Journal of Machine Learning Research 25 (2024) 1-68 Submitted 10/22; Revised 6/24; Published 10/24

Inference on High-dimensional Single-index Models with Streaming Data

Dongxiao Han* HANDONGXIAO@NANKAI.EDU.CN
School of Statistics and Data Science, KLMDASR, LEBPS, and LPMC
Nankai University
Tianjin, 300071,China

Jinhan Xie* JINHANXIE@163.COM
Yunnan Key Laboratory of Statistical Modeling and Data Analysis
Yunnan University
Kunming, 650091, China;
Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, AB, T6G 2G1, Canada

Jin Liu† LIUJIN@NANKAI.EDU.CN
School of Statistics and Data Science, KLMDASR, LEBPS, and LPMC
Nankai University
Tianjin, 300071,China

Liuquan Sun SLQ@AMT.AC.CN
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and School of Mathematical Sciences
University of Chinese Academy of Sciences
Beijing 100190, China

Jian Huang J.HUANG@POLYU.EDU.HK
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong, China

Bei Jiang BEI1@UALBERTA.CA
Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, AB, T6G 2G1, Canada

Linglong Kong LKONG@UALBERTA.CA
Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, AB, T6G 2G1, Canada

Editor: Debdeep Pati

Abstract

Traditional statistical methods are faced with new challenges due to streaming data. The major challenge is the rapidly growing volume and velocity of data, which makes storing

*. Co-first author
†. Corresponding author

such huge data sets in memory impossible. The paper presents an online inference framework for regression parameters in high-dimensional semiparametric single-index models with unknown link functions. The proposed online procedure updates only the current data batch and summary statistics of historical data instead of re-accessing the entire raw data set. At the same time, we do not need to estimate the unknown link function, which is a highly challenging task. In addition, a generalized convex loss function is used in the proposed inference procedure. To illustrate the proposed method, we use the Huber loss function and the negative log-likelihood of the logistic regression model. In this study, the asymptotic normality of the proposed online debiased Lasso estimators and the bounds of the proposed online Lasso estimators are investigated. To evaluate the performance of the proposed method, extensive simulation studies have been conducted. We provide applications to Nasdaq stock prices and financial distress data sets.

Keywords: high-dimensional data; lasso; single-index models; statistical inference; streaming data

1. Introduction

The rapid development of data collection techniques brings new challenges to developing online approaches to handle data in a streaming fashion. In such a data environment, it is often numerically challenging or sometimes infeasible to store the entire data set in memory. Consequently, the classical offline methods that involve the entire data set are less attractive or even infeasible due to computationally expensive. Instead, online methods can be used to process the out-of-memory data and make real-time decisions, which have been prevalent in economics, finance, machine learning, and statistics. Up to now, various online methods have been proposed. For example, the stochastic gradient descent (SGD) algorithm and its variants have been extended to the streaming settings; see Duchi and Singer (2009), Xiao (2010), Dekel et al. (2012), Chen et al. (2020), and Zhu et al. (2023). In addition, Lin and Xi (2011) considered an aggregated estimating equation for generalized linear models. Schifano et al. (2016) proposed online-updating algorithms and inferences applicable to linear models and estimation equations. Luo and Song (2020) suggested a renewable estimation and incremental inference to analyze streaming data sets using generalized linear models. The aforementioned online methods are developed for low-dimensional settings where the number of regressors is fixed and much smaller than the total sample size.

In recent years, a large amount of high-dimensional data streams, such as network flows, wireless sensor networks data, and multimedia streams have been generated; see Wang et al. (2017), Braverman et al. (2017), and Din et al. (2021). To analyze the above high-dimensional data streams, many online methods have been studied. For example, Langford et al. (2009) proposed an online ℓ_1-regularized method via a variant of the truncated SGD. Fan et al. (2018) developed the diffusion approximation approach to investigate SGD estimators. Gepperth and Pfülb (2021) presented an approach for the Gaussian mixture model via SGD with non-stationary, high-dimensional streaming data. Shi et al. (2021) introduced a valid inference method for single or low-dimensional regression coefficients via a recursive online-score estimation technique. Deshpande et al. (2023) considered a class of online estimators in a high-dimensional auto-regressive model. Han et al. (2021) proposed an online debiased lasso estimator for statistical inference with high-dimensional streaming data and further extended to the generalized linear models in Luo et al. (2023). The above existing estimation and inference procedures only focused on the linear or generalized linear models. However, much less is known under the potential misspecification of these commonly used models or more general models.

The single-index models (SIMs), which accommodate possible nonlinearity and avoid the curse of dimensionality simultaneously, are useful extensions of the linear regression model. Over the last few decades, the SIMs have been widely investigated in both the statistics and econometrics literature. In low-dimensional settings, the SIMs have been studied extensively in the literature, see Carroll et al. (1997), Xia et al. (2009), and Cui et al. (2011), among others. In high-dimensional settings, the SIMs have also attracted interest with various studies such as variable selection, estimation, and hypothesis testing. For example, Alquier and Biau (2013) introduced a PAC-Bayesian estimation approach for the sparse SIMs. Ganti et al. (2017) provided a suite of algorithms to learn the SIMs. Radchenko (2015) proposed a non-parametric least squares with an equality ℓ_1 constraint to simultaneous variable selection and estimation. Sign support recovery for the regression coefficient vector was studied by Neykov et al. (2016). Yang et al. (2017) considered the estimation problems of the parametric component of the SIMs. Zhang et al. (2020) proposed flexible regularized single-index quantile regression models for high-dimensional data. Eftekhari et al. (2021) conducted pointwise inference based on least squares. However, existing SIM estimation or inference methods have been studied on the fixed sample size before data collection and might not be suitable to implement the situations where data arrive in a streaming manner.

