1. the Student-t distribution with 5 degrees of freedom, t_5;
2. a log-normal distribution with parameters 0 and 0.25^2, $\text{logNormal}(0, 0.25^2)$.

For the distribution of error e_i, we choose:

1. $e_t = 20t_3$, i.e. a Student-t distribution with 3 degrees of freedom multiplied by 20, the standard deviation of which is about 34.64;
2. $e_t = 20\,\text{logNormal}(0, 0.5^2)$, i.e. a log-normal distribution with parameters 0 and 0.5^2 times 20, the standard deviation of which is about 12.08.

Again, the covariates and the errors are standardized to have *mean 0*, and the constants used are chosen to ensure appropriate signal-to-noise ratio for better presentation. We set

Table 1. Simulation results of Lasso and robust Lasso (RLasso) for logistic regression ($p = 400, \rho = 0.5$), where n is the sample size. The results are averages over 5000 replications.

		Student t_5		LogNormal$(0, 0.25)$	
n	Scenario	Lasso	RLasso	Lasso	RLasso
200	ℓ_2 loss	11.53	11.32	11.59	11.39
	ℓ_1 loss	50.14	48.38	50.30	48.60
	FP	0.96	0.88	0.88	0.79
	FN	11.30	10.85	11.84	11.41
	FE	28.11%	27.24%	29.08%	28.50%
300	ℓ_2 loss	10.22	9.85	10.28	9.94
	ℓ_1 loss	43.69	40.89	43.80	41.15
	FP	1.69	1.51	1.61	1.44
	FN	5.98	5.37	6.59	5.92
	FE	22.02%	20.82%	23.09%	22.00%
400	ℓ_2 loss	9.46	8.95	9.48	9.05
	ℓ_1 loss	40.21	36.62	40.10	36.85
	FP	2.28	2.04	2.24	1.97
	FN	3.53	3.08	4.08	3.44
	FE	20.33%	19.24%	21.34%	20.23%
500	ℓ_2 loss	8.87	8.28	8.88	8.34
	ℓ_1 loss	37.56	33.51	37.38	33.61
	FP	2.64	2.32	2.54	2.23
	FN	2.23	1.84	2.66	2.14
	FE	19.44%	18.37%	20.40%	19.33%
800	ℓ_2 loss	7.66	6.82	7.67	6.93
	ℓ_1 loss	32.39	27.29	32.15	27.52
	FP	3.28	2.94	3.12	2.71
	FN	0.56	0.41	0.79	0.53
	FE	18.18%	17.17%	19.22%	18.13%
1000	ℓ_2 loss	7.15	6.20	7.11	6.26
	ℓ_1 loss	30.27	24.71	29.85	24.82
	FP	3.49	3.13	3.38	2.96
	FN	0.24	0.17	0.35	0.22
	FE	17.83%	16.82%	18.83%	17.77%
2000	ℓ_2 loss	5.67	4.42	5.61	4.43
	ℓ_1 loss	24.15	17.48	23.69	17.51
	FP	3.82	3.50	3.69	3.34
	FN	0.0018	0.0008	0.0060	0.0014
	FE	17.19%	16.22%	18.12%	17.05%
3000	ℓ_2 loss	4.92	3.53	4.84	3.52
	ℓ_1 loss	21.07	13.87	20.50	13.90
	FP	3.90	3.56	3.79	3.41
	FN	0	0	0	0
	FE	16.94%	15.98%	17.97%	16.93%

$n = 50$, 100, 200, 300, 400, 800, 1000, 2000, and use the root mean squared forecast error (RMSE) to measure one-step-ahead forecasts of a total of 200 out-of-sample predictions. The results are reported in Table 2.

The results of Table 2 are also in agreement with the theorem. In particular, as expected, the RMSE approaches the standard deviation of e_i as the sample size increases. In general, similarly to the logistic regression, the robust estimator outperforms the non-robust one. This is particularly so for the case of heavy-tailed noises $e_i \sim 20t_3$. But as the sample size increases, the difference between the robust procedure and the standard procedure gradually decreases. The out-of-sample predictions seem to work well when the sample size $n \geq p$. For log-normal errors, FN seems to be higher than FP. Both are sizable when the sample size is small.

In conclusion, our robust method is more flexible than the standard Lasso. The above two simultation studies show clearly that the robust procedure outperforms the standard procedure under the setting with heavy-tailed covariates and errors. The truncation parameter enables the robust method to render consistently satisfactory results under all scenarios considered in our simulation.

Next, we assess the sensitivity of the robust procedures in linear regression with respect to the thresholds τ_1 and τ_2. We use the same model setting as before for (36), and consider a heavy-tail case with t_5 features and t_3 noises, and a light-tail case with normally distributed covariates and noises. We set $n = 100, 800$, $p = 400$, and choose different quantiles of the feature values and responses as the thresholds τ_1, τ_2. Table 3 shows the ℓ_2 and ℓ_1 loss of the estimated coefficients $\hat{\beta}$ using different values of threshold τ_1, τ_2 over 1000 replications. When τ_1 and τ_2 are set to be the 100% quantile, the procedure becomes the original Lasso method without any shrinkage. For the heavy-tail setting, our robust procedure has smaller ℓ_2 and ℓ_1 estimation error than the vanilla Lasso method in all cases. For the light-tail setting, our shrinkage method loses some efficiency on ℓ_2 and ℓ_1 estimation error. Moreover, if τ_1, τ_2 are set to be above the 90% quantiles of the original feature values and responses, the difference between robust and non-robust methods in terms of estimation error is less than 2%.

Finally, we compare our proposed truncation method with the standard method under different degrees of the cross-sectional dependence of the covariates. The setup is the same as that used before in (36). We choose $n = 200$, $p = 400$, but vary $\rho = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.98$. The numerical result is based on 1000 independent Monte Carlo experiments. It can be seen from Table 4 that the proposed truncation method shows significant improvement in the statistical performance over the standard method in all cases of ρ. In addition, the differences in ℓ_2 and ℓ_1 loss for different values of ρ are minor. Note that the condition number of the population covariance increases as ρ grows, but is not too large.

6. Discussion

In this paper, we first established the statistical error bounds and the support recovery guarantees of high dimensional generalized linear models in time series setting. Both finite

Table 2. Simulation results of Lasso and robust Lasso (RLasso) for linear regression ($p = 400, \rho = 0.5$), where n is the sample size and the results are averages over 5000 replications

		Student t		LogNormal	
n	Scenario	Lasso	RLasso	Lasso	RLasso
50	ℓ_2 loss	24.01	22.31	35.11	31.90
	ℓ_1 loss	148.01	140.39	218.33	201.09
	FP	44.26	44.12	47.52	47.31
	FN	12.23	12.16	15.12	14.94
	RMSE	49.86	48.10	16.46	15.78
100	ℓ_2 loss	25.29	23.44	40.86	37.04
	ℓ_1 loss	200.83	187.76	331.69	302.13
	FP	51.31	50.79	57.69	57.12
	FN	8.05	7.73	11.52	11.18
	RMSE	49.72	47.66	17.19	16.39
200	ℓ_2 loss	12.46	11.04	24.52	22.78
	ℓ_1 loss	66.12	52.68	205.78	187.80
	FP	11.89	7.83	33.56	31.73
	FN	7.19	5.85	11.06	10.64
	RMSE	40.84	39.13	15.10	14.73
300	ℓ_2 loss	9.67	8.03	12.64	10.36
	ℓ_1 loss	39.98	36.83	55.58	52.13
	FP	3.49	3.34	7.81	7.91
	FN	5.78	4.47	10.30	10.25
	RMSE	38.25	37.37	13.34	13.26
400	ℓ_2 loss	8.77	8.04	11.72	9.24
	ℓ_1 loss	37.53	32.60	50.27	47.80
	FP	2.99	2.82	1.50	1.71
	FN	3.67	2.67	10.27	9.51
	RMSE	37.04	36.18	13.13	13.01
800	ℓ_2 loss	6.42	5.78	9.44	8.95
	ℓ_1 loss	25.13	22.34	38.17	36.04
	FP	2.56	2.52	1.37	1.61
	FN	0.86	0.41	5.94	4.67
	RMSE	35.11	34.65	12.58	12.51
1000	ℓ_2 loss	5.79	5.15	8.75	8.28
	ℓ_1 loss	22.53	19.84	34.74	32.82
	FP	2.58	2.49	0.71	0.90
	FN	0.45	0.16	4.29	3.29
	RMSE	34.82	34.42	12.46	12.40
2000	ℓ_2 loss	4.11	3.56	6.69	6.22
	ℓ_1 loss	15.85	12.60	25.65	23.94
	FP	2.22	2.13	0.39	0.49
	FN	0.0376	0.0014	1.10	0.67
	RMSE	34.39	34.14	12.26	12.23

Table 3. Sensitivity of different values of thresholds τ_1 and τ_2 in linear regression case using the upper quantiles of the feature values and responses. The results are based on 1000 replications.

	70%	75%	80%	85%	90%	95%	98%	99%	99.5%	100%
Student t and $n = 100, p = 400$										
ℓ_2 loss	23.51	23.39	23.30	23.25	23.29	23.45	23.82	24.15	24.53	25.70
ℓ_1 loss	188.56	187.21	186.08	185.24	185.35	186.49	189.38	192.26	195.41	204.39
Student t and $n = 800, p = 400$										
ℓ_2 loss	6.08	5.98	5.88	5.82	5.76	5.80	5.91	6.00	6.11	6.62
ℓ_1 loss	23.59	23.15	22.77	22.48	22.35	22.44	22.91	23.29	23.72	26.05
Normal and $n = 100, p = 400$										
ℓ_2 loss	20.72	20.53	20.36	20.20	20.09	19.97	19.90	19.91	19.92	19.91
ℓ_1 loss	166.34	164.38	162.44	160.72	159.35	158.14	157.52	157.41	157.36	157.26
Normal and $n = 800, p = 400$										
ℓ_2 loss	5.77	5.62	5.50	5.36	5.24	5.10	5.03	5.02	5.00	5.00
ℓ_1 loss	22.50	21.90	21.34	20.79	20.32	19.80	19.52	19.45	19.39	19.38

Table 4. Sensitivity of different values of the cross-sectional dependence of the covariates (ρ) in linear regression. The results are based on 1000 replications.

ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.98
Standard procedure										
ℓ_2 loss	12.36	12.29	12.43	12.54	12.23	12.63	12.30	12.33	12.54	12.78
ℓ_1 loss	65.11	64.10	65.48	67.43	63.23	67.68	64.87	65.24	67.85	70.75
Robust procedure										
ℓ_2 loss	10.93	10.88	10.77	10.78	10.81	11.11	10.88	11.03	11.04	10.89
ℓ_1 loss	51.47	51.15	48.89	49.40	49.64	53.84	51.06	53.44	53.14	51.28

polynomial moment case and exponential moment case are studied. In the finite moment case with geometric moment contraction, we also proposed a new robust M-estimator by shrinking the feature variables (and the response in linear regression) before solving the empirical risk minimization. The shrinkage method works as if we have sub-Gaussian data, and does not require any algorithmic adaptation. Our robust procedure marks a significant improvement over the existing literature in the time series setting by relaxing the sub-Gaussian condition to the existence of finite moments but retaining a nearly i.i.d. sub-Gaussian deviation bound.

It is worth mentioning that high dimensional time series creates overwhelming technical difficulties in theoretical analysis. Commonly used concentration inequalities in the i.i.d. case do not hold. For example, the sharpest available Bernstein-type inequalities for weakly dependent random variables without any structural assumptions (such as linear process) involve additional logarithmic factors. Whether the large deviation bound under the i.i.d. setting is achievable or not remains a challenging open problem. We conjecture that the scaling condition in the main theorems may be improved by developing sharper concentration inequalities.

Besides the concentration inequalities, there are many future research directions to pursue. The first direction is to study the single index model or the additive regression in time series setting using the framework of the functional dependence measure. Another interesting problem is the robust regularized estimation in some special high dimensional time series models, such as vector autoregressive (VAR) model, vector autoregressions with heteroskedasticity, etc. By exploiting the model structure more explicitly, better statistical error bounds may be established. In addition, though many inferential theories have been developed recently for high dimensional VAR models, the statistical inference problem in high dimensional time series regression remains an important and challenging task to investigate.

Acknowledgements

We would like to thank the Editor, the Associate Editor and the anonymous referees for their detailed and constructive reviews, which helped to improve the paper substantially. We would also like to thank Professor Cun-Hui Zhang for helpful comments and discussions.

References

[1] Avella-Medina, M., and Ronchetti, E. Robust and consistent variable selection in high-dimensional generalized linear models. *Biometrika 105*, 1 (12 2017), 31–44.

[2] Basu, S., and Michailidis, G. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics 43*, 4 (08 2015), 1535–1567.

[3] Bianco, A. M., and Yohai, V. J. Robust estimation in the logistic regression model. In *Robust Statistics, Data Analysis, and Computer Intensive Methods*. Springer, 1996, pp. 17–34.

[4] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*, 4 (08 2009), 1705–1732.

[5] Blundell, C., Beck, J., and Heller, K. A. Modelling reciprocating relationships with hawkes processes. In *Advances in Neural Information Processing Systems* (2012), pp. 2600–2608.

[6] Bousquet, O. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique 334*, 6 (2002), 495 – 500.

[7] Brown, E. N., Kass, R. E., and Mitra, P. P. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience 7*, 5 (2004), 456.

[8] Brownlees, C., Joly, E., and Lugosi, G. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics 43*, 6 (2015), 2507–2536.

[9] Bühlmann, P., and van de Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

[10] Burkholder, D. L. Distribution function inequalities for martingales. *The Annals of Probability 1*, 1 (02 1973), 19–42.

[11] Cantoni, E., and Ronchetti, E. Robust inference for generalized linear models. *Journal of the American Statistical Association 96*, 455 (2001), 1022–1030.

[12] Catoni, O. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 48*, 4 (2012), 1148–1185.

[13] Chen, L., and Wu, W. B. Concentration inequalities for empirical processes of linear time series. *Journal of Machine Learning Research 18* (2017), 231–1.

[14] Chen, X., Xu, M., and Wu, W. B. Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics 41*, 6 (12 2013), 2994–3021.

[15] Chen, X., Xu, M., and Wu, W. B. Regularized estimation of linear functionals of precision matrices for high-dimensional time series. *IEEE Transactions on Signal Processing 64*, 24 (2016), 6459–6470.

[16] Chernozhukov, V., Chetverikov, D., and Kato, K. Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields 162*, 1-2 (2014), 47–70.

[17] Chernozhukov, V., Chetverikov, D., and Kato, K. Inference on causal and structural parameters using many moment inequalities. *The Review of Economic Studies 86*, 5 (2019), 1867–1900.

[18] Ding, M., Mo, J., Schroeder, C. E., and Wen, X. Analyzing coherent brain networks with granger causality. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2011), IEEE, pp. 5916–5918.

[19] Duchi, J., Khosravi, K., and Ruan, F. Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics 46*, 6B (2018), 3246–3275.

[20] Fan, J., Gong, W., and Zhu, Z. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics 212*, 1 (2019), 177–202.

[21] Fan, J., Li, Q., and Wang, Y. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*, 1 (2017), 247–265.

[22] Fan, J., Liu, H., Sun, Q., and Zhang, T. I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics 46*, 2 (2018), 814.

[23] Fan, J., Wang, W., and Zhu, Z. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *The Annals of Statistics 49*, 3 (2021), 1239.

[24] Fan, Y., and Lv, J. Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association 108*, 503 (2013), 1044–1061.

[25] Genkin, A., Lewis, D. D., and Madigan, D. Large-scale bayesian logistic regression for text categorization. *Technometrics 49*, 3 (2007), 291–304.

[26] Guo, S., Wang, Y., and Yao, Q. High-dimensional and banded vector autoregressions. *Biometrika* (2016), asw046.

[27] Hall, E. C., Raskutti, G., and Willett, R. M. Learning high-dimensional generalized linear autoregressive models. *IEEE Transactions on Information Theory 65*, 4 (2018), 2401–2422.

[28] Hampel, F. R. A general qualitative definition of robustness. *The Annals of Mathematical Statistics* (1971), 1887–1896.

[29] Hampel, F. R. The influence curve and its role in robust estimation. *Journal of the American Statistical Association 69*, 346 (1974), 383–393.

[30] Han, F., Lu, H., and Liu, H. A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research* (2015).

[31] Han, Y., and Tsay, R. S. High-dimensional linear regression for dependent observations with application to nowcasting. *Statistica Sinica 30* (2020), 1797–1827.

[32] Hausman, J. A., Lo, A. W., and MacKinlay, A. An ordered probit analysis of transaction stock prices. *Journal of Financial Economics 31*, 3 (1992), 319 – 379.

[33] Hill, R. W. *Robust regression when there are outliers in the carriers*. PhD thesis, Harvard University, 1977.

[34] Hodges Jr, J. L. Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (1967), vol. 1, pp. 163–186.

[35] Hsu, D., and Sabato, S. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research 17*, 1 (2016), 543–582.

[36] Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C.-H. Oracle inequalities for the lasso in the cox model. *The Annals of Statistics 41*, 3 (06 2013), 1142–1165.

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

[38] Huang, S.-J., and Shih, K.-R. Short-term load forecasting via arma model identification including non-gaussian process considerations. *IEEE Transactions on power systems 18*, 2 (2003), 673–679.

[39] Huber, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics* (1964), 73–101.

[40] Huber, P. J. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (1967), vol. 1, Berkeley, CA, pp. 221–233.

[41] Ivanoff, S., Picard, F., and Rivoirard, V. Adaptive lasso and group-lasso for functional poisson regression. *Journal of Machine Learning Research 17*, 55 (2016), 1–46.