In this paper, we develop an online framework for real-time estimation and inference of regression parameters in SIMs with streaming data. Our proposed procedure is established based on general convex loss functions. We consider the Huber loss function (Huber, 1964) and the negative log-likelihood of the logistic regression model as two special examples to illustrate the proposed method. Unlike previous works, the proposed online estimators are updated via the current data batch and summary statistics of historical data without accessing the entire raw data set. Meanwhile, we do not need to estimate any unknown link functions at each stage. In addition, the proposed online method accounts for the sparsity features in a candidate set of covariates and provides a valid statistical inference procedure for regression parameters. Under certain regular conditions, we also show the consistency and asymptotic normality of the proposed online estimators, which provides us with a theoretical basis for carrying out real-time statistical inference with streaming data. In summary, in comparison with the literature, our contributions lie in the following four-fold. (i) Unlike traditional high-dimensional offline SIMs (Neykov et al., 2016; Eftekhari et al., 2021; Han et al., 2022, 2023), which have access to the entire raw data set, our proposed method utilizes the current data batch along with summary statistics of historical data. (ii) In contrast to high-dimensional linear or generalized linear models with streaming data (Han et al., 2021; Luo et al., 2023) that presuppose a second-order differentiable loss function, our proposed method targets the SIMs that focus on accommodating possible nonlinearity. Moreover, it suffices for our method that the loss function only has a first-order derivative. The Huber loss, known for its robustness to responses, serves as a notable example within our framework. (iii) To conduct the inference procedure, we need to obtain an approximated inverse matrix estimator for the inverse of the second-order derivative of the expected loss function. Different from the works of Han et al. (2021) and Luo et al. (2023), we utilize the methodology of Cai et al. (2011) to obtain this estimator instead of imposing stronger exact ℓ_0 sparsity conditions on the population inverse of the second-order derivative of the expected loss function. (iv) Our work presents the upper bounds for the proposed online Lasso estimators with sub-Gaussian random covariates, thereby easing the constraints on bounded covariates as shown in Luo et al. (2023). In addition, we provide an improved understanding of how the number of data batches impacts oracle inequalities within an online framework, differing from traditional oracle inequalities (Negahban et al., 2012).

The rest of this paper is organized as follows. In Section 2.1, we present the model settings. The proposed online estimation procedure with its theoretical property is presented in Section 2.2. Section 2.3 introduces the proposed online one-step procedure. Some examples are provided to illustrate the proposed method in Section 3. We evaluate the performance of the proposed procedure through simulation studies in Section 4. In Section 5, we apply the proposed method to the Nasdaq stock and financial distress data sets. Some discussions are given in Section 6. Technical details are deferred to the Appendices.

2. Model and Methodology

2.1 Single-Index Models

We consider the following high-dimensional SIMs (Neykov et al., 2016):

$$Y = f(\boldsymbol{X}^\top \boldsymbol{\beta}_0, \epsilon), \tag{1}$$

where Y is a response variable, $\boldsymbol{X}$ is a p-dimensional covariate vector, $\boldsymbol{\beta}_0$ is a p-dimensional vector of regression parameters, f is an unknown link function, and ϵ is an error term whose distribution is unspecified. Without loss of generality, we assume $E(\boldsymbol{X}) = 0$. Assume that $E(\boldsymbol{\beta}_0^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_0) = 1$ (Neykov et al., 2016; Eftekhari et al., 2021) for identifiability, where $\boldsymbol{\Sigma} = E(\boldsymbol{X}\boldsymbol{X}^\top)$. Consider a time point $m \geq 2$ with a total of $N_m = \sum_{j=1}^m n_j$ independent copies of $(Y, \boldsymbol{X})$ arriving in a sequence of m data batches, denoted by $\{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$, where n_j is the size of the batch $\mathcal{D}_j$. For any $1 \leq j \leq m$, denote the observations in $\mathcal{D}_j$ by $\{Y_i^{(j)}, \boldsymbol{X}_i^{(j)}\}_{i=1}^{n_j}$. The SIMs involve many existing models as special cases, such as the linear regression model and the logistic regression model.

2.2 Online Consistent Estimation

The recovery of $\boldsymbol{\beta}_0$ up to a scale under model (1) often depends on the linearity of expectation assumption (Li and Duan, 1989; Li, 1991; Neykov et al., 2016) given below:
Definition 1 (Linearity of Expectation) A p-dimensional random variable $\boldsymbol{W}$ is said to satisfy linearity of expectation in the direction of $\boldsymbol{\beta}$ if for any direction $\boldsymbol{b} \in \mathbb{R}^p$:

$$E(\boldsymbol{W}^\top \boldsymbol{b} | \boldsymbol{W}^\top \boldsymbol{\beta}) = c_{\boldsymbol{b}} \boldsymbol{W}^\top \boldsymbol{\beta} + a_{\boldsymbol{b}},$$

where $a_{\boldsymbol{b}}$ and $c_{\boldsymbol{b}}$ are two constants which may depend on the direction $\boldsymbol{b}$.

We consider estimating $\boldsymbol{\beta}_0$ up to a scalar by using a loss function $l(Y, \boldsymbol{X}^\top \boldsymbol{\beta})$. The following conditions are for the following Proposition 2.

(C1) $\boldsymbol{X}$ satisfies the linearity of expectation assumption in the direction of $\boldsymbol{\beta}_0$. In addition, $\boldsymbol{X}$ is independent of ϵ.

(C2) The function $(Y, \boldsymbol{X}^\top\boldsymbol{\beta}) \to l(Y, \boldsymbol{X}^\top\boldsymbol{\beta})$ is convex in $\boldsymbol{X}^\top\boldsymbol{\beta} \in \mathbb{R}$, and the function $\boldsymbol{\beta} \to E\{l(Y, \boldsymbol{X}^\top\boldsymbol{\beta})\}$ has a unique minimizer $\boldsymbol{\beta}^* \neq 0$.

The linearity of expectation assumption for $\boldsymbol{X}$ in condition (C1) is commonly used for the SIMs (Li and Duan, 1989; Neykov et al., 2016). Moreover, the independence between $\boldsymbol{X}$ and ϵ in condition (C1) is also adopted by Neykov et al. (2016). Condition (C2) is for the parameter identification. Based on conditions (C1) and (C2) and the Jensen's inequality, we can obtain that $\boldsymbol{\beta}^*$ equals to $\boldsymbol{\beta}_0$ up to a scalar.

Remark 1 *The linearity of expectation assumption in condition (C1) is widely assumed in the sufficient dimension reduction literature, including SIMs as special cases; see Li and Duan (1989), Li (1991), Eftekhari et al. (2021), Cai et al. (2023) and references therein for further discussions on such assumptions and their applicability. It is worth that this linearity of expectation is satisfied uniformly in all directions when $\boldsymbol{W}$ has an elliptical symmetric distribution, including the multivariate normal distribution and Student's t distribution; see Cambanis et al. (1981). The assumption of elliptical symmetry plays an important role in numerous theoretical developments and applications. Various tests have been proposed to test whether that assumption holds true or not; see Cassart et al. (2008) and Babić et al. (2021).*

To test the linearity of expectation assumption in condition (C1), one promising way is to test whether several principal components of covariates $\boldsymbol{X}$ is an elliptical symmetric distribution. When the assumption of elliptical symmetry for covariates $\boldsymbol{X}$ is violated, we can apply coordinatewise Gaussianization to transform covariates $\boldsymbol{X}$ into normal distributions, i.e., $\widehat{T}_j = \Phi^{-1}\{n\widehat{F}_j/(n+1)\}$. Here, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and $\widehat{F}_j$ denotes the empirical cumulative distribution function of the jth component of $\boldsymbol{X}$. Further details on coordinatewise Gaussianization can be found in Mai et al. (2023).

Proposition 2 *Suppose that conditions (C1) and (C2) hold. Then there exists some non-zero constant k_1 depending on $l(Y, \boldsymbol{X}^\top\boldsymbol{\beta})$ such that $\boldsymbol{\beta}^* = k_1\boldsymbol{\beta}_0$.*

Proposition 2 indicates that the loss function $l(Y, \boldsymbol{X}^\top\boldsymbol{\beta})$ can provide a leeway to perform estimation and inference for $\boldsymbol{\beta}_0$ up to the scalar k_1.

Remark 3 *Notice that our objective is to conduct estimation and inference for $\boldsymbol{\beta}_0$ up to the scalar k_1, it is not essential to let $k_1 \to 1$ or determine k_1. In fact, since it is impossible to derive the explicit expression of k_1, determining it is not feasible. In addition, as the expression of the loss function $l(Y, \boldsymbol{X}^\top\boldsymbol{\beta})$ does not incorporate the link function f, the estimation of f could be avoided. The ℓ_1 and ℓ_2 bounds of the differences between $\boldsymbol{\beta}_0$ up to the scalar k_1 and its corresponding Lasso estimators, and the asymptotic distributions of the debiased Lasso estimators are provided in the following Theorems 4 and 5, respectively.*

By Proposition 2, a consistent estimator of $\boldsymbol{\beta}_0$ up to the scalar k_1 can be derived by minimizing the following penalized empirical version of $E\{l(Y, \boldsymbol{X}^\top\boldsymbol{\beta})\}$ under some mild condition:

$$\frac{1}{N_m}\sum_{j=1}^{m}\sum_{i=1}^{n_j} l(Y_i^{(j)}, \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}) + \lambda_n\|\boldsymbol{\beta}\|_1,$$

where λ_n is a tuning parameter, $\|\boldsymbol{\beta}\|_1 = \sum_{l=1}^{p} |\beta_l|$ is the ℓ_1-norm of $\boldsymbol{\beta}$, and β_l is the lth element of $\boldsymbol{\beta}$. However, under the streaming data setting, since new data arrives continually, data volume accumulates very fast over time. This leads to the result that the raw data can not be stored in memory for a long time and we can not access the entire data set $\{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$ at the time point m, making it impossible to implement the algorithm above. To tackle this problem, we consider an online updating procedure which just exploit the current data and the summary statistics from the historical raw data for estimating $\boldsymbol{\beta}^*$. To remove the dependence between an estimator of $\boldsymbol{\beta}^*$ and the observed data, we employ a sample-splitting technique. Without this technique, it is difficult to obtain an upper bound for the $\|\cdot\|_\infty$ of the difference between $\boldsymbol{H}$ and its corresponding estimator when the second order derivative of $l(Y, \boldsymbol{X}^\top \boldsymbol{\beta})$ does not exist, where $\|\cdot\|_\infty$ is the maximum absolute value of the entries in a matrix. Without loss of generality, assume that $n_1, \ldots, n_m$ are all even numbers. Let $\mathcal{D}_{j,1} = \{Y_i^{(j)}, \boldsymbol{X}_i^{(j)}\}_{i=1}^{n_j/2}$, and $\mathcal{D}_{j,2} = \{Y_i^{(j)}, \boldsymbol{X}_i^{(j)}\}_{i=n_j/2+1}^{n_j}$, for $j = 1, \ldots, m$. Define

$$\boldsymbol{H} = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} E\{l(Y, \boldsymbol{X}^\top \boldsymbol{\beta})\}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}.$$