[42] Jiang, X., Raskutti, G., and Willett, R. Minimax optimal rates for poisson inverse problems with physical constraints. *IEEE Transactions on Information Theory 61*, 8 (2015), 4458–4474.

[43] Krishnapuram, B., Carin, L., Figueiredo, M. A. T., and Hartemink, A. J. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 6 (2005), 957–968.

[44] Linderman, S., and Adams, R. Discovering latent network structure in point process data. In *International Conference on Machine Learning* (2014), pp. 1413–1421.

[45] Loh, P.-L. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics 45*, 2 (04 2017), 866–896.

[46] Loh, P.-L., and Wainwright, M. J. Support recovery without incoherence: A

case for nonconvex regularization. *The Annals of Statistics 45*, 6 (2017), 2455–2482.

[47] Lokhorst, J. The lasso and generalised linear models. *Honors Project, The University of Adelaide, Australia* (1999).

[48] Mallows, C. L. On some topics in robustness. *Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ* (1975).

[49] Mark, B., Raskutti, G., and Willett, R. Network estimation from point process data. *IEEE Transactions on Information Theory 65*, 5 (2018), 2953–2975.

[50] Massart, P. About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability 28*, 2 (2000), 863–884.

[51] McCullagh, P., and Nelder, J. A. *Generalized Linear Models*, vol. 37. Cambridge University Press, Cambridge, 1989.

[52] Meier, L., Van De Geer, S., and Bühlmann, P. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*, 1 (2008), 53–71.

[53] Merlevède, F., Peligrad, M., and Rio, E. A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields 151*, 3-4 (2011), 435–474.

[54] Merrill, H. M., and Schweppe, F. C. Bad data suppression in power system static state estimation. *IEEE Transactions on Power Apparatus and Systems*, 6 (1971), 2718–2725.

[55] Minsker, S. Geometric median and robust estimation in banach spaces. *Bernoulli 21*, 4 (2015), 2308–2335.

[56] Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science 27*, 4 (11 2012), 538–557.

[57] Ogata, Y. Seismicity analysis through point-process modeling: A review. In *Seismicity Patterns, Their Statistical Significance and Physical Meaning*. Springer, 1999, pp. 471–507.

[58] Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., and Simoncelli, E. P. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature 454*, 7207 (2008), 995.

[59] Pinelis, I. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability* (1994), 1679–1706.

[60] Priestley, M. B. *Non-linear and Non-stationary Time Series Analysis*. Academic Press, 1988.

[61] Raginsky, M., Jafarpour, S., Harmany, Z. T., Marcia, R. F., Willett, R. M., and Calderbank, R. Performance bounds for expander-based compressed sensing in poisson noise. *IEEE Transactions on Signal Processing 59*, 9 (2011), 4139–4153.

[62] Raginsky, M., Willett, R. M., Harmany, Z. T., and Marcia, R. F. Compressed sensing performance bounds under poisson noise. *IEEE Transactions on Signal Processing 58*, 8 (2010), 3990–4002.

[63] Raginsky, M., Willett, R. M., Horn, C., Silva, J., and Marcia, R. F. Sequential anomaly detection in the presence of noise and limited feedback. *IEEE*

Transactions on Information Theory 58, 8 (2012), 5544–5562.

[64] Raskutti, G., Wainwright, M. J., and Yu, B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory 57*, 10 (2011), 6976–6994.

[65] Rosenblatt, M. *Markov Processes: Structure and Asymptotic Behavior.* Springer, 1971.

[66] Rosenthal, H. On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics 8*, 3 (1970), 273–303.

[67] Roth, V. The generalized lasso. *IEEE transactions on neural networks 15*, 1 (2004), 16–28.

[68] Shao, X., and Wu, W. B. Asymptotic spectral theory for nonlinear time series. *The Annals of Statistics 35*, 4 (08 2007), 1773–1801.

[69] Shevade, S. K., and Keerthi, S. S. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics 19*, 17 (2003), 2246–2253.

[70] Silva, J., and Willett, R. Hypergraph-based anomaly detection of high-dimensional co-occurrences. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 3 (2008), 563–569.

[71] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 1 (1996), 267–288.

[72] Tibshirani, R. The lasso method for variable selection in the cox model. *Statistics in Medicine 16*, 4 (1997), 385–395.

[73] Tong, H. *Non-linear Time Series: A Dynamical System Approach.* Oxford University Press, 1990.

[74] Tsay, R. S. *Analysis of Financial Time Series*, vol. 543. John Wiley & Sons, 2005.

[75] Tsay, R. S., and Chen, R. *Nonlinear Time Series Analysis.* Wiley Series in Probability and Statistics. Wiley, 2018.

[76] Tukey, J. W. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics* (1960), 448–485.

[77] Tukey, J. W. The future of data analysis. *The Annals of Mathematical Statistics 33*, 1 (1962), 1–67.

[78] van de Geer, S., and Müller, P. Quasi-likelihood and/or robust estimation in high dimensions. *Statistical Science* (2012), 469–480.

[79] van de Geer, S. A. High-dimensional generalized linear models and the lasso. *The Annals of Statistics 36*, 2 (04 2008), 614–645.

[80] Vere-Jones, D., and Ozaki, T. Some examples of statistical estimation applied to earthquake data. *Annals of the Institute of Statistical Mathematics 34*, 1 (1982), 189–207.

[81] Wiener, N. *Nonlinear Problems in Random Theory.* Wiley, New York, 1958.

[82] Wu, W. B. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America 102*, 40 (2005), 14150–14154.

[83] Wu, W. B. M-estimation of linear models with dependent errors. *The Annals of Statistics 35*, 2 (2007), 495–521.

[84] Wu, W. B., and Min, W. On linear processes with dependent innovations.

Stochastic Processes and their Applications 115, 6 (2005), 939–958.

[85] Wu, W. B., and Shao, X. Limit theorems for iterated random functions. *Journal of Applied Probability 41*, 2 (2004), 425–436.

[86] Wu, W.-B., and Wu, Y. N. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics 10*, 1 (2016), 352–379.

[87] Zhang, C., Guo, X., Cheng, C., and Zhang, Z. Robust-BD estimation and inference for varying-dimensional general linear models. *Statistica Sinica* (2014), 653–673.

[88] Zhang, D. Robust estimation of the mean and covariance matrix for high dimensional time series. *Statistica Sinica 31*, 2 (2021), 797–820.

[89] Zhang, D., and Wu, W. B. Gaussian approximation for high dimensional time series. *The Annals of Statistics 45*, 5 (2017), 1895–1919.

[90] Zhou, H. H., and Raskutti, G. Non-parametric sparse additive auto-regressive network models. *IEEE Transactions on Information Theory 65*, 3 (2018), 1473–1492.

[91] Zhou, K., Zha, H., and Song, L. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics* (2013), pp. 641–649.

Supplementary Material to “High Dimensional Generalized Linear Models for Temporal Dependent Data”

YUEFENG HAN, RUEY S. TSAY, and WEI BIAO WU

Appendix A: Concentration Inequalities for High Dimensional Time Series

The proofs of main theorems require additional new concentration inequalities for high dimensional time series. Analogously to Bousquet’s inequality ([6]) for i.i.d data, we present concentration inequalities for both heavy-tailed and light-tailed high dimensional time series under the framework of functional dependence measure. The result may be of independent interest. Without loss of generality, assume $\mathbb{E}X_{ij} = 0$ in this section. Denote $\mathcal{F}_i^l = \sigma(\varepsilon_l, ..., \varepsilon_i)$ with $l \le i$, $\mathcal{F}_i = \sigma(\cdots, \varepsilon_{i-1}, \varepsilon_i)$. Write $\mathcal{P}_l(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_l) - \mathbb{E}(\cdot|\mathcal{F}_{l-1})$. The same notation as that in the main paper is used.

Theorem A.1. *Let $t = \log p \vee 1$ and $C_{q,\alpha}$ be a constant depending on q and α. Assume $\||X_{\cdot}|_\infty\|_{q,\alpha} < \infty$, where $q > 2$ and $\alpha > 0$. (i) If $\alpha > 1/2 - 1/q$, then for $x \gtrsim \sqrt{nt}\max_{1\le j\le p}\|X_{\cdot j}\|_{2,\alpha} + n^{1/q}t^{3/2}\||X_{\cdot}|_\infty\|_{q,\alpha}$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right|_\infty \ge x\right) \le \frac{C_{q,\alpha} n t^{q/2} \||X_{\cdot}|_\infty\|_{q,\alpha}^q}{x^q} + C_{q,\alpha}\exp\left(-\frac{C_{q,\alpha}x^2}{n\max_{1\le j\le p}\|X_{\cdot j}\|_{2,\alpha}^2}\right). \tag{37}$$

(ii) If $0 < \alpha < 1/2 - 1/q$, then for $x \gtrsim \sqrt{nt}\max_{1\le j\le p}\|X_{\cdot j}\|_{2,\alpha} + n^{1/2-\alpha}t^{3/2}\||X_{\cdot}|_\infty\|_{q,\alpha}$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right|_\infty \ge x\right) \le \frac{C_{q,\alpha} n^{q/2-\alpha q} t^{q/2} \||X_{\cdot}|_\infty\|_{q,\alpha}^q}{x^q} + C_{q,\alpha}\exp\left(-\frac{C_{q,\alpha}x^2}{n\max_{1\le j\le p}\|X_{\cdot j}\|_{2,\alpha}^2}\right). \tag{38}$$

Proof. For $t = 1 \vee \log p$, note that for any vector $u = (u_1, ..., u_p)^\top$, $|u|_\infty \le |u|_t \le p^{1/t}|u|_\infty = e|u|_\infty$. Let $L = \lfloor \log n/\log 2 \rfloor$, $\tau_l = 2^l$ if $1 \le l \le L$, $\tau_L = n$ and $\tau_0 = 0$. To simplify the notation, let $W_n = \sum_{i=1}^n X_i$, $W_{n,k} = \sum_{i=1}^n \mathbb{E}(X_i|\varepsilon_{i-k}, ..., \varepsilon_i)$, $X_{i,k} = \mathbb{E}(X_i|\varepsilon_{i-k}, ..., \varepsilon_i)$. Define $Q_{n,l} = W_{n,\tau_l} - W_{n,\tau_{l-1}}$ for $1 \le l \le L$, and write

$$W_n = W_{n,0} + W_n - W_{n,n} + \sum_{l=1}^L Q_{n,l}. \tag{39}$$

Note that $W_{n,n} - W_{n,0} = \sum_{l=1}^{L} Q_{n,l}$. By Lemma A.1 and Jensen's inequality,

$$\begin{aligned}\||W_{n,k+1} - W_{n,k}|_\infty\|_q &= \left\|\left|\sum_{i=1}^{n}[\mathbb{E}(X_i|\mathcal{F}_i^{i-k-1}) - \mathbb{E}(X_i|\mathcal{F}_i^{i-k})]\right|_t\right\|_q \\ &\le C_q\sqrt{nt}\,\left\|\left|\mathbb{E}(X_i|\mathcal{F}_i^{i-k-1}) - \mathbb{E}(X_i|\mathcal{F}_i^{i-k})\right|_t\right\|_q \\ &\le C_q\sqrt{nt}\,\left\||X_i - X_{i,i-k-1}|_t\right\|_q \\ &\le C_q\sqrt{nt}\omega_{k+1,q}.\end{aligned}$$

Then,

$$\||W_n - W_{n,n}|_\infty\|_q \le \sum_{k=n}^{\infty}\||W_{n,k+1} - W_{n,k}|_\infty\|_q \le \sum_{k=n}^{\infty} C_q\sqrt{nt}\omega_{k+1,q} = C_q\sqrt{nt}\Omega_{n+1,q}.$$

By Markov's inequality, we have

$$\mathbb{P}(|W_n - W_{n,n}|_t \ge x) \le \frac{\||W_n - W_{n,n}|_t\|_q^q}{x^q} \le \frac{C_q(nt)^{q/2}\Omega_{n+1,q}^q}{x^q}. \tag{40}$$

Note that $\Omega_{n+1,q} \le \||X_\cdot|_\infty\|_{q,\alpha} n^{-\alpha}$.

Recall $X_{i,0} = \mathbb{E}(X_i|\varepsilon_i)$ and $W_{n,0} = \sum_{i=1}^{n}\mathbb{E}[X_i|\varepsilon_i]$. Note that $\mathbb{E}[X_i|\varepsilon_i]$ are independent for different i. By Fuk-Nagaev inequality in Lemma D.2 of [17], we have

$$\begin{aligned}\mathbb{P}(|W_{n,0}|_\infty - 2\mathbb{E}|W_{n,0}|_\infty \ge x) \le& \frac{C_{q,\alpha}\sum_{i=1}^{n}\mathbb{E}\max_{1\le j\le p}|\mathbb{E}[X_i|\varepsilon_i]|^q}{x^q} \\ &+ \exp\left(-\frac{x^2}{3\max_{1\le j\le p}\sum_{i=1}^{n}\mathbb{E}|\mathbb{E}[X_i|\varepsilon_i]|^2}\right) \\ \le& \frac{C_{q,\alpha}n\||X_{i,0}|_\infty\|_q^q}{x^q} + \exp\left(-\frac{x^2}{3n\max_{1\le j\le p}\|X_{ij,0}\|_2^2}\right).\end{aligned}$$

Then, by Lemma D.3 of [17],

$$\begin{aligned}\mathbb{E}|W_{n,0}|_\infty &\lesssim \sqrt{t}\sqrt{\max_j\sum_{i=1}^{n}\mathbb{E}(\mathbb{E}[X_i|\varepsilon_i])^2} + t\sqrt{\mathbb{E}\max_i\max_j|\mathbb{E}[X_i|\varepsilon_i]|^2} \\ &\lesssim \sqrt{nt}\max_j\|X_{ij,0}\|_2 + tn^{1/q}\left[\mathbb{E}\max_j|\mathbb{E}[X_i|\varepsilon_i]|^q\right]^{1/q} \\ &\lesssim \sqrt{nt}\max_j\|X_{ij,0}\|_2 + tn^{1/q}\||X_{i,0}|_\infty\|_q.\end{aligned}$$

Hence $\mathbb{E}|W_{n,0}|_\infty \lesssim x$, which implies that

$$\mathbb{P}(|W_{n,0}|_\infty \ge x) \quad \le \quad \frac{C_1 n\||X_{i,0}|_\infty\|_q^q}{x^q} + C_2\exp\left(-\frac{x^2}{n\max_j\|X_{ij,0}\|_2^2}\right). \tag{41}$$

Finally, for each $1 \le l \le L$ and $1 \le i \le \lfloor n/\tau_l \rfloor$, define

$$
\begin{aligned}
Z_{i,l} &= \sum_{k=(i-1)\tau_l+1}^{i\tau_l \wedge n} \left[\mathbb{E}(X_i | \mathcal{F}_k^{k-\tau_l}) - \mathbb{E}(X_i | \mathcal{F}_k^{k-\tau_{l-1}}) \right];\\
U_{n,l}^e &= \sum_{i \text{ is even}} Z_{i,l} \quad \text{and} \quad U_{n,l}^o = \sum_{i \text{ is odd}} Z_{i,l}.
\end{aligned}
$$

Let $c = q/2 - 1 - \alpha q$, $\lambda_l = l^2/(\pi^2/3)$ if $1 \le l \le L/2$ and $\lambda_l = (L+1-l)^{-2}/(\pi^2/3)$ if $L/2 < l \le L$. Then, $\sum_{l=1}^L \lambda_l < 1$. Since $Z_{i,l}$ and $Z_{i',l}$ are independent for $|i - i'| > 1$, by Lemma A.2

$$
\mathbb{P}(|U_{n,l}^e|_t - 2\mathbb{E}|U_{n,l}^e|_t \ge \lambda_l x) \le \frac{C_q \sum\limits_{i \text{ is even}} \mathbb{E}|Z_{i,l}|_t^q}{(\lambda_l x)^q} + \exp\left(-\frac{(\lambda_l x)^2}{3 \sum\limits_{i \text{ is even}} |\sigma_{Z_{i,l}}|_t^2} \right),
$$

where $\sigma_{Z_{i,l}} = (\|Z_{i1,l}\|_2, ..., \|Z_{im,l}\|_2)^\top$.

By Lemma A.1, we can obtain

$$
\| |Z_{i,l}|_t \|_q \le C_q (\tau_l t)^{1/2} \tilde{\omega}_{l,q}, \quad \text{where } \tilde{\omega}_{l,q} = \sum_{\tau_{l-1}+1}^{\tau_l} \omega_{k,q} \le \tau_{l-1}^{-\alpha} \| |X_\cdot|_\infty \|_{q,\alpha}.
$$

Similarly, we can define $\tilde{\delta}_{l,2,j}$. By Theorem 3.2 of [10] and Jensen's inequality,

$$
\begin{aligned}
\|Z_{ij,l}\|_2 &\le \sqrt{\tau_l} \sum_{k=\tau_{l-1}+1}^{\tau_l} \left\| \mathbb{E}(X_{ij} | \mathcal{F}_i^{i-k}) - \mathbb{E}(X_{ij} | \mathcal{F}_i^{i-k+1}) \right\|_2 \\
&\le \sqrt{\tau_l} \sum_{k=\tau_{l-1}+1}^{\tau_l} \|X_{ij,k} - X_{ij,k-1}\|_2 \\
&\le \sqrt{\tau_l} \sum_{k=\tau_{l-1}+1}^{\tau_l} \delta_{k,2,j} \\
&= \sqrt{\tau_l} \tilde{\delta}_{l,2,j}.
\end{aligned}
$$

This implies that $|\sigma_{Z_{i,l}}|_t \lesssim \max_j \sqrt{\tau_l} \tau_{l-1}^{-\alpha} \|X_{\cdot j}\|_{2,\alpha}$. So we obtain, for $x > 0$,

$$
\begin{aligned}
\mathbb{P}(|U_{n,l}^e|_t - 2\mathbb{E}|U_{n,l}^e|_t \ge \lambda_l x) &\le \frac{C_q n t^{q/2} \tau_l^{q/2-1} \tilde{\omega}_{l,q}^q}{\lambda_l^q x^q} + \exp\left(-\frac{C_q \lambda_l^2 x^2 \tau_{l-1}^{2\alpha}}{n \max_j \|X_{\cdot j}\|_{2,\alpha}} \right) \\
&\le \frac{C_q n t^{q/2}}{x^q} \cdot \frac{\tau_l^{q/2-1} \tau_{l-1}^{-\alpha q} \| |X_\cdot|_\infty \|_{q,\alpha}}{\lambda_l^q} + \exp\left(-\frac{C_q x^2 \lambda_l^2 \tau_{l-1}^{2\alpha}}{n \max_j \|X_{\cdot j}\|_{2,\alpha}} \right).
\end{aligned}
$$

By Lemma 8 in [16],

$$\begin{aligned}
\mathbb{E}|U_{n,l}^e|_t &\lesssim \sqrt{t}\sqrt{\max_j \sum_{i \text{ is even}} \mathbb{E}Z_{ij,l}^2} + t\sqrt{\mathbb{E}\max_i \max_j Z_{ij,l}^2} \\
&\lesssim \sqrt{t}\sqrt{\max_j \sum_{i \text{ is even}} \mathbb{E}Z_{ij,l}^2} + t(n/\tau_l)^{1/q}\||Z_{i,l}|_t\|_q \\
&\lesssim \sqrt{nt}\max_j \tilde{\delta}_{l,2,j} + n^{1/q}t^{3/2}\tau_l^{1/2-1/q}\tilde{\omega}_{l,q} \\
&\lesssim \sqrt{nt}\max_j \tau_l^{-\alpha}\|X_{\cdot j}\|_{2,\alpha} + n^{1/q}t^{3/2}\tau_l^{-c/q}\||X_{\cdot}|_\infty\|_{q,\alpha}.
\end{aligned}$$

Notice that $\lambda_l^{-1}\tau_l^{c/q} \lesssim n^{c/q}$ for $c > 0$ and $\min_{c\geq 0}\lambda_l\tau_l^{-c/q} > 1$ for $c < 0$ and $\min_{l\geq 0}\lambda_l\tau_l^{\alpha} > 1$. Hence $\mathbb{E}|U_{n,l}^e|_t \lesssim \lambda_l x$ always holds. Therefore,

$$\mathbb{P}(|U_{n,l}^e|_t \geq \lambda_l x) \leq \frac{C_3 n t^{q/2}}{x^q}\cdot\frac{\tau_l^{q/2-1}\tau_{l-1}^{-\alpha q}\||X_{\cdot}|_\infty\|_{q,\alpha}}{\lambda_l^q} + \exp\left(-\frac{C_4 x^2\lambda_l^2\tau_{l-1}^{2\alpha}}{n\max_j\|X_{\cdot j}\|_{2,\alpha}}\right). \tag{42}$$

A similar inequality holds for $U_{n,l}^o$. Let

$$A = \sum_{l=1}^{L}\frac{\tau_l^c}{\lambda_l^q} \quad \text{and} \quad B = \sum_{l=1}^{L}\exp\left\{-\frac{C_5 x^2\lambda_l^2\tau_l^{2\alpha}}{n\max_j\|X_{\cdot j}\|_{2,\alpha}^2}\right\}.$$

Since $\sum_{l=1}^{L}\lambda_l \leq 1$ and $|Q_{n,l}|_t \leq |U_{n,l}^e|_t + |U_{n,l}^o|_t$, by (42),

$$\begin{aligned}
\mathbb{P}(|\sum_{l=1}^{L}Q_{n,l}|_t \geq 2x) &\leq \sum_{l=1}^{L}\mathbb{P}(|Q_{n,l}|_t \geq 2\lambda_l x) \\
&\leq \sum_{l=1}^{L}[\mathbb{P}(|U_{n,l}^e|_t \geq \lambda_l x) + \mathbb{P}(|U_{n,l}^o|_t \geq \lambda_l x)] \\
&\leq \frac{C_6 n t^{q/2}\||X_{\cdot}|_\infty\|_{q,\alpha}^q}{x^q}A + C_7 B. \qquad (43)
\end{aligned}$$

Let $\psi := \min_{l\geq 1}\lambda_l^2\tau_l^{2\alpha} > 0$. By the definition of τ_l and λ_l and by elementary calculations, there exists a constant $C_8 > 1$ such that for all $y \geq 1$,

$$\sum_{l=1}^{L}\exp\left\{-C_5 y\lambda_l^2\tau_l^{2\alpha}\right\} \leq C_8\exp\{-C_5 y\psi\}. \tag{44}$$

We apply (44) with $y = x^2/(n\max_j\|X_{\cdot j}\|_{2,\alpha}^2)$. If $c > 0$, it can be obtained that $A \leq C_9\lambda_l^c \leq C_9 n^c$. If $c < 0$, then $A \leq C_{10}$. Hence, combining (39), (40), (41), (43) and (44), if $c > 0$ and $x \gtrsim \sqrt{nt}\max_j\|X_{\cdot j}\|_{2,\alpha} + n^{1/q+c/q}t^{3/2}\||X_{\cdot}|_\infty\|_{q,\alpha}$,

$$\mathbb{P}(|W_n|_t \geq x) \leq \exp\left\{-\frac{C_{q,\alpha}x^2}{n\max_j\|X_{\cdot j}\|_{2,\alpha}^2}\right\} + \frac{C_{q,\alpha}n^{c+1}t^{q/2}\||X_{\cdot}|_\infty\|_{q,\alpha}^q}{x^q}, \tag{45}$$

if $c<0$ and $x \gtrsim \sqrt{nt}\max_j \|X_{\cdot j}\|_{2,\alpha} + n^{1/q}t^{3/2}\||X_\cdot|_\infty\|_{q,\alpha}$,

$$\mathbb{P}(|W_n|_t \geq x) \leq \exp\left\{-\frac{C_{q,\alpha}x^2}{n\max_j \|X_{\cdot j}\|_{2,\alpha}^2}\right\} + \frac{C_{q,\alpha}nt^{q/2}\||X_\cdot|_\infty\|_{q,\alpha}^q}{x^q}. \tag{46}$$

By (45) and (46), both cases with $c<0$ and $c>0$ of Theorem A.1 follow.

□

Theorem A.2. *(i). Assume* $\||X_\cdot|_\infty\|_{\psi_\nu} < \infty$, *where* $\nu \geq 0$. *Let* $\alpha = 2/(1+2\nu)$, *then there exists a constant* $C_\nu > 0$ *depending on* ν *such that*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right|_\infty \geq x\right) \leq C_\nu \exp\left(-\frac{x^\alpha}{2e\alpha(n\log p)^{\alpha/2}\||X_\cdot|_\infty\|_{\psi_\nu}^\alpha}\right). \tag{47}$$

(ii). Assume $\|X_{\cdot j}\|_{\psi_\nu} < \infty$, *where* $\nu \geq 0$. *Let* $\alpha = 2/(1+2\nu)$, *then there exists a constant* $C_\nu > 0$ *depending on* ν *such that*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_{ij}\right| \geq x\right) \leq C_\nu \exp\left(-\frac{x^\alpha}{2e\alpha(n)^{\alpha/2}\|X_{\cdot j}\|_{\psi_\nu}^\alpha}\right). \tag{48}$$

Proof. We only prove part (i), as the proof of part (ii) is similar. Let $u_0 = (e\alpha\||X_\cdot|_\infty\|_{\psi_\nu}^\alpha)^{-1}$ and $t = 1 \vee \log p$. Note that for any vector $u = (u_1,...,u_p)^\top$, $|u|_\infty \leq |u|_t \leq p^{1/t}|u|_\infty = e|u|_\infty$. Let $L = \lfloor \log n/\log 2 \rfloor$, $\tau_l = 2^l$ if $1 \leq l \leq L$, $\tau_L = n$ and $\tau_0 = 0$. Let $W_n = \sum_{i=1}^n X_i$, $W_{n,k} = \sum_{i=1}^n \mathbb{E}(X_i|\varepsilon_{i-k},...,\varepsilon_i)$, $X_{i,k} = \mathbb{E}(X_i|\varepsilon_{i-k},...,\varepsilon_i)$. Define $Q_{n,l} = \sum_{i=1}^n \mathcal{P}_{i-l}X_i$. Then, $Q_{n,l}$ is a martingale. By Lemma A.1 and Jensen's inequality,

$$\begin{aligned}\||Q_{n,l}|_\infty\|_q &\leq C\sqrt{qnt}\,\||\mathbb{E}(X_i|\mathcal{F}_{i-l}) - \mathbb{E}(X_i|\mathcal{F}_{i-l-1})|_t\|_q \\ &\leq C\sqrt{qnt}\,\left\||X_i - X_{i,i-l}|_t\right\|_q \\ &= C\sqrt{qnt}\omega_{l,q}.\end{aligned}$$

Then,

$$\||W_n|_t\|_q \leq \sum_{l=0}^\infty \||Q_{n,l}|_t\|_q \leq \sum_{l=0}^\infty C\sqrt{qnt}\omega_{l,q} = C\sqrt{qnt}\Omega_{0,q}.$$

Let $Z_n = W_n/(\sqrt{nt})$. Then $\||Z_n|_t\|_q \leq C\sqrt{q}\Omega_{0,q}$. Write the negative binomial expansion $(1-s)^{-1/2} = 1 + \sum_{k=1}^\infty a_k s^k$, where $|s| < 1$ and $a_k = (2k)!/(2^{2k}(k!)^2)$. By Stirling's formula, as $k \to \infty$, $a_k \sim (k\pi)^{-1/2}$. Hence $k! \sim \sqrt{2}(k/e)^k a_k^{-1}$, and there exist absolute constants $c_1, c_2 > 0$ such that $c_1(k/e)^k a_k^{-1} \leq k! \leq c_2(k/e)^k a_k^{-1}$ holds for all $k \geq 1$. If $\alpha k > 2$, we have $\Omega_{0,\alpha k} \leq (\alpha k)^\nu \||X_\cdot|_\infty\|_{\psi_\nu}$. Hence, by elementary manipulations,

$$\frac{u^k\||Z_n|_t^\alpha\|_k^k}{k!} \leq \frac{u^k(\alpha k)^{\alpha k/2}\Omega_{0,\alpha k}^{\alpha k}}{c_1(k/e)^k a_k^{-1}} \leq \frac{a_k u^k}{c_1 u_0^k} \tag{49}$$

If $\alpha k < 2$, then $\||Z_n|_t\|_{\alpha k} \le \||Z_n|_t\|_2 \le 2^\nu \||X_\cdot|_\infty\|_{\psi_\nu}$. Using $e^x = \sum_{k=0}^\infty x^k/k!$, we obtain,

$$\begin{aligned}
\sup_n \mathbb{E}\exp\{u|Z_n|_t^\alpha\} &\le 1 + \sum_{1\le k<2/\alpha} \frac{u^k(2^\nu\||X_\cdot|_\infty\|_{\psi_\nu})^{\alpha k}}{k!} + \sum_{k\ge 2/\alpha} \frac{a_k u^k}{c_1 u_0^k} \\
&\le 1 + c'_\alpha \sum_{k=1}^\infty a_k \frac{u^k}{u_0^k} \\
&\le 1 + c_\alpha \frac{u/u_0}{(1-u/u_0)^{1/2}},
\end{aligned}$$

where constants $c_\alpha, c'_\alpha > 0$ only depend on α. Let $u = u_0/2$, then $\sup_n \mathbb{E}\exp\{u|Z_n|_t^\alpha\} \le 1 + c_\alpha/\sqrt{2}$. Hence,

$$\mathbb{P}(|W_n|_t \ge x) = \mathbb{P}(|Z_n|_t \ge \frac{x}{\sqrt{nt}}) \le (1 + c_\alpha/\sqrt{2})\exp\left\{-\frac{x^\alpha}{2e\alpha(nt)^{\alpha/2}\||X_\cdot|_\infty\|_{\psi_\nu}^\alpha}\right\}. \quad (50)$$

Clearly, Theorem A.2(i) follows from (50).

□

Theorem A.3. *Assume $\mathbb{E}X_i = 0$, $|X_i| \le M$ for all i, and $\|X_\cdot\|_{2,\mathrm{GMC}} < \infty$ for some $\rho \in (0,1)$. Also assume $n \ge 4 \vee (\log(\rho^{-1})/2)$. For any $x > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge x\right) \le \exp\Big\{-\frac{x^2}{4C_1(n\|X_\cdot\|_{2,\mathrm{GMC}}^2 + M^2) + 2C_2 M(\log n)^2 x}\Big\}, \quad (51)$$

where $C_1 = 2\max\{(e^4-5)/4, [\rho(1-\rho)\log(\rho^{-1})]^{-1}\}\cdot(8\vee\log(\rho^{-1}))^2$, $C_2 = \max\{(c\log 2)^{-1}, [1\vee(\log(\rho^{-1})/8)]\}$ with $c = [\log(\rho^{-1})/8] \wedge \sqrt{(\log 2)\log(\rho^{-1})/4}$.

Proof. See Theorem 2.1 in [88].

□

Theorem A.4. *Assume $\mathbb{E}X_i = 0$, $\mathbb{P}(|X_i| > x) \le c_1\exp(-c_2 x^\nu)$ for some $\nu > 0, c_1, c_2 > 0$, and $\|X_\cdot\|_{2,\mathrm{GMC}} < \infty$ for some $\rho \in (0,1)$. For any $x > 0$,*

$$\begin{aligned}
\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \ge x\right) &\le \exp\left(-\frac{x^2}{C_1(1+n\|X_\cdot\|_{2,\mathrm{GMC}}^2)}\right) + n\exp\left(-\frac{x^{\nu/(1+\nu)}}{C_2}\right) \\
&\quad + \exp\left(-\frac{x^2}{C_3 n}\exp\left(\frac{x^{\nu/(1+\nu)^2}}{C_4(\log x)^{\nu/(1+\nu)}}\right)\right), \quad (52)
\end{aligned}$$

where $C_1, C_2, C_3, C_4 > 0$ are constants only depending on ρ.

Proof. Consider the coefficient $\tau(M, X)$ of weak dependence in [53], which is defined by

$$\tau(M, X) = \left\| \sup_{f\in\Lambda_1(\mathbb{R}^k)} |\mathbb{E}f(X|M) - \mathbb{E}f(X)| \right\|_1,$$

where $\Lambda_1(\mathbb{R}^k)$ is the set of 1-Lipschitz functions from $\mathbb{R}^k$ to $\mathbb{R}$. The τ-mixing coefficients $\tau(i)$ of a sequence of random variables X_i are then defined by

$$\tau(i) = \sup_{k\geq 0} \max_{1\leq \ell \leq k} \frac{1}{\ell} \sup \left\{ \tau\big(\sigma(X_j, j\leq 0), (X_{j_1}, ..., X_{j_\ell})\big), i \leq j_1 < \cdots < j_\ell \right\}.$$

By the definition of geometric moment contraction and $\|X_\cdot\|_{2,\mathrm{GMC}} < \infty$, we have

$$\tau(i) \leq c_0 \rho^i,$$

for some positive constant c_0.

For any $M > 0$, let $h_M(x) = (x \wedge M) \vee (-M)$. Define the projection operator $\mathcal{P}_j(\cdot) = \mathbb{E}(\cdot|\varepsilon_j, \varepsilon_{j-1}, \ldots) - \mathbb{E}(\cdot|\varepsilon_{j-1}, \varepsilon_{j-2}, \ldots)$. Then we can write $X_i = \sum_{h=0}^{\infty} \mathcal{P}_{i-h} X_i$. By the orthogonality of $\mathcal{P}_j$, the triangle inequality and the Hölder inequality, we have

$$\begin{aligned} |\mathrm{Cov}(X_0, X_k)| &= \Big| \sum_{h=0}^{\infty} \mathbb{E}[(\mathcal{P}_{-h}X_0)(\mathcal{P}_{-h}X_k)] \Big| \leq \sum_{h=0}^{\infty} \Big| \mathbb{E}[(\mathcal{P}_{-h}X_0)(\mathcal{P}_{-h}X_k)] \Big| \\ &\leq \sum_{h=0}^{\infty} \|\mathcal{P}_{-h}X_0\|_2 \|\mathcal{P}_{-h}X_k\|_2 \leq \sum_{h=0}^{\infty} \delta_{h,2}\delta_{h+k,2}, \end{aligned}$$

where the last step follows by the fact that $\|\mathcal{P}_j X_i\|_2 \leq \delta_{i-j,2} = \|X_{i-j} - X_{i-j,\{0\}}\|_2$ in view of Jensen's inequality. It follows that

$$\sum_{k=-\infty}^{\infty} |\mathrm{Cov}(X_0, X_k)| \leq 2 \sum_{k=0}^{\infty}\sum_{h=0}^{\infty} \delta_{h,2}\delta_{h+k,2} \leq 2\|X_\cdot\|_{2,\mathrm{GMC}}^2. \tag{53}$$

Hence, by the Lipschitz continuity of the function $h_M(x)$ and the bound $|h_M(x)| \leq M$,

$$\begin{aligned} V &= \sup_{M>0} \sup_{i>0} \left(\mathrm{Var}(h_M(X_i)) + 2\sum_{j>i} |\mathrm{Cov}(h_M(X_i), h_M(X_j))| \right) \\ &\leq \sum_{k=-\infty}^{\infty} |\mathrm{Cov}(X_0, X_k)| \leq 2\|X_\cdot\|_{2,\mathrm{GMC}}^2. \end{aligned}$$

As $\mathbb{P}(|X_i| > x) \leq \exp(-cx^\nu)$ for some $\nu > 0$, applying Theorem 1 in [53] with $1/\gamma = 1 + 1/\nu$ therein, we have the desired results.