When the batch $\mathcal{D}_1$ arrives, let $\hat{\boldsymbol{\beta}}_1^{(1)}$ be the minimizer of

$$\frac{2}{n_1} \sum_{i=1}^{n_1/2} l(Y_i^{(1)}, \boldsymbol{X}_i^{(1)\top} \boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1, \tag{2}$$

and $\hat{\boldsymbol{\beta}}_2^{(1)}$ be the minimizer of

$$\frac{2}{n_1} \sum_{i=n_1/2+1}^{n_1} l(Y_i^{(1)}, \boldsymbol{X}_i^{(1)\top} \boldsymbol{\beta}) + \gamma_1 \|\boldsymbol{\beta}\|_1, \tag{3}$$

where λ_1 and γ_1 are two tuning parameters. Then we store $\{\hat{\boldsymbol{\beta}}_1^{(1)}, \hat{\boldsymbol{\beta}}_2^{(1)}, n_1 \boldsymbol{H}_1^{(1)}, n_1 \boldsymbol{H}_2^{(1)}\}$, where $\boldsymbol{H}_1^{(1)}$, and $\boldsymbol{H}_2^{(1)}$ are empirical versions of $\boldsymbol{H}$ which are obtained by using $\{\mathcal{D}_{1,1}, \hat{\boldsymbol{\beta}}_2^{(1)}\}$, and $\{\mathcal{D}_{1,2}, \hat{\boldsymbol{\beta}}_1^{(1)}\}$, respectively. For any time point $2 \leq s \leq m$, as the raw data $\{\mathcal{D}_1, \ldots \mathcal{D}_{s-1}\}$ is not stored, we consider replacing the cumulative objective function

$$\frac{2}{N_s} \sum_{j=1}^{s} \sum_{i=1}^{n_j/2} l(Y_i^{(j)}, \boldsymbol{X}_i^{(j)\top} \boldsymbol{\beta}) + \lambda_s \|\boldsymbol{\beta}\|_1, \tag{4}$$

with another function just including historical summary statistics $\{\hat{\boldsymbol{\beta}}_2^{(s-1)}, \sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)}\}$, and the current data set $\mathcal{D}_{s,1}$ to estimate $\boldsymbol{\beta}^*$ at the sth time point, where λ_s is a tuning parameter, $N_s = \sum_{j=1}^{s} n_j$, $\hat{\boldsymbol{\beta}}_2^{(s-1)}$ is an estimator of $\boldsymbol{\beta}^*$ at the $(s-1)$th time point by using $\{\hat{\boldsymbol{\beta}}_1^{(s-2)}, \mathcal{D}_{s-1,2}, \sum_{j=1}^{s-2} n_j \boldsymbol{H}_2^{(j)}\}$, and $\boldsymbol{H}_1^{(j)}$ is an empirical version of $\boldsymbol{H}$ which is acquired by using $\{\mathcal{D}_{j,1}, \hat{\boldsymbol{\beta}}_2^{(j)}\}$ at the jth time point, $j = 1, \ldots, s-1$. $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2^{(s-1)})^\top \boldsymbol{H}_1^{(j)} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2^{(s-1)})/2 + 2 \sum_{i=1}^{n_j/2} l(Y_i^{(j)}, \boldsymbol{X}_i^{(j)\top} \hat{\boldsymbol{\beta}}_2^{(s-1)})/n_j$ can be considered as an approximated second-order Taylor expansion of $2 \sum_{i=1}^{n_j/2} l(Y_i^{(j)}, \boldsymbol{X}_i^{(j)\top} \boldsymbol{\beta})/n_j$ at $\hat{\boldsymbol{\beta}}_2^{(s-1)}$. Then, motivated by Luo and Song

(2020), by replacing $2\sum_{i=1}^{n_j/2} l(Y_i^{(j)}, \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta})/n_j$ with $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2^{(s-1)})^\top \boldsymbol{H}_1^{(j)}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2^{(s-1)})/2 + 2\sum_{i=1}^{n_j/2} l(Y_i^{(j)}, \boldsymbol{X}_i^{(j)\top}\hat{\boldsymbol{\beta}}_2^{(s-1)})/n_j$ in (4), for $j = 1\ldots, s-1$, and removing constant terms, we can obtain the updating estimator $\hat{\boldsymbol{\beta}}_1^{(s)}$ at the sth time point by minimizing the following objective function:

$$L_{1s}(\boldsymbol{\beta}) + \lambda_s\|\boldsymbol{\beta}\|_1, \tag{5}$$

where $L_{1s}(\boldsymbol{\beta}) = [(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2^{(s-1)})^\top \sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2^{(s-1)})/2 + 2\sum_{i=1}^{n_s/2} l(Y_i^{(s)}, \boldsymbol{X}_i^{(s)\top}\boldsymbol{\beta})]/N_s$. Similarly, the updating estimator $\hat{\boldsymbol{\beta}}_2^{(s)}$ is given by

$$\hat{\boldsymbol{\beta}}_2^{(s)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\{L_{2s}(\boldsymbol{\beta}) + \gamma_s\|\boldsymbol{\beta}\|_1\}, \tag{6}$$