□

Lemma A.1. *Let D_i, $1 \leq i \leq n$, be p-dimensional martingale difference vectors with respect to the σ-field $\mathcal{G}_i$. Let $s > 1$ and $q \geq 2$. Then*

$$\||D_1 + \ldots + D_n|_s\|_q \leq c \left\{ q\|\sup_i |D_i|_s\|_q + \sqrt{q(s-1)} \left\| \left[\sum_{i=1}^{n} \mathbb{E}(|D_i|_s^2|\mathcal{G}_{i-1}) \right]^{1/2} \right\|_q \right\},$$

where c is an absolute constant.

Lemma A.1 provides a Rosenthal-Burkholder type bound on moments of Banach-spaced martingales and follows from Theorem 4.1 of [59].

Lemma A.2. *Assume $s > 1$. Let $X_1, \ldots, X_n$ be p-dimensional independent random vectors with mean zero such that for some $q > 2$, $\| |X_i|_s \|_q < \infty$, $1 \leq i \leq n$. Let $T_n = \sum_{i=1}^n X_i$ and $\sigma_i = (\|X_{i1}\|_2, \ldots, \|X_{ip}\|_2)^\top$. Then, for any $y > 0$,*

$$\mathbb{P}\left(|T_n|_s \geq 2\mathbb{E}|T_n|_s + y\right) \leq C_q y^{-q} \sum_{i=1}^n \mathbb{E}|X_i|_s^q + \exp\left(-\frac{y^2}{3\sum_{i=1}^n |\sigma_i|_s^2}\right), \tag{54}$$

where C_q is a positive constant depending only on q.

Proof. See Lemma C.6 in [89]. □

Appendix B: Quasi log-likelihood Loss

More generally than the commonly used maximum log-likelihood function, we may consider the following quasi-(log)likelihood function

$$R(z, y) := -\int_y^{H(z)} \frac{y-u}{\mathcal{V}(u)} du, \quad y \in \mathcal{Y}, \quad z \in \mathbb{R},$$

where $\mathcal{V} : \mathbb{R} \to (0, \infty)$ is a given variance function, and H is the inverse link function; see also [51]. The canonical link function (up to an additive constant) is

$$g(t) := \int_{y_0}^t \frac{1}{\mathcal{V}(u)} du, \quad t \in \mathcal{Y},$$

where y_0 is an arbitrary but fixed constant. Let

$$r(z) := \int_{y_0}^{H(z)} \frac{u}{\mathcal{V}(u)} du, \quad z \in \mathbb{R}.$$

Then the loss function is $R(z, y) = -yg(H(z)) + r(z)$. In this sense, we can define the loss function

$$R(z, y) = -yh(z) + r(z), \tag{55}$$

and assume h and r satisfy some uniform continuity conditions. It is worth noting that the main theorems in Section 3 and their proofs in Section D can be extended to quasi-loglikelihood loss function with minor modifications.

Appendix C: Real Data Analysis

In this section, we use a real dataset to illustrate the application of the Lasso procedures. Consider the high frequency financial dataset studied by [75], which consisting of the high-frequency trading of Walgreens stock on February 6, 2017. The data are available from the TAQ database of the New York Stock Exchange. Let y_i^* be the observed price change of the ith trade during the normal trading hours between 9:30AM to 4:00PM, Eastern Time. Due to the discreteness of y_i^*, as suggested by [75] Example 4.2, we divide the price changes into 7 categories, namely,

$$(-\infty, -0.02), \quad [-0.02, -0.01), \quad [-0.01, 0), \quad 0, \quad (0, 0.01], \quad (0.01, 0.02], \quad (0.02, \infty),$$

where the unit is one U.S. dollar. The category associated with y_i^* is thus defined as Y_i. If $y_i^* < -0.02$, we have $Y_i = 1$, if $-0.02 \leq y_i^* < -0.01$, $Y_i = 2$, and so on. We let t_i be the time duration between $(i-1)$th and ith transactions, which is measured in seconds. Let s_i be the normalized size of the transaction, which is the trading volume (number of shares) of the ith trade divided by 100. We also define six dummy variables for the price changes. Specifically, let

$$z_{i,j} = \begin{cases} 1 & \text{if } Y_i = j, \\ 0 & \text{if } Y_i \neq j, \end{cases} \qquad j = 2, ..., 7. \tag{56}$$

Denote $z_i = (z_{i,2}, ..., z_{i,7})'$. Then in our study, we employ the following $9d$ input variables,

$$X_i := \{z_{i-l}, y_{i-l}^*, t_{i-l}, s_{i-l} | l = 1, 2, ..., d\},$$

where d denotes the largest lag used in the time series. For this dataset, we want to predict trade-by-trade price change. On February 6, 2017, there were 29275 transactions available for the Walgreens stock. We use the first 27275 observations as the training subsample and reserve the last 2000 observations for out-of-sample prediction for comparison.

The well known ordered probit model [32] with $d = 3$ is used as benchmark. Setting $d = 3$, [75] compare the benchmark with several network models. In this particular instance, a 27-10-1 (feedforward) neural network appears to perform the best among the network models considered. The prediction results for both models are reported in Tables 5 and 6, respectively. In comparison, we apply Lasso methods with multinomial logistic regression to the data. Besides the main effects X_i, we also add two-way interactions between $y_{i-l}^*, t_{i-l}, s_{i-l}$ and z_{i-l}, $l = 1, 2, ..., d$. That is, there are a total of $27d$ input variables. Note that adding two-way interactions does not improve the predictions of the benchmark. In both standard and robust Lasso procedures, we choose $d = 16$. The optimal values of tuning parameters are chosen by a two-dimensional grid search using BIC; see also Section 2.2. The prediction results are summarized in Tables 7 and 8.

Table 5 shows that the ordered probit model does not perform well in prediction. As a matter of fact, the model predicts no price change for all of the last 2000 transactions. This is not surprising as the probability of no price change in the training subsample is 71.3%. The forecasting results in Table 6 show that the 27-10-1 neural network is able to correctly predict 3, 47, 1389, and 11 times for Categories 1, 3, 4, and 5, respectively. Its misclassification rate is 27.5%. In comparison, the standard Lasso procedure for multinomial logistic regression correctly predicts 2, 5, 35, 1378, 20, and 3 times for Categories 1, 2, 3, 4, 5, and 6, respectively. The corresponding misclassification rate is 27.85%. And the proposed robust Lasso procedure for multinomial logistic regression correctly predicts 3, 6, 32, 1378, 20, 3 and 1 times for Categories 1 to 7, respectively. Its misclassification rate is also 27.85%. The standard Lasso procedure and the robust Lasso procedure perform almost the same for the misclassification rate but the latter improves the predictions in most categories.

In some scenarios of multiclass classification, researchers assign different costs for classifying certain classes (see [19]); for example, it may be less costly to misclassify a benign tumor as cancerous than the opposite. In our particular example, the classes are unbalanced and we are more interested in big price changes. To this end, we use a cost matrix $W = (w_{jk})_{j,k=1}^{7} \in \mathbb{R}_{+}^{7\times 7}$, where $w_{jk} \geq 0$ is the cost (weights) for classifying an observation of class j as class k. We assume $w_{jj} = 0$ for each j. Then, the weighted empirical error is defined as

$$err_W(g) = \frac{1}{n}\sum_{i\leq n}\sum_{j,k=1}^{7} w_{jk}\mathbf{1}_{\{g(X_i)=j,Y_i=k\}}, \tag{57}$$

where g is a classifier. We further define $w_{jk}(k \neq j)$ as the reciprocal of the proportion of class j among all the 29275 observations, *i.e.,*

$$w_{jk} = \left(\frac{1}{29275}\sum_{i=1}^{29275}\mathbf{1}_{\{Y_i=j\}}\right)^{-1}, \quad k \neq j \tag{58}$$

Then, the weighted empirical errors for the 27-10-1 network, standard Lasso procedure and robust Lasso procedure are 4.09, 4.11 and 3.99, respectively. Therefore, both the standard Lasso procedure and robust Lasso procedure are compatible to the neural networks. From the weighted empirical error perspective, our robust Lasso procedure fares best. This example demonstrates that the standard Lasso procedure and robust Lasso procedure can be helpful in modeling trade-by-trade price changes in the financial market.

Table 5. Forecast tabulation for the ordered Probit model

		Predicted Categories						
	Y_i	1	2	3	4	5	6	7
Real Categories	1	0	0	0	4	0	0	0
	2	0	0	0	100	0	0	0
	3	0	0	0	180	0	0	0
	4	0	0	0	1437	0	0	0
	5	0	0	0	174	0	0	0
	6	0	0	0	100	0	0	0
	7	0	0	0	5	0	0	0

Table 6. Forecast tabulation for a 27-10-1 feedforward neural network

		Predicted Categories						
	Y_i	1	2	3	4	5	6	7
Real Categories	1	3	0	0	1	0	0	0
	2	0	0	2	98	0	0	0
	3	0	0	47	129	4	0	0
	4	2	0	38	1389	8	0	0
	5	0	0	11	152	11	0	0
	6	0	0	5	95	0	0	0
	7	0	0	0	5	0	0	0

Table 7. Forecast tabulation for the standard Lasso method

		Predicted Categories						
	Y_i	1	2	3	4	5	6	7
Real Categories	1	2	1	0	1	0	0	0
	2	0	5	2	91	1	1	0
	3	0	0	35	134	11	0	0
	4	1	7	23	1378	20	6	2
	5	0	1	6	146	20	1	0
	6	0	1	0	94	2	3	0
	7	1	0	0	4	0	0	0

Table 8. Forecast tabulation for the robust Lasso method

		Predicted Categories						
	Y_i	1	2	3	4	5	6	7
Real Categories	1	3	0	0	1	0	0	0
	2	0	6	2	90	1	1	0
	3	0	0	32	141	7	0	0
	4	0	7	24	1378	22	5	1
	5	0	1	5	147	20	1	0
	6	0	1	0	96	0	3	0
	7	1	0	0	3	0	0	1

Appendix D: Proofs of Main Theorems

Lemma D.1. *Suppose Assumption 3 holds and $\lambda \geq 2|\nabla\mathcal{R}_n(\beta^*)|_\infty$. Let $\mathcal{N} = \{\beta \in \mathbb{R}^p : |\beta - \beta^*|_2^2 \leq c_1 s\lambda^2, \beta - \beta^* \in \mathcal{C}(S)\}$, where $\mathcal{C}(S)$ is defined in (14) and $c_1 > 0$ is a constant. Suppose $\mathcal{R}_n(\beta)$ satisfies LRSC$(\mathcal{C}(S), \mathcal{N}, \kappa_\mathcal{R}, \varphi_\mathcal{R})$, where $\varphi_\mathcal{R} = c_0 s\lambda^2$ for some positive constant c_0 and $\kappa_\mathcal{R} > 0$ is a constant. Then,*

$$|\hat\beta - \beta^*|_2^2 \leq c_2 s(\lambda/\kappa_\mathcal{R})^2, \tag{59}$$

$$|\hat\beta - \beta^*|_1 \leq c_3 s(\lambda/\kappa_\mathcal{R}), \tag{60}$$

where $c_2, c_3 > 0$ are constants.

***Remark* D.1.** *Note that the results of Lemma D.1 can be extended to weakly sparsity case; see [20].*

Proof. Let $\hat\beta_t = t\hat\beta + (1-t)\beta^*$. We set $t = 1$ if $|\hat\beta - \beta^*|_2 \leq \ell$ and $t = \ell/|\hat\beta - \beta^*|_2$ if $|\hat\beta - \beta^*|_2 > \ell$. Denote $\hat\Delta = \hat\beta - \beta^*$ and $\hat\Delta_t = \hat\beta_t - \beta^*$. By [56] and Assumption 3, when $\lambda \geq 2|\nabla\mathcal{R}_n(\beta^*)|_\infty$, $\hat\Delta$ falls in the cone $\mathcal{C}(S)$.

Note that $\hat\Delta_t \in \mathcal{C}(S)$ as $t \leq 1$, and $|\hat\Delta_t|_1 = |\hat\beta_t - \beta^*|_1 \leq \ell$. By LRSC$(\mathcal{C}(S), \mathcal{N}, \kappa_\mathcal{R}, \varphi_\mathcal{R})$, the symmetric Bregman divergence satisfies

$$D_\mathcal{R}(\hat\beta_t, \beta^*) = (\hat\beta_t - \beta^*)^\top(\nabla\mathcal{R}_n(\hat\beta_t) - \nabla\mathcal{R}_n(\beta^*)) \geq \kappa_\mathcal{R}|\hat\Delta_t|_2^2 - \varphi_\mathcal{R}.$$

By Lemma F.2 in [22], $D_\mathcal{R}(\hat\beta_t, \beta^*) \leq tD_\mathcal{R}(\hat\beta, \beta^*)$. The Karush-Kuhn-Tucker (KKT) condition gives that $\nabla\mathcal{R}_n(\hat\beta) + \lambda\zeta = 0$ for some subgradient ζ of $|\beta|_1$ at $\beta = \hat\beta$. It follows that

$$\begin{aligned}\kappa_\mathcal{R}|\hat\Delta_t|_2^2 - \varphi_\mathcal{R} &\leq tD_\mathcal{R}(\hat\beta, \beta^*) = \hat\Delta_t^\top(\nabla\mathcal{R}_n(\hat\beta) - \nabla\mathcal{R}_n(\beta^*)) \leq \hat\Delta_t^\top(-\nabla\mathcal{R}_n(\beta^*) - \lambda\zeta) \\ &\leq 1.5\lambda|\hat\Delta_t|_1 \leq 6\lambda|(\hat\Delta_t)_S|_1 \leq 6\lambda\sqrt{s}|(\hat\Delta_t)_S|_2 \leq 6\lambda\sqrt{s}|\hat\Delta_t|_2.\end{aligned}$$

Hence, given $\varphi_\mathcal{R} = c_0' s\lambda^2/\kappa_\mathcal{R}$,

$$|\hat\Delta_t|_2 \leq C_1\sqrt{s}(\lambda/\kappa_\mathcal{R}).$$

If choose $\ell > C_1\sqrt{s}(\lambda/\kappa_\mathcal{R})$, then $\hat\Delta_t = \hat\Delta$. It follows that $|\hat\beta - \beta^*|_2^2 \leq C_1^2 s(\lambda/\kappa_\mathcal{R})^2$.

Moreover,over the cone $\mathcal{C}(S)$,

$$|\hat\beta - \beta^*|_1 \leq 4|(\hat\beta - \beta^*)_S|_1 \leq 4\sqrt{s}|\hat\beta - \beta^*|_2 \leq C_1 s(\lambda/\kappa_\mathcal{R}).$$

□

Lemma D.2. *Assume $|\beta^*|_1 \leq L < \infty$, $|r'(x)| \leq M_1 < \infty$ and $|r''(x)| \leq M_2 < \infty$ for any $x \in \mathbb{R}$. Also assume $\||X_\cdot|_\infty\|_{\gamma,\alpha_X} < \infty, \|Y_\cdot\|_{q,\alpha_Y} < \infty$, where $\gamma > 16/7, q > 2$, $\alpha_X > 21/2 - 8/\gamma$, $\alpha_Y > 1/2 - 1/\gamma - 1/q$. Let $1/\chi = 1/\gamma + 1/q < 1/2$, $\alpha_1 = \min\{\alpha_X, \alpha_Y\}$*

and $\alpha_2 = \alpha_X/7 - 1$*. It holds that, in an event with probability at most* $C_1(\log p)^{-\chi} + C_2(\log p)^{-7\gamma/8} + p^{-C_3}$

$$\left|\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - r'(X_i^\top\beta^*)\big)X_i\right|_\infty > C_4 a_{p,1}\sqrt{\frac{\log p}{n}} + \frac{C_5 a_{p,2}(\log p)^{3/2}}{n^{1-1/q-1/\gamma}} + \frac{C_6 a_{p,3}(\log p)^{3/2}}{n^{1-8/(7\gamma)}}, \tag{61}$$

where $C_1, C_2, C_3, C_4, C_5, C_6 > 0$ *are constants,* C_3 *and* C_5 *only depend on* L*, and*

$$\begin{aligned}
a_{p,1} &= \max_j \|X_{\cdot j}\|_{\gamma,\alpha_1}\|Y_\cdot\|_{q,\alpha_1} + \||X_\cdot|_\infty\|_{\gamma,\alpha_X}^{1/7} \max_j \|X_{\cdot j}\|_{\gamma,\alpha_X} + \max_j \|X_{\cdot j}\|_{\gamma,\alpha_2},\\
a_{p,2} &= \||X_\cdot|_\infty\|_{\gamma,\alpha_1}\|Y_\cdot\|_{q,\alpha_1},\\
a_{p,3} &= \||X_\cdot|_\infty\|_{\gamma,\alpha_X}^{1/7}\||X_\cdot|_\infty\|_{\gamma,\alpha_X} + \||X_\cdot|_\infty\|_{\gamma,\alpha_2}.
\end{aligned}$$

Proof. As $\mathbb{E}(Y_i|X_i) = r'(X_i^\top\beta^*)$,

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - r'(X_i^\top\beta^*)\big)X_i\right|_\infty &\le \left|\frac{1}{n}\sum_{i=1}^{n} Y_iX_i - \mathbb{E}Y_iX_i\right|_\infty + \left|\frac{1}{n}\sum_{i=1}^{n} r'(X_i^\top\beta^*)X_i - \mathbb{E}r'(X_i^\top\beta^*)X_i\right|_\infty\\
&:= \mathrm{I} + \mathrm{II}.
\end{aligned}$$

We bound the above two terms one by one.

We first consider I. Let $1/\chi = 1/\gamma + 1/q$. By Hölder's inequality, we have for $m \ge 0$ that

$$\begin{aligned}
&\sum_{l=m}^{\infty} \|\max_j |X_{lj}Y_l - X_{lj,\{0\}}Y_{l,\{0\}}|\|_\chi\\
&\le \sum_{l=m}^{\infty}\left(\|\max_j |X_{lj}(Y_l - Y_{l,\{0\}})|\|_\chi + \|\max_j |(X_{lj} - X_{lj,\{0\}})Y_{l,\{0\}}|\|_\chi\right)\\
&= \sum_{l=m}^{\infty}\left(\|\max_j |X_{lj}|\|_\gamma \|Y_l - Y_{l,\{0\}}\|_q + \|\max_j |X_{lj} - X_{lj,\{0\}}|\|_\gamma \|Y_{l,\{0\}}\|_q\right).
\end{aligned}$$

Since $\alpha_1 = \min(\alpha_X, \alpha_Y)$, the dependence adjusted norm satisfies for χ

$$\|\max_j |X_{\cdot j}Y_\cdot|\|_{\chi,\alpha_1} \le \|\max_j |X_{\cdot j}|\|_{\gamma,0}\|Y_\cdot\|_{q,\alpha_1} + \|\max_j |X_{\cdot j}|\|_{\gamma,\alpha_1}\|Y_\cdot\|_{q,0} \le 2\||X_\cdot|_\infty\|_{\gamma,\alpha_1}\|Y_\cdot\|_{q,\alpha_1}.$$

Similarly, we can obtain

$$\|X_{\cdot j}Y_\cdot\|_{\chi,\alpha_1} \le \|X_{\cdot j}\|_{\gamma,0}\|Y_\cdot\|_{q,\alpha_1} + \|X_{\cdot j}\|_{\gamma,\alpha_1}\|Y_\cdot\|_{q,0} \le 2\|X_\cdot\|_{\gamma,\alpha_1}\|Y_\cdot\|_{q,\alpha_1}.$$

Let $\alpha_X > 21/2 - 8/\gamma$ and $\alpha_Y > 1/2 - 1/\gamma - 1/q$. Then $\alpha_1 = \min\{\alpha_X, \alpha_Y\} > 1/2 - 1/\chi$. Applying Theorem A.1, we have

$$\mathrm{I} \gtrsim \sqrt{\frac{\log p}{n}} \max_j \|X_{\cdot j}\|_{\gamma,\alpha_1}\|Y_\cdot\|_{q,\alpha_1} + n^{1/q+1/\gamma-1}(\log p)^{3/2}\||X_\cdot|_\infty\|_{\gamma,\alpha_1}\|Y_\cdot\|_{q,\alpha_1}, \tag{62}$$

in an event with probability at least $1 - C_1(\log p)^{-\chi} - p^{-C_2}$.

Next, we consider II. Let $1/\phi = 1/\gamma + 1/K$, where $K \geq \gamma$ will be specified later. Again, by Hölder's inequality, we have for $m \geq 0$ that,

$$\begin{aligned}
&\sum_{i=m}^{\infty} \| \max_j |r'(X_i^\top \beta^*) X_{ij} - r'(X_{i,\{0\}}^\top \beta^*) X_{ij,\{0\}}| \|_\phi \\
&\leq \sum_{i=m}^{\infty} \left(\| \max_j |[r'(X_i^\top \beta^*) - r'(X_{i,\{0\}}^\top \beta^*)] X_{ij}| \|_\phi + \| \max_j |r'(X_{i,\{0\}}^\top \beta^*) [X_{ij} - X_{ij,\{0\}}]| \|_\phi \right) \\
&\leq \sum_{i=m}^{\infty} M_1 \| \max_j |X_i - X_{i,\{0\}}| \|_\phi + \sum_{i=m}^{\infty} \| r'(X_i^\top \beta^*) - r'(X_{i,\{0\}}^\top \beta^*) \|_K \| \max_j |X_{ij}| \|_\gamma .
\end{aligned}$$

Let $R_0 = L/(2M_1)$. As $\sum_{i=m}^{\infty} \| |X_i - X_{i,\{0\}}|_\infty \|_\gamma = O((m+1)^{-\alpha_X} \| |X_\cdot|_\infty \|_{\gamma,\alpha_X})$,

$$\begin{aligned}
\sum_{i=m}^{\infty} \| r'(X_i^\top \beta^*) - r'(X_{i,\{0\}}^\top \beta^*) \|_K &\leq \sum_{i=m}^{\infty} \| \min\{2M_1, L|X_i - X_{i,\{0\}}|_\infty\} \|_K \\
&\leq 2M_1 \sum_{i=m}^{\infty} \left(\mathbb{E} \min\{1, R_0 |X_i - X_{i,\{0\}}|_\infty\}^K \right)^{1/K} \\
&\leq 2M_1 \sum_{i=m}^{\infty} \left(\mathbb{E} \min\{1, R_0 |X_i - X_{i,\{0\}}|_\infty\}^\gamma \right)^{1/K} \\
&\leq 2M_1 \sum_{i=m}^{\infty} R_0^{\gamma/K} \frac{\| |X_\cdot|_\infty \|_{\gamma,\alpha_X}^{\gamma/K}}{(l+1)^{\alpha_X \gamma/K}} .
\end{aligned}$$

Thus, we need $\alpha_X \gamma / K > 1$. By setting $K = 7\gamma$, we have $1/\phi = 8/(7\gamma) < 1/2$ and $\alpha_2 = \alpha_X \gamma/K - 1 = \alpha_X/7 - 1 > 1/2 - 8/(7\gamma) = 1/2 - 1/\phi > 0$. Meanwhile,

$$\sum_{i=m}^{\infty} \| r'(X_i^\top \beta^*) - r'(X_{i,\{0\}}^\top \beta^*) \|_K \leq 2M_1 R_0^{1/7} \| |X_\cdot|_\infty \|_{\gamma,\alpha_X}^{1/7} .$$

It follows that

$$\begin{aligned}
\| \max_j |r'(X_\cdot^\top \beta^*) X_{\cdot j}| \|_{\phi,\alpha_2} &= \sup_{m \geq 0} (m+1)^{\alpha_2} \sum_{i=m}^{\infty} \| \max_j |r'(X_i^\top \beta^*) X_{ij} - r'(X_{i,\{0\}}^\top \beta^*) X_{ij,\{0\}}| \|_\phi \\
&\leq M_1 \| \max_j |X_{\cdot j}| \|_{\phi,\alpha_2} + 2M_1 R_0^{1/7} \| |X_\cdot|_\infty \|_{\gamma,\alpha_X}^{1/7} \| \max_j |X_{ij}| \|_\gamma \\
&\leq M_1 \| \max_j |X_{\cdot j}| \|_{\phi,\alpha_2} + (2M_1)^{6/7} L^{1/7} \| |X_\cdot|_\infty \|_{\gamma,\alpha_X}^{1/7} \| \max_j |X_{ij}| \|_\gamma .
\end{aligned}$$

As $\alpha_2 > 1/2 - 1/\phi$, Theorem A.1 implies that

$$\begin{aligned}
\mathrm{II} \gtrsim & \sqrt{\frac{\log p}{n}} \left(L^{1/7} \| |X_\cdot|_\infty \|_{\gamma,\alpha_X}^{1/7} \max_j \| X_{\cdot j} \|_{\gamma,\alpha_X} + \max_j \| X_{\cdot j} \|_{\gamma,\alpha_2} \right) \\
& + n^{8/(7\gamma)-1} (\log p)^{3/2} \left(L^{1/7} \| |X_\cdot|_\infty \|_{\gamma,\alpha_X}^{1/7} \| |X_\cdot|_\infty \|_{\gamma,\alpha_X} + \| |X_\cdot|_\infty \|_{\gamma,\alpha_2} \right), \qquad (63)
\end{aligned}$$

in an event with probability at least $1 - C_3(\log p)^{-7\gamma/8} - p^{-C_4}$.

Then Lemma D.2 follows from (62) and (63). □

Proof of Lemma 3.1. We first prove the RSC of $\mathcal{R}_n(\beta)$ at $\beta = \beta^*$ over the cone $\mathcal{C}(S)$

$$\mathcal{C}(S) = \{\Delta \in \mathbb{R}^p : |\Delta_{S^c}|_1 \leq 3|\Delta_S|_1\},$$

for $S \subset \{1, 2, ..., p\}$ and $|S|_0 = s$. Let $\gamma' = \gamma/2$, $K = 7\gamma' = 7\gamma/2$, $1/\phi' = 1/\gamma' + 1/K$ and $\alpha_2 = \alpha_X \gamma'/K - 1 = \alpha_X/7 - 1 > 1/2 - 8/(7\gamma) > 1/2 - 1/\phi' > 0$. Similarly to the proof of Lemma D.2, we can show that in an event with probability at least $1 - c_1(\log p)^{-7/(16\gamma)} - p^{-c_2}$,

$$\left|\frac{1}{n}\sum_{i=1}^n r''(X_i^\top \beta^*) X_i X_i^\top - \mathbb{E} r''(X_i^\top \beta^*) X_i X_i^\top\right|_\infty \leq c_3 a_{p,4}\sqrt{\frac{\log p}{n}} + c_4 a_{p,5} n^{16/(7\gamma)-1}(\log p)^{3/2},$$

where $c_1, c_2, c_3, c_4 > 0$ and c_3, c_4 only depend on L.

Next, we bound the difference of $n^{-1}\sum_{i=1}^n \Delta^\top r''(X_i^\top \beta) X_i X_i^\top \Delta$ and $n^{-1}\sum_{i=1}^n \Delta^\top r''(X_i^\top \beta^*) X_i X_i^\top \Delta$ to control LRSC within a neighborhood of β^*. Basic calculation shows that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \Delta^\top r''(X_i^\top \beta) X_i X_i^\top \Delta - \frac{1}{n}\sum_{i=1}^n \Delta^\top r''(X_i^\top \beta^*) X_i X_i^\top \Delta\right| \\
=&\left|\frac{1}{n}\sum_{i=1}^n [r''(X_i^\top \beta) - r''(X_i^\top \beta^*)] \Delta^\top X_i X_i^\top \Delta\right| \\
\leq& \max_i \left|r''(X_i^\top \beta) - r''(X_i^\top \beta^*)\right| \cdot \frac{1}{n}\sum_{i=1}^n (X_i^\top \Delta)^2. \qquad (64)
\end{aligned}$$

Let

$$\lambda_0 = c_3 a_{p,4}\sqrt{\frac{\log p}{n}} + c_4 a_{p,5} n^{16/(7\gamma)-1}(\log p)^{3/2}.$$

As $\sup_{|\nu|_2=1} \mathbb{E}|X_i^\top \nu|^\gamma \leq c_0$ and $\Delta \in \mathcal{C}(S)$, we can show that in an event with probability at least $1 - c_1(\log p)^{-7/(16\gamma)} - p^{-c_2}$,

$$\frac{1}{n}\sum_{i=1}^n (X_i^\top \Delta)^2 \leq c_0|\Delta|_2^2 + c'\lambda_0 \cdot |\Delta|_1^2 \leq c_0|\Delta|_2^2 + c' s\lambda_0 \cdot |\Delta|_2^2 \leq 2c_0|\Delta|_2^2.$$

By Markov inequality, setting $x \asymp n^{2/\gamma}$, for any $|\nu|_2 = 1$

$$\mathbb{P}(\max_i |X_i^\top \nu| > x) \leq n \cdot \frac{\mathbb{E}|X_i^\top \nu|^\gamma}{x^\gamma} \leq c_0 n^{-1}. \qquad (65)$$

As $|\beta - \beta^*|_2^2 \leq C_1 s\lambda^2$, over the cone $\mathcal{C}(S)$,

$$\max_i \left|r''(X_i^\top \beta) - r''(X_i^\top \beta^*)\right| \leq M_3 c' |\beta - \beta^*|_2 n^{2/\gamma} \leq M_3 c' n^{2/\gamma}\sqrt{s}\lambda.$$

Hence, we have, for any $\beta \in \mathcal{N}$,

$$\frac{1}{n}\sum_{i=1}^{n}\Delta^\top r''(X_i^\top\beta)X_iX_i^\top\Delta \geq \kappa_{\mathrm{H}}|\Delta|_2^2 - \lambda_0|\Delta|_1^2 - 2c_0'|\Delta|_2^2 \geq (\kappa_{\mathrm{H}}/2)|\Delta|_2^2. \tag{66}$$

□

Proof of Theorem 3.1. Theorem 3.1 follows from Lemma D.2, Lemma 3.1 and Lemma D.1. □

Proof of Proposition 3.1. The proof of Proposition 3.1 is similar to that of Theorem 3.1, and thus is omitted. Specifically, we need to replace the concentration inequality Theorem A.1 by Fuk-Nagaev inequality in Lemma D.2 of [17]. As we do not need to use the condition $|\beta^*|_1 \leq L$ to derive functional dependence measure, the proof can be simplified. □

Lemma D.3. *Assume $|\beta^*|_1 \leq L < \infty$, $|r'(x)| \leq M_1 < \infty$ and $|r''(x)| \leq M_2 < \infty$ for any $x \in \mathbb{R}$. Also assume $\max_{1\leq j\leq p}\|X_{\cdot j}\|_{\psi_\iota} < \infty$, $\|Y_\cdot\|_{\psi_\nu} < \infty$, where $\iota, \nu > 0$.*

(i). It holds that, in an event with probability at most p^{-C_1}

$$\left|\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - r'(X_i^\top\beta^*)\big)X_i\right|_\infty > C_2\max_j\|X_{\cdot j}\|_{\psi_\iota}\|Y_\cdot\|_{\psi_\nu}\frac{(\log p)^{1/2+\nu+\iota}}{\sqrt{n}} + C_3\max_j\|X_{\cdot j}\|_{\psi_\iota}^2\frac{(\log p)^{1/2+2\iota}}{\sqrt{n}}, \tag{67}$$

where $C_1, C_2 > 0$ are constants, and $C_3 > 0$ only depends on L.

(ii). Assume the process X_i and Y_i also satisfy geometric moment contraction $\|X_{\cdot j}\|_{4,\mathrm{GMC}} < \infty$ for some constant $0 < \rho_j < 1$, $\|Y_\cdot\|_{4,\mathrm{GMC}} < \infty$ for some constant $0 < \rho_y < 1$, and $\rho = \min\{\rho_j, \rho_y\} \in (0,1)$. Then, in an event with probability at most p^{-C_4},

$$\left|\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - r'(X_i^\top\beta^*)\big)X_i\right|_\infty > C_5 b_{p,1}\sqrt{\frac{\log p}{n}} + \frac{C_6(\log p + \log n)^{1+\iota+\nu}}{n} + \frac{C_7(\log p + \log n)^{1+\iota}}{n}, \tag{68}$$

where $C_4, C_5, C_6, C_7 > 0$ are constants, and

$$b_{p,1} = \max_j\|X_{\cdot j}\|_{4,\mathrm{GMC}}\|Y_\cdot\|_{4,\mathrm{GMC}} + L\max_j\|X_{\cdot j}\|_{4,\mathrm{GMC}}^2.$$

Proof. As $\mathbb{E}(Y_i|X_i) = r'(X_i^\top\beta^*)$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - r'(X_i^\top\beta^*)\big)X_i\right|_\infty \leq \left|\frac{1}{n}\sum_{i=1}^{n}Y_iX_i - \mathbb{E}Y_iX_i\right|_\infty + \left|\frac{1}{n}\sum_{i=1}^{n}r'(X_i^\top\beta^*)X_i - \mathbb{E}r'(X_i^\top\beta^*)X_i\right|_\infty$$
$$:= \mathrm{I} + \mathrm{II}.$$

We first consider the case (i). Let $\gamma = \phi(1+\iota/\nu)$ and $q = \phi(1+\nu/\iota)$. Then

$$\begin{aligned}\sum_{l=0}^{\infty} \|X_{lj}Y_l - X_{lj,\{0\}}Y_{l,\{0\}}\|_\phi &\le \sum_{l=0}^{\infty} \big(\|X_{lj}\|_\gamma \|Y_l - Y_{l,\{0\}}\|_q + \||X_{lj} - X_{lj,\{0\}}|\|_\gamma \|Y_{l,\{0\}}\|_q\big) \\ &\le 2\Delta_{0,\gamma,j}\Delta_{0,q,y}.\end{aligned}$$

By the definition of γ and q, we have

$$||X_{\cdot j}Y_{\cdot}\|_{\psi_{\nu+\iota}} \le \sup_{\phi\ge 2} \frac{2\Delta_{0,\gamma,j}\Delta_{0,q,y}}{\phi^{\iota+\nu}} \le 2\|X_{\cdot j}\|_{\psi_\iota}\|Y_{\cdot}\|_{\psi_\nu} \sup_{\phi\ge 2}\frac{\gamma^\iota q^\nu}{\phi^{\iota+\nu}} = 2c\|X_{\cdot j}\|_{\psi_\iota}\|Y_{\cdot}\|_{\psi_\nu},$$

where $c = (1+\iota/\nu)^\iota(1+\nu/\iota)^\nu$. Then, applying Theorem A.2(ii) with $\alpha = 2/(1+2\nu+2\iota)$, we have, in an event with probability at least $1-p^{-c_1}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} Y_iX_i - \mathbb{E}Y_iX_i\right|_\infty \lesssim \frac{(\log p)^{1/2+\nu+\iota}}{\sqrt{n}} \max_j \|X_{\cdot j}\|_{\psi_\iota}\|Y_{\cdot}\|_{\psi_\nu}. \tag{69}$$

For II,

$$\begin{aligned}&\sum_{i=0}^{\infty} \|r'(X_i^\top\beta^*)X_{ij} - r'(X_{i,\{0\}}^\top\beta^*)X_{ij,\{0\}}\|_\phi \\ \le &\sum_{i=0}^{\infty} \|r'(X_i^\top\beta^*)[X_{ij} - X_{ij,\{0\}}]\|_\phi + \sum_{i=0}^{\infty} \|[r'(X_i^\top\beta^*) - r'(X_{i,\{0\}}^\top\beta^*)]X_{ij,\{0\}}\|_\phi \\ \le &M_1\Delta_{0,\phi,j} + \sum_{i=0}^{\infty} \|r'(X_i^\top\beta^*) - r'(X_{i,\{0\}}^\top\beta^*)\|_{2\phi}\|X_{ij,\{0\}}\|_{2\phi} \\ \le &M_1\Delta_{0,\phi,j} + M_2L\max_k \sum_{i=0}^{\infty} \|X_{ik} - X_{ik,\{0\}}\|_{2\phi}\|X_{ij}\|_{2\phi} \\ \le &M_1\Delta_{0,\phi,j} + M_2L\max_k \Delta_{0,2\phi,k}\Delta_{0,2\phi,j}.\end{aligned}$$

It follows that

$$\begin{aligned}||r'(X_{\cdot}^\top\beta^*)X_{\cdot j}\|_{\psi_{2\nu}} &\le \sup_{\phi\ge 2}\frac{M_1\Delta_{0,\phi,j} + M_2L\max_k \Delta_{0,2\phi,k}\Delta_{0,2\phi,j}}{\phi^{2\iota}} \\ &\le M_1\|X_{\cdot j}\|_{\psi_\iota} + M_2L\|X_{\cdot j}\|_{\psi_\iota}\max_k \|X_{\cdot k}\|_{\psi_\iota}.\end{aligned}$$

Again, Theorem A.2(ii) implies that, in an event with probability at least $1-p^{-c_2}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} r'(X_i^\top\beta^*)X_i - \mathbb{E}r'(X_i^\top\beta^*)X_i\right|_\infty \lesssim \frac{(\log p)^{1/2+2\iota}}{\sqrt{n}} \max_j \|X_{\cdot j}\|_{\psi_\iota}^2. \tag{70}$$

Part (i) is completed by (69) and (70).