where $L_{2s}(\boldsymbol{\beta}) = [(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1^{(s-1)})^\top \sum_{j=1}^{s-1} n_j \boldsymbol{H}_2^{(j)}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1^{(s-1)})/2 + 2\sum_{i=n_s/2+1}^{n_s} l(Y_i^{(s)}, \boldsymbol{X}_i^{(s)\top}\boldsymbol{\beta})]/N_s$, γ_s is a tuning parameter, $\hat{\boldsymbol{\beta}}_1^{(s-1)}$ is an estimator of $\boldsymbol{\beta}^*$ at the $(s-1)$th time point by using $\{\hat{\boldsymbol{\beta}}_2^{(s-2)}, \mathcal{D}_{s-1,1}, \sum_{j=1}^{s-2} n_j\boldsymbol{H}_1^{(j)}\}$, and $\boldsymbol{H}_1^{(j)}$ is an empirical version of $\boldsymbol{H}$ which is got by using $\{\mathcal{D}_{j,2}, \hat{\boldsymbol{\beta}}_1^{(j)}\}$ at the jth time point, $j = 1, \ldots, s-1$. Then we take $\hat{\boldsymbol{\beta}}_{ave}^{(s)} = \{\hat{\boldsymbol{\beta}}_1^{(s)} + \hat{\boldsymbol{\beta}}_2^{(s)}\}/2$ as the final estimator at the sth step and store $\{\hat{\boldsymbol{\beta}}_1^{(s)}, \hat{\boldsymbol{\beta}}_2^{(s)}, \sum_{j=1}^{s} n_j\boldsymbol{H}_1^{(j)}, \sum_{j=1}^{s} n_j\boldsymbol{H}_2^{(j)}\}$, where $\boldsymbol{H}_1^{(s)}$, and $\boldsymbol{H}_2^{(s)}$ are empirical versions of $\boldsymbol{H}$ which are obtained by using $\{\mathcal{D}_{s,1}, \hat{\boldsymbol{\beta}}_2^{(s)}\}$, and $\{\mathcal{D}_{s,2}, \hat{\boldsymbol{\beta}}_1^{(s)}\}$, respectively. Since the loss function $l(Y, \boldsymbol{X}^\top\boldsymbol{\beta})$ does not incorporate the link function f, both $L_{1s}(\boldsymbol{\beta}) + \lambda_s\|\boldsymbol{\beta}\|_1$ and $L_{2s}(\boldsymbol{\beta}) + \gamma_s\|\boldsymbol{\beta}\|_1$ also exclude f. Consequently, the estimation of f is avoided in our proposed estimation procedure, which is detailed in Algorithm 1.

Algorithm 1 Online estimation for the SIMs.

Input: Streaming data sets $\mathcal{D}_1 \ldots \mathcal{D}_s \ldots$, and the tuning parameters $\lambda_1 \ldots \lambda_s \ldots$, $\gamma_1 \ldots \gamma_s \ldots$;

1. Calculate the offline lasso penalized estimators $\widehat{\boldsymbol{\beta}}_1^{(1)}$, $\widehat{\boldsymbol{\beta}}_2^{(1)}$ via (2) and (3) based on $\mathcal{D}_1$;
2. Update $n_1 H_1^{(1)}$ and $n_2 H_2^{(1)}$;
3. **for** $s = 2, 3, \ldots,$ **do**

(i). Read the current data set $\mathcal{D}_s$;

(ii). Calculate the online lasso penalized estimators $\widehat{\boldsymbol{\beta}}_1^{(s)}$ and $\widehat{\boldsymbol{\beta}}_2^{(s)}$ via (5) and (6);

(iii). Update and store the summary statistics $\{\widehat{\boldsymbol{\beta}}_1^{(s)}, \widehat{\boldsymbol{\beta}}_2^{(s)}, \sum_{j=1}^{s} n_j\boldsymbol{H}_1^{(j)}, \sum_{j=1}^{s} n_j\boldsymbol{H}_2^{(j)}\}$;

(iv). Calculate $\hat{\boldsymbol{\beta}}_{ave}^{(s)} = \{\hat{\boldsymbol{\beta}}_1^{(s)} + \hat{\boldsymbol{\beta}}_2^{(s)}\}/2$;

(v). Release data set $\mathcal{D}_s$ from the memory;

end for

Output: $\widehat{\boldsymbol{\beta}}_{ave}^{(s)}$ for $s = 1, 2, \ldots$

In what follows, we will provide the convergence rates of $\hat{\beta}_1^{(s)}$, $\hat{\beta}_2^{(s)}$, and $\hat{\boldsymbol{\beta}}_{ave}^{(s)}$, for $s = 1, \cdots, m$. Let $\|\cdot\|_2$ be the ℓ_2-norm (Euclidean norm). Define

$$N_1 = n_1,\ g_{\boldsymbol{\beta}}(Y, \boldsymbol{X}) = \partial l(Y, \boldsymbol{X}^\top \boldsymbol{\beta})/\partial \boldsymbol{\beta}\,, \boldsymbol{Z} = g_{\boldsymbol{\beta}^*}(Y, \boldsymbol{X}),$$
$$l_1^{(j)}(\boldsymbol{\beta}) = 2\sum_{i=1}^{n_j/2} l(Y_i^{(j)},\ \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta})/n_j,\ l_2^{(j)}(\boldsymbol{\beta}) = 2\sum_{i=n_j/2+1}^{n_j} l(Y_i^{(j)}, \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta})/n_j,$$
$$\nabla l_1^{(j)}(\boldsymbol{\beta}) = 2\sum_{i=1}^{n_j/2} g_{\boldsymbol{\beta}}(Y_i^{(j)}, \boldsymbol{X}_i^{(j)})/n_j, \text{ and } \nabla l_2^{(j)}(\boldsymbol{\beta}) = 2\sum_{i=n_j/2+1}^{n_j} g_{\boldsymbol{\beta}}(Y_i^{(j)}, \boldsymbol{X}_i^{(j)})/n_j.$$

For a p-dimensional random vector $\boldsymbol{\xi}$, define

$$||\boldsymbol{\xi}||_{\psi_2} = \sup_{\boldsymbol{a}\in\mathbb{R}^p, ||\boldsymbol{a}||_2=1} \sup_{k\geq 1} (E|\boldsymbol{a}^\top \boldsymbol{\xi}|^k)^{1/k}/\sqrt{k}.$$

In addition to conditions (C1) and (C2), the following conditions are required.