Next, we shall prove (ii). Let $1/\phi_0 = \iota + \nu$. Elementary calculation shows that

$$\begin{aligned}\mathbb{P}\left(|X_{ij}Y_i| > x\right) &\le \mathbb{P}\left(|X_{ij}| > x^{\phi_0\iota}\right) + \mathbb{P}\left(|Y_i| > x^{\phi_0\nu}\right) \le c_3 \exp\left(-c_4 x^{\phi_0}\right) + c_5 \exp\left(-c_6 x^{\phi_0}\right) \\ &\le c_7 \exp\left(-c_8 x^{\phi_0}\right).\end{aligned}$$

By Hölder's inequality, for $\rho = \min\{\rho_j, \rho_y\}$,

$$\begin{aligned}&\rho^{-m}\sum_{i=m}^{\infty}\|Y_iX_{ij} - Y_{i,\{0\}}X_{ij,\{0\}}\|_2 \\ \le&\rho^{-m}\sum_{i=m}^{\infty}\|Y_i\|_4\|X_{ij} - X_{ij,\{0\}}\|_4 + \rho^{-m}\sum_{i=m}^{\infty}\|Y_i - Y_{i,\{0\}}\|_4\|X_{ij,\{0\}}\|_4 \\ \le&\|X_{.j}\|_{4,\mathrm{GMC}}\|Y_i\|_4 + \|X_{ij}\|_4\|Y_.\|_{4,\mathrm{GMC}}.\end{aligned}$$

It follows that

$$\|Y_.X_{.j}\|_{2,\mathrm{GMC}} \le 2\|X_{.j}\|_{4,\mathrm{GMC}}\|Y_.\|_{4,\mathrm{GMC}}.$$

Let $1/\phi_1 = 1 + 1/\phi_0 = 1 + \iota + \nu$. Applying Theorem A.4 with $x \asymp \sqrt{n\log p}\max_j\|Y_.X_{.j}\|_{2,\mathrm{GMC}} + (\log p + \log n)^{1/\phi_1}$, we have, in an event with probability at most p^{-c_9},

$$\max_j\left|\frac{1}{n}\sum_{i=1}^{n}X_{ij}Y_i - \mathbb{E}X_{ij}Y_i\right| \ge c_{10}\max_j\|X_{.j}\|_{4,\mathrm{GMC}}\|Y_.\|_{4,\mathrm{GMC}}\sqrt{\frac{\log p}{n}} + \frac{c_{11}(\log p + \log n)^{1+\iota+\nu}}{n}.$$

Similarly, we can show

$$\|r'(X_.^\top\beta^*)X_{.j}\|_{2,\mathrm{GMC}} \le M_2L\max_k\|X_{.k}\|_{4,\mathrm{GMC}}\|X_{.j}\|_{4,\mathrm{GMC}} + M_1\|X_{.j}\|_{2,\mathrm{GMC}} < \infty.$$

As $|r'(X_.^\top\beta^*)| \le M_1$, Theorem A.4 implies that, in an event with probability at most $p^{-c_{12}}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}r'(X_i^\top\beta^*)X_i - \mathbb{E}r'(X_i^\top\beta^*)X_i\right|_\infty \ge c_{13}L\max_j\|X_{.j}\|_{4,\mathrm{GMC}}^2\sqrt{\frac{\log p}{n}} + \frac{c_{14}(\log p + \log n)^{1+\iota}}{n}.$$

Then, part (ii) is also proved.

□

Proof of Lemma 3.2. (i). Similarly to the proof of part (i) in Lemma D.3, we can show that in an event with probability at least $1 - p^{-c_1}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}r''(X_i^\top\beta^*)X_iX_i^\top - \mathbb{E}r''(X_i^\top\beta^*)X_iX_i^\top\right|_\infty \le c_2L\max_j\|X_{.j}\|_{\psi_\iota}^3\frac{(\log p)^{1/2+3\iota}}{\sqrt{n}},$$

where $c_1, c_2 > 0$. In addition, as $\sup_{\gamma\geq 1}\sup_{|\nu|_2=1}\gamma^{-\iota}\|X_i^\top \nu\|_\gamma \leq c_0$, for any $|\theta|_2 = 1$,

$$\mathbb{P}(\max_i |X_i^\top \theta| \geq cx) \leq c' n \exp\left(-c' x^{1/\iota}\right).$$

Setting $x \asymp (\log n)^\iota$, we have

$$\mathbb{P}(\max_i |X_i^\top \theta| \geq c(\log n)^\iota) \leq n^{-c_3}.$$

As $|\beta - \beta^*|_2^2 \leq C_1 s\lambda^2$, over the cone $\mathcal{C}(S)$, in an event with probability at least $1 - n^{-c_3}$

$$\max_i \left|r''(X_i^\top \beta) - r''(X_i^\top \beta^*)\right| \leq M_3 c(\log n)^\iota |\beta - \beta^*|_2 \leq M_3 c(\log n)^\iota \sqrt{s}\lambda.$$

Then, employing the same arguments in the proof of Lemma 3.1, we can establish LRSC.

(ii). Similarly to the proof of part (ii) in Lemma D.3, we can show that in an event with probability at least $1 - p^{-c_1}$,

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n r''(X_i^\top \beta^*) X_i X_i^\top - \mathbb{E}r''(X_i^\top \beta^*) X_i X_i^\top\right|_\infty \\ \leq & c_2 L \max_j \|X_{\cdot j}\|_{6,\mathrm{GMC}}^3 \sqrt{\frac{\log p}{n}} + \frac{c_3(\log p + \log n)^{1+2\iota}}{n},\end{aligned}$$

where $c_1, c_2, c_3 > 0$. Then, adopting the same procedures in the proof of Lemma 3.1, we can establish LRSC.

□

Proof of Theorem 3.2. Theorem 3.2 follows from Lemma D.3, Lemma 3.2 and Lemma D.1. □

Proof of Proposition 3.2. The proof of Proposition 3.2 is also similar to that of Theorem 3.2(ii), and thus is omitted. Specifically, we need to replace Theorem A.2 by the concentration inequality in (1.4) of [53], which concerns exponential inequality for independent random variables. As we do not need to use the condition $|\beta^*|_1 \leq L$ to derive functional dependence measure, the proof can be much simplified. □

Proof of Corollary 3.1. We only prove the case under the conditions of Theorem 3.1, and proofs of the other two cases are similar. The specific technique was developed by [46]. The following proof is similar to that of Corollary 3 in [46]. Let $\hat{Q} := \int_0^1 \nabla^2 \mathcal{R}_n(\beta^* + t(\hat{\beta} - \beta^*))\mathrm{d}t$, $\mathcal{R}(\beta) = \mathbb{E}\mathcal{R}_n(\beta)$ and $\nabla^2\mathcal{R}(\beta) = \mathbb{E}H_n(\beta)$. By the fundamental theorem of calculus for vector-valued functions, $\hat{Q}\ (\hat{\beta} - \beta^*) = \nabla\mathcal{R}_n(\hat{\beta}) - \nabla\mathcal{R}_n(\beta^*)$.

By Theorem 3.1, in an event Ω with probability at most $1 - C_1(\log p)^{-\chi} - C_2(\log p)^{-7\gamma/16} - n^{-1} - p^{-C_3}$,

$$|\hat{\beta} - \beta^*|_2 \leq \sqrt{s}\lambda.$$

Note that

$$\hat{Q} - \nabla^2\mathcal{R}_n(\beta^*) = \int_0^1 \frac{1}{n}\sum_{i=1}^n \left(r''\left(X_i^\top \left(\beta^* + t(\hat{\beta} - \beta^*) \right) \right) - r''(X_i^\top \beta^*) \right) X_i X_i^\top \mathrm{d}t.$$

For each pair $v, w \in \mathbb{R}^p$, we have

$$\left| v^\top \left(\hat{Q} - \nabla^2\mathcal{R}_n(\beta^*) \right) w \right| \leq \frac{M_3}{2} \left| \frac{1}{n}\sum_{i=1}^n \left(X_i^\top(\hat{\beta} - \beta^*) \right) (X_i^\top v)(X_i^\top w) \right|.$$

Employing similar arguments in the proof of Lemma 3.1, we can show, in the event Ω,

$$\begin{aligned} \left\| \hat{Q}_{SS} - \nabla^2\mathcal{R}_n(\beta^*)_{SS} \right\|_2 &\lesssim n^{2/\gamma}\sqrt{s}\lambda, \\ \left\| \nabla^2\mathcal{R}_n(\beta^*)_{SS} - \nabla^2\mathcal{R}(\beta^*)_{SS} \right\|_2 &\lesssim s\lambda. \end{aligned}$$

If follows that

$$\left\| \hat{Q}_{SS} - \nabla^2\mathcal{R}(\beta^*)_{SS} \right\|_2 \lesssim n^{2/\gamma}\sqrt{s}\lambda + s\lambda.$$

By the matrix inequality that $\|A^{-1} - B^{-1}\|_2 \leq \|A^{-1}\|_2^2 \|A - B\|_2 / [1 - \|A^{-1}\|_2 \|A - B\|_2]$, we have

$$\left\| \left(\hat{Q}_{SS} \right)^{-1} - \left(\nabla^2\mathcal{R}(\beta^*)_{SS} \right)^{-1} \right\|_2 \lesssim n^{2/\gamma}\sqrt{s}\lambda + s\lambda. \tag{71}$$

A similar argument shows that in the event Ω,

$$\begin{aligned} \max_{j \in S^c} \left| e_j^\top \left(\hat{Q}_{S^cS} - \nabla^2\mathcal{R}_n(\beta^*)_{S^cS} \right) \right|_2 &\lesssim n^{2/\gamma}\sqrt{s}\lambda, \\ \max_{j \in S^c} \left| e_j^\top \left(\nabla^2\mathcal{R}_n(\beta^*)_{S^cS} - \nabla^2\mathcal{R}(\beta^*)_{S^cS} \right) \right|_2 &\lesssim \sqrt{s}\lambda. \end{aligned}$$

Then, in the event Ω,

$$\max_{j \in S^c} \left| e_j^\top \left(\hat{Q}_{S^cS} - \nabla^2\mathcal{R}(\beta^*)_{S^cS} \right) \right|_2 \lesssim n^{2/\gamma}\sqrt{s}\lambda. \tag{72}$$

Let

$$\left| \hat{Q}_{S^cS} \left(\hat{Q}_{SS} \right)^{-1} \nabla\mathcal{R}_n(\beta^*)_S \right|_\infty \leq \mathrm{I} + \mathrm{II},$$

where

$$\begin{aligned} \mathrm{I} &= \left| \nabla^2\mathcal{R}(\beta^*)_{S^cS} \left(\nabla^2\mathcal{R}(\beta^*)_{SS} \right)^{-1} \nabla\mathcal{R}_n(\beta^*)_S \right|_\infty, \\ \mathrm{II} &= \left| \left\{ \hat{Q}_{S^cS} \left(\hat{Q}_{SS} \right)^{-1} - \nabla^2\mathcal{R}(\beta^*)_{S^cS} \left(\nabla^2\mathcal{R}(\beta^*)_{SS} \right)^{-1} \right\} \nabla\mathcal{R}_n(\beta^*)_S \right|_\infty. \end{aligned}$$

By Lemma D.2, in the event Ω,

$$|\nabla R_n(\beta^*)|_\infty = \left|\frac{1}{n}\sum_{i=1}^n \left(Y_i - r'(X_i^\top \beta^*)\right) X_i\right|_\infty \leq \frac{1}{2}\lambda. \tag{73}$$

By the incoherence condition, it follows that I $\lesssim \lambda$. Turning to II, we have,

$$\text{II} \leq \max_{j\in S^c}\left|e_j^\top \left\{\hat{Q}_{S^cS}\left(\hat{Q}_{SS}\right)^{-1} - \nabla^2\mathcal{R}(\beta^*)_{S^cS}\left(\nabla^2\mathcal{R}(\beta^*)_{SS}\right)^{-1}\right\}\right|_2 |\nabla\mathcal{R}_n(\beta^*)_S|_2 .$$

Elementary calculation shows that

$$\begin{aligned}&\max_{j\in S^c}\left|e_j^\top \left\{\hat{Q}_{S^cS}\left(\hat{Q}_{SS}\right)^{-1} - \nabla^2\mathcal{R}(\beta^*)_{S^cS}\left(\nabla^2\mathcal{R}(\beta^*)_{SS}\right)^{-1}\right\}\right|_2 \\ &\leq \left|e_j^\top \nabla^2\mathcal{R}(\beta^*)_{S^cS}\Delta_1\right|_2 + \left|e_j^\top \Delta_2\left(\nabla^2\mathcal{R}(\beta^*)_{SS}\right)^{-1}\right|_2 + \left|e_j^\top\Delta_2\Delta_1\right|_2 ,\end{aligned}$$

where

$$\Delta_1 := \left(\hat{Q}_{SS}\right)^{-1} - \left(\nabla^2\mathcal{R}(\beta^*)_{SS}\right)^{-1}, \quad \text{and} \quad \Delta_2 := \left(\hat{Q}_{S^cS} - \nabla^2\mathcal{R}(\beta^*)_{S^cS}\right).$$

By (71) and (72), in the event Ω, we have

$$\text{II} \lesssim (n^{2/\gamma}\sqrt{s}\lambda + s\lambda)\sqrt{s}\lambda.$$

In addition,

$$\left\|\hat{Q}_{S^cS}\left(\hat{Q}_{SS}\right)^{-1} - \nabla^2\mathcal{R}(\beta^*)_{S^cS}\left(\nabla^2\mathcal{R}(\beta^*)_{SS}\right)^{-1}\right\|_\infty \lesssim (n^{2/\gamma}\sqrt{s}\lambda + s\lambda)\sqrt{s}.$$

Then, by (73), the incoherence condition and Proposition 2 in [46], strict dual feasibility holds under the scaling condition $n^{2/\gamma}s\lambda + s^{3/2}\lambda \lesssim 1$. We are able to apply Theorem 1 and Theorem 2 in [46].

Turning to ℓ_∞ error bounds, we have in the event Ω,

$$\begin{aligned}|\hat{\beta} - \beta^*|_\infty &\leq \left|\left(\hat{Q}_{SS}\right)^{-1}\nabla\mathcal{R}_n(\beta^*)_S\right|_\infty + \lambda\left\|\left(\hat{Q}_{SS}\right)^{-1}\right\|_\infty \\ &\leq \sqrt{s}\|\Delta_1\|_2\,|\nabla\mathcal{R}_n(\beta^*)_S|_\infty + \left|\left(\nabla^2\mathcal{R}(\beta^*)_{SS}\right)^{-1}\nabla\mathcal{R}_n(\beta^*)_S\right|_\infty + \lambda\left\|\left(\nabla^2\mathcal{R}(\beta^*)_{SS}\right)^{-1}\right\|_\infty \\ &\quad + \lambda\left\|\left(\hat{Q}_{SS}\right)^{-1} - \left(\nabla^2\mathcal{R}(\beta^*)_{SS}\right)^{-1}\right\|_\infty \\ &\lesssim (n^{2/\gamma}\sqrt{s}\lambda + s\lambda)\sqrt{s}\lambda + \left\|\left(\nabla^2\mathcal{R}(\beta^*)_{SS}\right)^{-1}\right\|_\infty \lambda \\ &\lesssim \left\|\left(\nabla^2\mathcal{R}(\beta^*)_{SS}\right)^{-1}\right\|_\infty \lambda,\end{aligned}$$

using the scaling condition. This implies the desired result by Theorem 2 in [46].