(C3) There exists a positive constant M_1 such that

$$||\boldsymbol{Z}||_{\psi_2} \leq M_1.$$

(C4) $\boldsymbol{\beta}_0$ is s_0-sparse with $s_0^3 \log p = o(n_1^{\alpha_1})$ for some $0 < \alpha_1 < 1$, where s_0 is the number of nonzero elements in $\boldsymbol{\beta}_0$.

(C5) There exist two positive constant M_2 and M_3 such that

$$M_2 \leq \inf_{\|\boldsymbol{\Delta}\|_2=1} \|\boldsymbol{H}^{1/2}\boldsymbol{\Delta}\|_2^2 \leq \sup_{\|\boldsymbol{\Delta}\|_2=1} \|\boldsymbol{H}^{1/2}\boldsymbol{\Delta}\|_2^2 \leq M_3.$$

(C6) There exist two positive constants M_4 and M_5 such that for any $1 \leq s \leq m$, with probability at least $1 - P(n_s, p)$,

$$l_1^{(s)}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - l_1^{(s)}(\boldsymbol{\beta}^*) - \boldsymbol{\Delta}^\top \nabla l_1^{(s)}(\boldsymbol{\beta}^*) \geq M_4||\boldsymbol{\Delta}||_2^2 - M_5\sqrt{\frac{\log p}{n_s}}||\boldsymbol{\Delta}||_1||\boldsymbol{\Delta}||_2,$$

and

$$l_2^{(s)}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - l_2^{(s)}(\boldsymbol{\beta}^*) - \boldsymbol{\Delta}^\top \nabla l_2^{(s)}(\boldsymbol{\beta}^*) \geq M_4||\boldsymbol{\Delta}||_2^2 - M_5\sqrt{\frac{\log p}{n_s}}||\boldsymbol{\Delta}||_1||\boldsymbol{\Delta}||_2,$$

for all $||\boldsymbol{\Delta}||_2 \leq 1$, where $\Omega(n_j, p)$ is a function of n_j.

(C7) There exists a positive number $M_6 \geq 1$ such that for any $1 \leq s \leq m$, with probability at least $1 - P_s(n_1, \cdots, n_s, p)$,

$$\begin{aligned}&\max\left\{\left\|\frac{1}{N_s}\sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)} - \boldsymbol{H}\right\|_\infty, \left\|\frac{1}{N_s}\sum_{j=1}^{s} n_j \boldsymbol{H}_2^{(j)} - \boldsymbol{H}\right\|_\infty\right\}\\ \leq&\frac{1}{N_s}\sum_{j=1}^{s} n_j M_6^j \sqrt{s_0}\max\{\frac{\log p}{n_j}, \sqrt{\frac{\log p}{n_j}}\}.\end{aligned}$$

where $P_s(n_1, \cdots, n_s, p)$ is a function of $n_1, \cdots, n_s$ and p.

(C8) Suppose that for some positive constant a_0 and any $1 \leq s \leq m$, $2^s s_0\sqrt{\log p/N_s} = o(1)$ and
$$\lim_{p\to\infty} 1 - P(n_s, p) - P_{s-1}(n_1, \ldots, n_{s-1}, p) - 2ep^{-a_0 N_s/n_s} = 1.$$

Condition (C3) assumes that $\boldsymbol{Z}$ has a sub-Gaussian tail. Condition (C4) is similar to the assumption in Janková and Van De Geer (2016). Condition (C5) indicates that $\boldsymbol{H}$ is positive definite and has finite eigenvalues. Many commonly-used loss functions such as the Huber loss (Huber, 1964) and the negative log-likelihood of generalized linear models satisfy condition (C6). The compliance of the Huber loss and the negative log-likelihood associated with the logistic regression model with condition (C6) is demonstrated in Lemmas 14 and 15 of the Appendix B, respectively. Moreover, condition (C7) is verifiable through mathematical induction, as detailed in the proofs of Corollaries 6 and 10. Condition (C8) can ensure the consistency of our online Lasso estimators. In Section 3, we provide the concrete forms of $P(n_s, p)$ and $P_s(n_1, \cdots, n_s, p)$ for specific examples and show that the condition $\lim_{p\to\infty} 1-P(n_s, p)-P_{s-1}(n_1, \ldots, n_{s-1}, p)-2ep^{-a_0 N_s/n_s} = 1$ in (C8) is satisfied under some mild conditions. Neykov et al. (2016) concentrated on variable selection consistency, while our work focuses on point estimation and pointwise inference for the regression parameter vector. In addition, Neykov et al. (2016) investigated the ordinary high-dimensional data, whereas our research is centered on high-dimensional streaming data. These distinctions markedly distinguish condition (C5) from the assumptions 2.3.1 and 2.3.2 presented in Neykov et al. (2016). Similarly, conditions (C4) and (C8) are notably different from the assumptions regarding n, p and s_0 in Neykov et al. (2016). It is worth pointing out that the condition (C5) has been used in high dimensional statistical inference, see Fan et al. (2017), van de Geer et al. (2014), Eftekhari et al. (2021) and references therein. The following Theorem 4 provides the consistency of $\hat{\boldsymbol{\beta}}_1^{(s)}$, $\hat{\boldsymbol{\beta}}_2^{(s)}$ and $\hat{\boldsymbol{\beta}}_{ave}^{(s)}$, for $s = 1, \cdots, m$.