□

Lemma D.4. *Assume $|\beta^*|_1 \le L < \infty$, $|r'(x)| \le M_1 < \infty$ and $|r''(x)| \le M_2 < \infty$ for any $x \in \mathbb{R}$. Also assume $\mathbb{E}|X_{ij}^4| \le C < \infty$, for any $1 \le j \le p$, and $\mathbb{E}|Y_i|^4 \le C < \infty$. Let $\|X_{.j}\|_{4,\mathrm{GMC}} < \infty$ for some constant $0 < \rho_j < 1$, $\|Y_{.}\|_{4,\mathrm{GMC}} < \infty$ for some constant $0 < \rho_y < 1$, $\|Y_{.}\|_{\psi_\nu} < \infty$ and $\rho = \min\{\rho_j, \rho_y\} \in (0,1)$. Choose $\tau \asymp n^{1/4}(\log p)^{-1/4}(\log n)^{-1/2}$. As long as $(\log n)^{2\nu+1}(\log p/n)^{1/2} \le C_1$, it holds that*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \big(Y_i - r'(\widetilde{X}_i^\top \beta^*)\big)\widetilde{X}_i\right|_\infty > C_2(\log n)\sqrt{\frac{\log p}{n}}\right) \le n^{-C_3} + p^{-C_4}, \qquad (74)$$

where $C_1, C_3 > 0$ depend on $\|Y_{.}\|_{\psi_\nu}$, $C_2, C_4 > 0$ depend on L, ρ, $\max_j \|X_{.j}\|_{4,\mathrm{GMC}}$ and $\|Y_{.}\|_{4,\mathrm{GMC}}$.

Proof. Note that

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n \big(Y_i - r'(\widetilde{X}_i^\top \beta^*)\big)\widetilde{X}_i\right|_\infty &\le \left|\frac{1}{n}\sum_{i=1}^n Y_i\widetilde{X}_i - \mathbb{E}Y_i\widetilde{X}_i\right|_\infty + \left|\frac{1}{n}\sum_{i=1}^n r'(\widetilde{X}_i^\top\beta^*)\widetilde{X}_i - \mathbb{E}r'(\widetilde{X}_i^\top\beta^*)\widetilde{X}_i\right|_\infty \\ &\quad + \left|\mathbb{E}Y_i\widetilde{X}_i - \mathbb{E}r'(\widetilde{X}_i^\top\beta^*)\widetilde{X}_i\right|_\infty \\ &:= \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

We bound the above three terms one by one.

We first bound I. Let $\widetilde{Y}_i = \mathrm{sgn}(Y_i)(|Y_i| \wedge \tau_1)$, with $\tau_1 \asymp \tau \asymp n^{1/4}(\log p)^{-1/4}(\log n)^{-1/2}$. Then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n Y_i\widetilde{X}_{ij} - \mathbb{E}Y_i\widetilde{X}_{ij} &= \Big(\frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i\widetilde{X}_{ij} - \mathbb{E}\widetilde{Y}_i\widetilde{X}_{ij}\Big) + \mathbb{E}[(Y_i - \widetilde{Y}_i)\widetilde{X}_{ij}] + \frac{1}{n}\sum_{i=1}^n (Y_i - \widetilde{Y}_i)\widetilde{X}_{ij} \\ &= \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3.\end{aligned}$$

For I_1, by Hölder's inequality, for $\rho = \min\{\rho_j, \rho_y\}$,

$$\begin{aligned}&\rho^{-m}\sum_{i=m}^\infty \|\widetilde{Y}_i\widetilde{X}_{ij} - \widetilde{Y}_{i,\{0\}}\widetilde{X}_{ij,\{0\}}\|_2 \\ \le&\rho^{-m}\sum_{i=m}^\infty \|\widetilde{Y}_i\|_4\|\widetilde{X}_{ij} - \widetilde{X}_{ij,\{0\}}\|_4 + \rho^{-m}\sum_{i=m}^\infty \|\widetilde{Y}_i - \widetilde{Y}_{i,\{0\}}\|_4\|\widetilde{X}_{ij,\{0\}}\|_4 \\ \le&\|X_{.j}\|_{4,\mathrm{GMC}}\|\widetilde{Y}_i\|_4 + \|\widetilde{X}_{ij}\|_4\|Y_{.}\|_{4,\mathrm{GMC}}.\end{aligned}$$

The last inequality follows from the property that the function $\mathrm{sgn}(x)(|x| \wedge \tau_1) = (x \wedge \tau_1) \vee (-\tau_1)$ is Lipschitz continuous and bounded. It follows that

$$\|\widetilde{Y}_{.}\widetilde{X}_{.j}\|_{2,\mathrm{GMC}} \le 2\|X_{.j}\|_{4,\mathrm{GMC}}\|Y_{.}\|_{4,\mathrm{GMC}}.$$

Applying Theorem A.3 with $x \asymp \sqrt{n\log p(\log n)^2}$ and as $|\widetilde{X}_{ij}\widetilde{Y}_i| \le \tau_1\tau$, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_j \left|\sum_{i=1}^n \widetilde{X}_{ij}\widetilde{Y}_i - \mathbb{E}\widetilde{X}_{ij}\widetilde{Y}_i\right| \ge c_1\sqrt{n\log p(\log n)^2}\right)\\
&\le 2p\exp\left(-\frac{x^2}{C_\rho\|\widetilde{X}_{\cdot j}\widetilde{Y}_{\cdot}\|_{2,\mathrm{GMC}}^2 + C_\rho(\tau_1\tau)^2 + C_\rho(\tau_1\tau)(\log n)^2 x}\right)\\
&\le p^{-c_2},
\end{aligned}$$

for some $c_1, c_2 > 0$ depend on ρ and $\|\widetilde{Y}_{\cdot}\widetilde{X}_{\cdot j}\|_{2,\mathrm{GMC}}$. Hence, with probability at least $1 - p^{-c_2}$,

$$\max_j \left| n^{-1}\sum_{i=1}^n \widetilde{X}_{ij}\widetilde{Y}_i - \mathbb{E}\widetilde{X}_{ij}\widetilde{Y}_i\right| \le c_1\sqrt{\frac{(\log n)^2\log p}{n}}.$$

For I_2, we have

$$\mathbb{E}[(Y_i - \widetilde{Y}_i)\widetilde{X}_{ij}] \le \mathbb{E}|\widetilde{X}_{ij}Y_i\mathbf{1}_{\{|Y_i|>\tau_1\}}| \le \frac{\mathbb{E}|\widetilde{X}_{ij}Y_i^2|}{\tau_1} \le \frac{C}{\tau_1}.$$

For I_3 with any $1 \le j \le p$,

$$\mathbb{P}(\mathrm{I}_3 \ne 0) \le \mathbb{P}\left(\bigcup_{i=1}^n \{|Y_i| > \tau_1\}\right) \le \sum_{i=1}^n \mathbb{P}\left(|Y_i| > \tau_1\right).$$

As $\|Y_{\cdot}\|_{\psi_\nu} < \infty$, for any $x > 0$,

$$\mathbb{P}(|Y_i| > x) \le c_3\exp(-c_4 x^{1/\nu}).$$

If $x \asymp (\log n)^\nu \lesssim \tau_1 \asymp n^{1/4}(\log p)^{-1/4}(\log n)^{-1/2}$, we have

$$\mathbb{P}(\mathrm{I}_3 \ne 0) \le n^{-c_5}.$$

Next, we consider II. Similarly to I_1, setting $\rho = \min\{\rho_j, \rho_y\}$, we can show

$$\begin{aligned}
&\rho^{-m}\sum_{i=m}^\infty \|r'(\widetilde{X}_i^\top\beta^*)\widetilde{X}_{ij} - r'(\widetilde{X}_{i,\{0\}}^\top\beta^*)\widetilde{X}_{ij,\{0\}}\|_2\\
\le&\rho^{-m}\sum_{i=m}^\infty \|[r'(\widetilde{X}_i^\top\beta^*) - r'(\widetilde{X}_{i,\{0\}}^\top\beta^*)]\widetilde{X}_{ij}\|_2 + \rho^{-m}\sum_{i=m}^\infty \|r'(\widetilde{X}_{i,\{0\}}^\top\beta^*)(\widetilde{X}_{ij} - \widetilde{X}_{ij,\{0\}})\|_2\\
\le&\rho^{-m}\sum_{i=m}^\infty\sum_{k=1}^p M_2|\beta_k^*|\|\widetilde{X}_{ik} - \widetilde{X}_{ik,\{0\}}\|_4\|\widetilde{X}_{ij}\|_4 + \rho^{-m}\sum_{i=m}^\infty M_1\|\widetilde{X}_{ij} - \widetilde{X}_{ij,\{0\}}\|_2\\
\le&\sum_{k=1}^p M_2|\beta_k^*|\|X_{\cdot k}\|_{4,\mathrm{GMC}}\|\widetilde{X}_{ij}\|_4 + M_1\|X_{\cdot j}\|_{2,\mathrm{GMC}}\\
\le&M_2 L\max_k \|X_{\cdot k}\|_{4,\mathrm{GMC}}\|X_{\cdot j}\|_{4,\mathrm{GMC}} + M_1\|X_{\cdot j}\|_{2,\mathrm{GMC}}.
\end{aligned}$$

Hence,

$$\|r'(\widetilde{X}_{\cdot}^{\top}\beta^*)X_{\cdot j}\|_{2,\mathrm{GMC}} \le M_2 L \max_k \|X_{\cdot k}\|_{4,\mathrm{GMC}}\|X_{\cdot j}\|_{4,\mathrm{GMC}} + M_1\|X_{\cdot j}\|_{2,\mathrm{GMC}} < \infty.$$

Applying Theorem A.3 again with $x \asymp \sqrt{n\log p(\log n)^2}$ and $|r'(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_{ij}| \le M_1\tau$, with probability at least $1-p^{-c_7}$, we have

$$\max_j \left|\frac{1}{n}\sum_{i=1}^n r'(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_{ij} - \mathbb{E}r'(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_{ij}\right| \le c_6(\log n)\sqrt{\frac{\log p}{n}},$$

where $c_6, c_7 > 0$ depend on M_1, M_2, L, ρ and $\|r'(\widetilde{X}_{\cdot}^{\top}\beta^*)X_{\cdot j}\|_{2,\mathrm{GMC}}$.

Finally, we bound III. As $\mathbb{E}(Y_i|X_i) = r'(X_i^{\top}\beta^*)$, for any $1 \le j \le p$,

$$\begin{aligned}\left|\mathbb{E}Y_i\widetilde{X}_{ij} - \mathbb{E}r'(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_{ij}\right| &= \mathbb{E}\left|[r'(X_i^{\top}\beta^*) - r'(\widetilde{X}_i^{\top}\beta^*)]\widetilde{X}_{ij}\right| \\ &\le \sum_{k=1}^p M_2|\beta_k^*|\mathbb{E}|(X_{ik}-\widetilde{X}_{ik})\widetilde{X}_{ij}| \\ &\le \sum_{k=1}^p M_2|\beta_k^*|\mathbb{E}|\widetilde{X}_{ij}X_{ik}\mathbf{1}_{\{|X_{ik}|>\tau\}}| \\ &\le \sum_{k=1}^p M_2|\beta_k^*|\mathbb{E}|\widetilde{X}_{ij}X_{ik}^2|/\tau \\ &\le M_2LC\tau^{-1}.\end{aligned}$$

Combining $\mathrm{I}_1, \mathrm{I}_2, \mathrm{I}_3$, II and III, we have the desired results. □

***Remark* D.2.** *If* $|Y_i| \le C < \infty$, *for example* Y_i *is a categorical variable, then the condition* $(\log n)^{2\nu+1}(\log p/n)^{1/2} \le C_1$ *can be removed.*

Proof of Lemma 3.3. We first prove the RSC of $\mathcal{R}_n(\beta)$ at $\beta = \beta^*$ over the cone $\mathcal{C}(S)$

$$\mathcal{C}(S) = \{\Delta \in \mathbb{R}^p : |\Delta_{S^c}|_1 \le 3|\Delta_S|_1\},$$

for $S \subset \{1, 2, ..., p\}$ and $|S|_0 = s$. Note that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \Delta^{\top} r''(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_i\widetilde{X}_i^{\top}\Delta \\ =&\Delta^{\top}\left(\mathbb{E}r''(X_i^{\top}\beta^*)X_iX_i^{\top}\right)\Delta + \Delta^{\top}\left(\frac{1}{n}\sum_{i=1}^n r''(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_i\widetilde{X}_i^{\top} - \mathbb{E}r''(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_i\widetilde{X}_i^{\top}\right)\Delta \\ &+ \Delta^{\top}\left(\mathbb{E}r''(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_i\widetilde{X}_i^{\top} - \mathbb{E}r''(X_i^{\top}\beta^*)X_iX_i^{\top}\right)\Delta. \qquad (75)\end{aligned}$$

For the third term in (75), we can show for any $1 \leq j, k \leq p$,

$$
\begin{aligned}
&\left|\mathbb{E}r''(\widetilde{X}_i^\top\beta^*)\widetilde{X}_{ij}\widetilde{X}_{ik} - \mathbb{E}r''(X_i^\top\beta^*)X_{ij}X_{ik}\right| \\
\leq &\left|\mathbb{E}\left(r''(\widetilde{X}_i^\top\beta^*) - \mathbb{E}r''(X_i^\top\beta^*)\right)\widetilde{X}_{ij}\widetilde{X}_{ik}\right| + \left|\mathbb{E}r''(X_i^\top\beta^*)(\widetilde{X}_{ij}\widetilde{X}_{ik} - X_{ij}X_{ik})\right| \\
\leq &\sum_{l=1}^{p} M_3|\beta_l^*| \cdot \mathbb{E}|(X_{il} - \widetilde{X}_{il})\widetilde{X}_{ij}\widetilde{X}_{ik}| + M_2\mathbb{E}|\widetilde{X}_{ij}\widetilde{X}_{ik} - X_{ij}X_{ik}| \\
\leq &\sum_{l=1}^{p} M_3|\beta_l^*| \cdot \mathbb{E}|X_{il}\widetilde{X}_{ij}\widetilde{X}_{ik}\mathbf{1}_{\{|X_{il}|>\tau\}}| + M_2\mathbb{E}|X_{ij}X_{ik}(\mathbf{1}_{\{|X_{ij}|>\tau\}} + \mathbf{1}_{\{|X_{ik}|>\tau\}})| \\
\leq &\sum_{l=1}^{p} M_3|\beta_l^*| \cdot \frac{\mathbb{E}|X_{il}^2\widetilde{X}_{ij}\widetilde{X}_{ik}|}{\tau} + M_2\frac{\mathbb{E}|X_{ij}^2X_{ik}|}{\tau} + M_2\frac{\mathbb{E}|X_{ij}X_{ik}^2|}{\tau} \\
\leq &\frac{2CM_2 + CM_3L}{\tau}.
\end{aligned}
$$

As for the second term in (75), we shall bound the $\|r''(\widetilde{X}_\cdot^\top\beta^*)\widetilde{X}_{\cdot j}\widetilde{X}_{\cdot k}\|_{2,\mathrm{GMC}}$. By Hölder's inequality, for $\rho = \min_j\{\rho_j\}$,

$$
\begin{aligned}
&\rho^{-m}\sum_{i=m}^{\infty}\|r''(\widetilde{X}_i^\top\beta^*)\widetilde{X}_{ij}\widetilde{X}_{ik} - r''(\widetilde{X}_{i,\{0\}}^\top\beta^*)\widetilde{X}_{ij,\{0\}}\widetilde{X}_{ik,\{0\}}\|_2 \\
\leq &\rho^{-m}\sum_{i=m}^{\infty}\|[r''(\widetilde{X}_i^\top\beta^*) - r''(\widetilde{X}_{i,\{0\}}^\top\beta^*)]\widetilde{X}_{ij}\widetilde{X}_{ik}\|_2 + \rho^{-m}\sum_{i=m}^{\infty}\|r''(\widetilde{X}_{i,\{0\}}^\top\beta^*)[\widetilde{X}_{ij}\widetilde{X}_{ik} - \widetilde{X}_{ij,\{0\}}\widetilde{X}_{ik,\{0\}}]\|_2 \\
\leq &\rho^{-m}\sum_{i=m}^{\infty}\|r''(\widetilde{X}_i^\top\beta^*) - r''(\widetilde{X}_{i,\{0\}}^\top\beta^*)\|_6\|\widetilde{X}_{ij}\widetilde{X}_{ik}\|_3 + M_2\cdot\rho^{-m}\sum_{i=m}^{\infty}\|\widetilde{X}_{ij}\widetilde{X}_{ik} - \widetilde{X}_{ij,\{0\}}\widetilde{X}_{ik,\{0\}}\|_2 \\
\leq &\rho^{-m}\sum_{i=m}^{\infty}\sum_{l=1}^{p} M_3|\beta_l^*| \cdot \|\widetilde{X}_{il} - \widetilde{X}_{il,\{0\}}\|_6\|\widetilde{X}_{ij}\widetilde{X}_{ik}\|_3 + M_2\cdot\rho^{-m}\sum_{i=m}^{\infty}\|\widetilde{X}_{ij} - \widetilde{X}_{ij,\{0\}}\|_4\|\widetilde{X}_{ik}\|_4 \\
&\quad + M_2\cdot\rho^{-m}\sum_{i=m}^{\infty}\|\widetilde{X}_{ij,\{0\}}\|_4\|\widetilde{X}_{ik} - \widetilde{X}_{ik,\{0\}}\|_4 \\
\leq &M_3L\max_l\|X_{\cdot l}\|_{6,\mathrm{GMC}}\|X_{ij}\|_6\|X_{ik}\|_6 + M_2\|X_{\cdot j}\|_{4,\mathrm{GMC}}\|X_{ik}\|_4 + M_2\|X_{ij}\|_4\|X_{\cdot k}\|_{4,\mathrm{GMC}}.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\|r''(\widetilde{X}_\cdot^\top\beta^*)\widetilde{X}_{\cdot j}\widetilde{X}_{\cdot k}\|_{2,\mathrm{GMC}} \\
\leq &M_3L\max_l\|X_{\cdot l}\|_{6,\mathrm{GMC}}\|X_{ij}\|_6\|X_{ik}\|_6 + M_2\|X_{\cdot j}\|_{4,\mathrm{GMC}}\|X_{ik}\|_4 + M_2\|X_{ij}\|_4\|X_{\cdot k}\|_{4,\mathrm{GMC}}.
\end{aligned}
$$

Applying Theorem A.3 with $x \asymp \sqrt{n\log p(\log n)^2}$, as $|r''(\widetilde{X}_i^\top\beta^*)\widetilde{X}_{ij}\widetilde{X}_{ik}| \leq M_2\tau^2$ and $\tau \asymp n^{1/4}(\log p)^{-1/4}(\log n)^{-1/2}$, we have with probability at least $1 - p^{-c_1}$

$$
\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n} r''(\widetilde{X}_i^\top\beta^*)\widetilde{X}_{ij}\widetilde{X}_{ik} - \mathbb{E}r''(\widetilde{X}_i^\top\beta^*)\widetilde{X}_{ij}\widetilde{X}_{ik}\right| \leq c_2(\log n)\sqrt{\frac{\log p}{n}}, \tag{76}
$$

where c_1, c_2 depend on ρ and $\|r''(\widetilde{X}_{\cdot}^{\top}\beta^*)\widetilde{X}_{\cdot j}\widetilde{X}_{\cdot k}\|_{2,\mathrm{GMC}}$. Hence, employing (76) and the upper bound of $\left|\mathbb{E}r''(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_{ij}\widetilde{X}_{ik} - \mathbb{E}r''(X_i^{\top}\beta^*)X_{ij}X_{ik}\right|$ into (75), we have, in an event with probability at least $1 - p^{-c_1}$,

$$\frac{1}{n}\sum_{i=1}^{n}\Delta^{\top}r''(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_i\widetilde{X}_i^{\top}\Delta \geq \kappa_{\mathrm{H}}|\Delta|_2^2 - c_3(\log n)\sqrt{\frac{\log p}{n}}|\Delta|_1^2, \tag{77}$$

where $c_1, c_3 > 0$ depend on ρ, M_2, M_3, L and $\max_j \|X_{\cdot j}\|_{6,\mathrm{GMC}}$.