Theorem 4 *Suppose that conditions (C1)-(C8) are satisfied. For any $1 \leq s \leq m$, assume $\lambda_s = c_{1s}\sqrt{\log p/N_s}$, and $\gamma_s = c_{2s}\sqrt{\log p/N_s}$, where c_{1s} and c_{2s} could be any constants which belong to $[2M_1\sqrt{2(a_0+1)/a_1}, a_2]$, a_1 is a positive constant not depending on any parameter, and a_2 could be any constant no less than $2M_1\sqrt{2(a_0+1)/a_1}$. If*
$$\max_{1\leq s\leq m-1} d_1^2 a_3^{2s-2} N_s^{\alpha_1/2-1/2} s M_6^s \leq A_1,$$
where A_1 could be any positive constant, $d_1 = \max\{3a_2/M_4, 4\}$, and
$$a_3 = \max\{(2M_3 + 3a_2/2)/\min\{M_2/3, M_4/2\}, 8 + 2M_3/\{M_1\sqrt{2(a_0+1)/a_1}\}\}.$$
Then for any $1 \leq s \leq m$, we have that with probability at least $1-P(n_s, p)-P_{s-1}(n_1, \ldots, n_{s-1}, p) - 2ep^{-a_0 N_s/n_s}$,
$$\begin{aligned}
&||\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}^*||_2 \leq d_s\sqrt{\frac{s_0\log p}{N_s}}, \quad ||\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}^*||_1 \leq d_s^2 s_0\sqrt{\frac{\log p}{N_s}},\\
&||\hat{\boldsymbol{\beta}}_2^{(s)} - \boldsymbol{\beta}^*||_2 \leq d_s\sqrt{\frac{s_0\log p}{N_s}}, \quad ||\hat{\boldsymbol{\beta}}_2^{(s)} - \boldsymbol{\beta}^*||_1 \leq d_s^2 s_0\sqrt{\frac{\log p}{N_s}},\\
&||\hat{\boldsymbol{\beta}}_{ave}^{(s)} - \boldsymbol{\beta}^*||_2 \leq d_s\sqrt{\frac{s_0\log p}{N_s}}, \quad \textit{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(s)} - \boldsymbol{\beta}^*||_1 \leq d_s^2 s_0\sqrt{\frac{\log p}{N_s}},
\end{aligned}$$
where e is Euler's number and $d_s = d_1 a_3^{s-1}$.

It is inevitable that the constants d_s and d_s^2 in Theorem 4 inherently depend on s due to the propagation of the estimation errors in the previous step to the current estimators. This dependency marks a deviation from the approach in traditional oracle inequalities (Van de Geer, 2008; Huang et al., 2013). This phenomenon is also observed in Theorem 1 of Luo et al. (2023). More details can be found in Remark 3 of Luo et al. (2023). In addition, we also conduct simulation studies in Section 4.1 to gain a clearer insight into how the upper bounds of the proposed estimator are influenced by the number of data batches m, in contrast to the traditional offline lasso estimator.

2.3 Online Pointwise Inference

We construct pointwise inference for the lth component of the regression parameter vector $\boldsymbol{\beta}^*$, for $l = 1, \cdots, p$. Since $\hat{\boldsymbol{\beta}}_1^{(s)}, \hat{\boldsymbol{\beta}}_2^{(s)}$ and $\hat{\boldsymbol{\beta}}_{ave}^{(s)}$ are not $N_s^{1/2}$ consistent by Theorem 4, we cannot obtain the asymptotic normalities of these estimators. Let β_l^* be the lth element of $\boldsymbol{\beta}^*$, $\boldsymbol{\Omega} = \boldsymbol{H}^{-1}$, and $\hat{\boldsymbol{\Omega}}_1^{(s)}$ and $\hat{\boldsymbol{\Omega}}_2^{(s)}$ be two estimators of $\boldsymbol{\Omega}$ which will be specified later. To tackle this issue, we first consider the following one-step estimator for β_l^* based on $\hat{\boldsymbol{\beta}}_1^{(s)}$ to increase the convergence rate:

$$\hat{\beta}_{1,l}^{one} = \hat{\beta}_{1,l}^{(s)} - \hat{\boldsymbol{\Omega}}_{1,l}^{(s)\top} \left\{ \sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)} (\hat{\boldsymbol{\beta}}_1^{(s)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) + n_s \nabla l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)}) \right\} / N_s,$$

where $\hat{\beta}_{1,l}^{(s)}$ is the lth element of $\hat{\boldsymbol{\beta}}_1^{(s)}$, and $\hat{\boldsymbol{\Omega}}_{1,l}^{(s)}$ is the lth column of $\hat{\boldsymbol{\Omega}}_1^{(s)}$. It can be shown that