Next, we bound the difference of $n^{-1}\sum_{i=1}^{n}\Delta^{\top}r''(\widetilde{X}_i^{\top}\beta)\widetilde{X}_i\widetilde{X}_i^{\top}\Delta$ and $n^{-1}\sum_{i=1}^{n}\Delta^{\top}r''(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_i\widetilde{X}_i^{\top}\Delta$ to control LRSC within a neighborhood of β^*. Basic calculation shows that

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}\Delta^{\top}r''(\widetilde{X}_i^{\top}\beta)\widetilde{X}_i\widetilde{X}_i^{\top}\Delta - \frac{1}{n}\sum_{i=1}^{n}\Delta^{\top}r''(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_i\widetilde{X}_i^{\top}\Delta\right|\\ =&\left|\frac{1}{n}\sum_{i=1}^{n}[r''(\widetilde{X}_i^{\top}\beta) - r''(\widetilde{X}_i^{\top}\beta^*)]\Delta^{\top}\widetilde{X}_i\widetilde{X}_i^{\top}\Delta\right|\\ \leq&\max_i\left|r''(\widetilde{X}_i^{\top}\beta) - r''(\widetilde{X}_i^{\top}\beta^*)\right|\cdot\frac{1}{n}\sum_{i=1}^{n}(\widetilde{X}_i^{\top}\Delta)^2.\end{aligned} \tag{78}$$

By the decomposition

$$\frac{1}{n}\sum_{i=1}^{n}(\widetilde{X}_i^{\top}\Delta)^2 = \Delta^{\top}\big(\mathbb{E}X_iX_i^{\top}\big)\Delta + \Delta^{\top}\big(\mathbb{E}\widetilde{X}_i\widetilde{X}_i^{\top} - \mathbb{E}X_iX_i^{\top}\big)\Delta + \Delta^{\top}\big(\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i\widetilde{X}_i^{\top} - \mathbb{E}\widetilde{X}_i\widetilde{X}_i^{\top}\big)\Delta,$$

as $\sup_{|\nu|_2=1}\mathbb{E}|X_i^{\top}\nu|^2 \leq c_0$ and $\Delta \in \mathcal{C}(S)$, similarly to the proof of (75), we can obtain with probability at least $1 - p^{-c_4}$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\widetilde{X}_i^{\top}\Delta)^2 &\leq c_0|\Delta|_2^2 + c'\log n\sqrt{\frac{\log p}{n}}\cdot|\Delta|_1^2 \leq c_0|\Delta|_2^2 + c's\log n\sqrt{\frac{\log p}{n}}\cdot|\Delta|_2^2\\ &\leq 2c_0|\Delta|_2^2.\end{aligned}$$

Meanwhile, as $|\beta - \beta^*|_2^2 \leq C_1 s\lambda^2$, over the cone $\mathcal{C}(S)$,

$$\max_i\left|r''(\widetilde{X}_i^{\top}\beta) - r''(\widetilde{X}_i^{\top}\beta^*)\right| \leq M_3\tau|\beta - \beta^*|_1 \leq M_3\tau\cdot 4|\beta_S - \beta_S^*|_1 \leq 4M_3\tau s\lambda.$$

Hence, it follows from (78) that

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}\Delta^{\top}r''(\widetilde{X}_i^{\top}\beta)\widetilde{X}_i\widetilde{X}_i^{\top}\Delta - \frac{1}{n}\sum_{i=1}^{n}\Delta^{\top}r''(\widetilde{X}_i^{\top}\beta^*)\widetilde{X}_i\widetilde{X}_i^{\top}\Delta\right|\\ \leq&8c_0M_3s\sqrt{\lambda}|\Delta|_2^2 \leq (\kappa_{\mathrm{H}}/3)|\Delta|_2^2.\end{aligned} \tag{79}$$

Combining (77) and (79), we conclude that

$$\frac{1}{n}\sum_{i=1}^{n}\Delta^\top r''(\widetilde{X}_i^\top\beta)\widetilde{X}_i\widetilde{X}_i^\top\Delta \geq \kappa_{\mathrm{H}}|\Delta|_2^2 - (\kappa_{\mathrm{H}}/3)|\Delta|_2^2 - c_3 s(\log n)\sqrt{\frac{\log p}{n}}|\Delta|_2^2$$
$$\geq (\kappa_{\mathrm{H}}/2)|\Delta|_2^2.$$

□

Proof of Theorem 3.3. Theorem 3.3 follows from Lemma D.4, Lemma 3.3 and Lemma D.1. □

Proof of Corollary 3.3. The proof is similar to Corollary 3.1. Thus it is omitted. □

Lemma D.5. *Assume $|\beta^*|_1 \leq L < \infty$. Also assume $\mathbb{E}|X_{ij}^4| \leq C < \infty$, for any $1 \leq j \leq p$, and $\mathbb{E}|Y_i|^4 \leq C < \infty$. Let $\|X_{\cdot j}\|_{4,\mathrm{GMC}} < \infty$ for some constant $0 < \rho_j < 1$, $\|Y_\cdot\|_{4,\mathrm{GMC}} < \infty$ for some constant $0 < \rho_y < 1$, and $\rho = \min\{\rho_j, \rho_y\} \in (0,1)$. Choose $\tau_1, \tau_2 \asymp n^{1/4}(\log p)^{-1/4}(\log n)^{-1/2}$. It holds that*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i(\widetilde{Y}_i - \widetilde{X}_i^\top\beta^*)\right|_\infty > C_1(\log n)\sqrt{\frac{\log p}{n}}\right) \leq p^{-C_2}, \tag{80}$$

where $C_1, C_2 > 0$ only depend on L, ρ, $\max_j \|X_{\cdot j}\|_{4,\mathrm{GMC}}$ and $\|Y_\cdot\|_{4,\mathrm{GMC}}$.

Proof. Note that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i(\widetilde{Y}_i - \widetilde{X}_i^\top\beta^*)\right|_\infty \leq \left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i\widetilde{Y}_i - \mathbb{E}\widetilde{X}_i\widetilde{Y}_i\right|_\infty + \left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i\widetilde{X}_i^\top\beta^* - \mathbb{E}\widetilde{X}_i\widetilde{X}_i^\top\beta^*\right|_\infty$$
$$+ \left|\mathbb{E}\widetilde{X}_i\widetilde{Y}_i - \mathbb{E}X_iY_i\right|_\infty + \left|\mathbb{E}X_iY_i - \mathbb{E}\widetilde{X}_i\widetilde{X}_i^\top\beta^*\right|_\infty$$
$$:= \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}.$$

We bound the four terms one by one. For I, by Hölder's inequality,

$$\|\widetilde{X}_{ij}\widetilde{Y}_i - \widetilde{X}_{ij,\{0\}}\widetilde{Y}_{i,\{0\}}\|_2 \leq \|\widetilde{X}_{ij} - \widetilde{X}_{ij,\{0\}}\|_4\|\widetilde{Y}_{i,\{0\}}\|_4 + \|\widetilde{X}_{ij}\|_4\|\widetilde{Y}_i - \widetilde{Y}_{i,\{0\}}\|_4.$$

It follows that, for $\rho = \min\{\rho_j, \rho_y\}$,

$$\rho^{-m}\sum_{i=m}^{\infty}\|\widetilde{X}_{ij}\widetilde{Y}_i - \widetilde{X}_{ij,\{0\}}\widetilde{Y}_{i,\{0\}}\|_2$$
$$\leq \rho^{-m}\sum_{i=m}^{\infty}\|\widetilde{X}_{ij} - \widetilde{X}_{ij,\{0\}}\|_4\|\widetilde{Y}_{i,\{0\}}\|_4 + \rho^{-m}\sum_{i=m}^{\infty}\|\widetilde{X}_{ij}\|_4\|\widetilde{Y}_i - \widetilde{Y}_{i,\{0\}}\|_4$$
$$\leq \|X_{\cdot j}\|_{4,\mathrm{GMC}}\|\widetilde{Y}_i\|_4 + \|\widetilde{X}_{ij}\|_4\|Y_\cdot\|_{4,\mathrm{GMC}}.$$

That is, $\|\widetilde{X}_{.j}\widetilde{Y}_{.}\|_{2,\mathrm{GMC}} \leq 2\|X_{.j}\|_{4,\mathrm{GMC}}\|Y_{.}\|_{4,\mathrm{GMC}}$. Applying Theorem A.3 with $x \asymp \sqrt{n\log p(\log n)^2}$ and $|\widetilde{X}_{ij}\widetilde{Y}_i| \leq \tau_1\tau_2$, we have

$$\begin{aligned}\mathbb{P}\left(\max_{j}\left|\sum_{i=1}^{n}\widetilde{X}_{ij}\widetilde{Y}_i - \mathbb{E}\widetilde{X}_{ij}\widetilde{Y}_i\right| \geq x\right) &\leq 2p\exp\left(-\frac{x^2}{C\|\widetilde{X}_{.j}\widetilde{Y}_{.}\|^2_{2,\mathrm{GMC}} + C(\tau_1\tau_2)^2 + C(\tau_1\tau_2)(\log n)^2 x}\right)\\ &\leq p^{-C_1},\end{aligned}$$

for some $C_1 > 2$.

For II, we can obtain

$$\mathrm{II} \leq \max_{1\leq j,k\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{ij}\widetilde{X}_{ik} - \mathbb{E}\widetilde{X}_{ij}\widetilde{X}_{ik}\right|\|\beta^*\|_1 \leq L\max_{1\leq j,k\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{ij}\widetilde{X}_{ik} - \mathbb{E}\widetilde{X}_{ij}\widetilde{X}_{ik}\right|.$$

Similarly to I, $\|\widetilde{X}_{.j}\widetilde{X}_{.k}\|_{2,\mathrm{GMC}} \leq 2\|X_{.j}\|_{4,\mathrm{GMC}}\|X_{.k}\|_{4,\mathrm{GMC}}$. Applying Theorem A.3 again with $x \asymp \sqrt{n\log p(\log n)^2}$ and $|\widetilde{X}_{ij}\widetilde{X}_{ik}| \leq \tau_2^2$, we have

$$\mathbb{P}\left(\max_{1\leq j,k\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{ij}\widetilde{X}_{ik} - \mathbb{E}\widetilde{X}_{ij}\widetilde{X}_{ik}\right| \geq x\right) \leq p^{-C_2}.$$

Next, we bound III. For any $1 \leq j \leq p$,

$$\begin{aligned}\mathbb{E}\widetilde{X}_{ij}\widetilde{Y}_i - \mathbb{E}X_{ij}Y_i &= \mathbb{E}\widetilde{Y}_i(\widetilde{X}_{ij} - X_{ij}) + \mathbb{E}(\widetilde{Y}_i - Y_i)X_{ij}\\ &\leq \mathbb{E}|\widetilde{Y}_iX_{ij}\mathbf{1}_{\{|X_{ij}|\geq\tau_2\}}| + \mathbb{E}|Y_iX_{ij}\mathbf{1}_{\{|Y_i|\geq\tau_1\}}|\\ &\leq \frac{\mathbb{E}|Y_iX_{ij}^2|}{\tau_2} + \frac{\mathbb{E}|Y_i^2X_{ij}|}{\tau_2}\\ &\leq C\tau_2^{-1} \leq C\tau_2^{-2}.\end{aligned}$$

Finally, we bound IV. For any $1 \leq j \leq p$,

$$\begin{aligned}\mathbb{E}X_{ij}Y_i - \mathbb{E}\widetilde{X}_{ij}\widetilde{X}_i^\top\beta^* = \mathbb{E}X_{ij}X_i^\top\beta^* - \mathbb{E}\widetilde{X}_{ij}\widetilde{X}_i^\top\beta^* &= \sum_{k=1}^{p}|\beta_k^*|\mathbb{E}|X_{ij}X_{ik} - \widetilde{X}_{ij}\widetilde{X}_{ik}|\\ &\leq \sum_{k=1}^{p}|\beta_k^*|\mathbb{E}\big(|(X_{ij} - \widetilde{X}_{ij})X_{ik}| + |\widetilde{X}_{ij}(X_{ik} - \widetilde{X}_{ik})|\big)\\ &\leq \sum_{k=1}^{p}|\beta_k^*|\mathbb{E}|X_{ij}X_{ik}|\big(\mathbf{1}_{\{|X_{ij}|\geq\tau_2\}} + \mathbf{1}_{\{|X_{ik}|\geq\tau_2\}}\big)\\ &\leq \sum_{k=1}^{p}|\beta_k^*|\big(\mathbb{E}|X_{ij}^2X_{ik}|/\tau_2 + \mathbb{E}|X_{ij}X_{ik}^2|/\tau_2\big)\\ &\leq CL\tau_2^{-1} \leq CL\tau_2^{-2}.\end{aligned}$$

The the desired results follows from the upper bounds of $\mathrm{I}, \mathrm{II}, \mathrm{III}$ and IV. □

Proof of Lemma 4.1. Note that

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i\widetilde{X}_i^\top = \Big(\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i\widetilde{X}_i^\top - \mathbb{E}\widetilde{X}_i\widetilde{X}_i^\top\Big) + \Big(\mathbb{E}\widetilde{X}_i\widetilde{X}_i^\top - \mathbb{E}X_iX_i^\top\Big) + \mathbb{E}X_iX_i^\top.$$

We can decompose $n^{-1}\sum_{i=1}^{n}\beta^\top\widetilde{X}_i\widetilde{X}_i^\top\beta$ into three terms. For the first term, similarly to the proof of II in Lemma D.5, with $x \asymp \sqrt{n\log p(\log n)^2}$ and $|\widetilde{X}_{ij}\widetilde{X}_{ik}| \leq \tau_2^2$, we can show,

$$\mathbb{P}\left(\max_{1\leq j,k\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{ij}\widetilde{X}_{ik} - \mathbb{E}\widetilde{X}_{ij}\widetilde{X}_{ik}\right| \geq c_1(\log n)\sqrt{\frac{\log p}{n}}\right) \leq p^{-C_2}.$$

For the second term,

$$\begin{aligned}\mathbb{E}|X_{ij}X_{ik} - \widetilde{X}_{ij}\widetilde{X}_{ik}| &\leq \mathbb{E}|X_{ij}X_{ik}|\big(\mathbf{1}_{\{|X_{ij}|\geq\tau_2\}} + \mathbf{1}_{\{|X_{ik}|\geq\tau_2\}}\big)\\ &\leq \big(\mathbb{E}|X_{ij}^2X_{ik}|/\tau_2 + \mathbb{E}|X_{ij}X_{ik}^2|/\tau_2\big)\\ &\leq C\tau_2^{-1} \leq C\tau_2^{-2}.\end{aligned}$$

Therefore, we have for any $\beta \in \mathbb{R}^p$, with probability at least $1 - p^{-C_2}$,

$$\frac{1}{n}\sum_{i=1}^{n}\beta^\top\widetilde{X}_i\widetilde{X}_i^\top\beta \geq \beta^\top(\mathbb{E}X_iX_i^\top)\beta - c_1(\log n)\sqrt{\frac{\log p}{n}}\|\beta\|_1^2.$$

□

Proof of Theorem 4.1. Employing the same arguments as those in the proof of Theorem 2 in [23], Theorem 4.1 then follows from Lemma D.5 and Lemma 4.1. □

Proof of Corollary 3.3. The proof is similar to Corollary 3.1 in the paper and Corollary 1 in [46], and thus is omitted. Note that in linear regression, the hessian matrix does not depends on β, which will simplify the proof. □