$$\begin{aligned}
\hat{\beta}_{1,l}^{one} - \beta_l^* &= \hat{\beta}_{1,l}^{(s)} - \beta_l^* - \hat{\boldsymbol{\Omega}}_{1,l}^{(s)\top} \left\{ \sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)} (\hat{\boldsymbol{\beta}}_1^{(s)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) + n_s \nabla l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)}) \right\} / N_s \\
=& \boldsymbol{\Omega}_l^\top \boldsymbol{H} (\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}^*) - \hat{\boldsymbol{\Omega}}_{1,l}^{(s)\top} \left\{ \sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)} (\hat{\boldsymbol{\beta}}_1^{(s)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) + n_s \nabla l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)}) \right\} / N_s \\
=& \boldsymbol{\Omega}_l^\top \sum_{j=1}^{s} n_j (\boldsymbol{H} - \boldsymbol{H}_1^{(j)}) (\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}^*) / N_s \\
& - (\hat{\boldsymbol{\Omega}}_{1,l}^{(s)} - \boldsymbol{\Omega}_l)^\top \left\{ \sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)} (\hat{\boldsymbol{\beta}}_1^{(s)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) + n_s \nabla l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)}) \right\} / N_s \\
& - \boldsymbol{\Omega}_l^\top \left\{ \sum_{j=1}^{s-1} n_j H_1^{(j)} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(s-1)}) + n_s \nabla l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)}) - n_s H_1^{(s)} (\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}^*) \right\} / N_s, \qquad (7)
\end{aligned}$$

where $\boldsymbol{\Omega}_l$ is the lth column of $\boldsymbol{\Omega}$. By the proof of Theorem 5 in the Appendix A, the first two terms in (7) are $o_p(N_s^{-1/2})$ under some mild conditions. By the Taylor expansion, $\sum_{j=1}^{s} n_j H_1^{(j)} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(j)})$ can be estimated by $\sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\boldsymbol{\beta}^*) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)})$. Inspired by this, we consider the following decomposition for the third term,

$$
\begin{aligned}
&\boldsymbol{\Omega}_l^\top \left\{ \sum_{j=1}^{s-1} n_j H_1^{(j)} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(s-1)}) + n_s \nabla l_1^{(s)} (\hat{\boldsymbol{\beta}}_1^{(s)}) - n_s H_1^{(s)} (\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}^*) \right\} / N_s \\
=&\boldsymbol{\Omega}_l^\top \left\{ \sum_{j=1}^{s} n_j H_1^{(j)} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(j)}) \right\} / N_s \\
&+ \boldsymbol{\Omega}_l^\top \left\{ \sum_{j=1}^{s-1} n_j H_1^{(j)} (\hat{\boldsymbol{\beta}}_2^{(j)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) + n_s \nabla l_1^{(s)} (\hat{\boldsymbol{\beta}}_1^{(s)}) + n_s H_1^{(s)} (\hat{\boldsymbol{\beta}}_2^{(s)} - \hat{\boldsymbol{\beta}}_1^{(s)}) \right\} / N_s, \\
=&\boldsymbol{\Omega}_l^\top \left\{ \sum_{j=1}^{s} n_j H_1^{(j)} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(j)}) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)} (\boldsymbol{\beta}^*) + \sum_{j=1}^{s} n_j \nabla l_1^{(j)} (\hat{\boldsymbol{\beta}}_2^{(j)}) \right\} / N_s \\
&+ \boldsymbol{\Omega}_l^\top \left\{ \sum_{j=1}^{s-1} n_j H_1^{(j)} (\hat{\boldsymbol{\beta}}_2^{(j)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)} (\hat{\boldsymbol{\beta}}_2^{(j)}) \right. \\
&\left. + n_s \nabla l_1^{(s)} (\hat{\boldsymbol{\beta}}_1^{(s)}) + n_s H_1^{(s)} (\hat{\boldsymbol{\beta}}_2^{(s)} - \hat{\boldsymbol{\beta}}_1^{(s)}) \right\} / N_s \\
&+ \boldsymbol{\Omega}_l^\top \sum_{j=1}^{s} n_j \nabla l_1^{(j)} (\boldsymbol{\beta}^*) / N_s \\
=&\boldsymbol{\Omega}_l^\top \left\{ \sum_{j=1}^{s} n_j H_1^{(j)} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(j)}) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)} (\boldsymbol{\beta}^*) + \sum_{j=1}^{s} n_j \nabla l_1^{(j)} (\hat{\boldsymbol{\beta}}_2^{(j)}) \right\} / N_s \\
&+ (\boldsymbol{\Omega}_l - \hat{\boldsymbol{\Omega}}_{1,l}^{(s)})^\top \left\{ \sum_{j=1}^{s-1} n_j H_1^{(j)} (\hat{\boldsymbol{\beta}}_2^{(j)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)} (\hat{\boldsymbol{\beta}}_2^{(j)}) \right. \\
&\left. + n_s \nabla l_1^{(s)} (\hat{\boldsymbol{\beta}}_1^{(s)}) + n_s H_1^{(s)} (\hat{\boldsymbol{\beta}}_2^{(s)} - \hat{\boldsymbol{\beta}}_1^{(s)}) \right\} / N_s \\
&+ \hat{\boldsymbol{\Omega}}_{1,l}^{(s)\top} \left\{ \sum_{j=1}^{s-1} n_j H_1^{(j)} (\hat{\boldsymbol{\beta}}_2^{(j)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)} (\hat{\boldsymbol{\beta}}_2^{(j)}) \right. \\
&\left. + n_s \nabla l_1^{(s)} (\hat{\boldsymbol{\beta}}_1^{(s)}) + n_s H_1^{(s)} (\hat{\boldsymbol{\beta}}_2^{(s)} - \hat{\boldsymbol{\beta}}_1^{(s)}) \right\} / N_s \\
&+ \boldsymbol{\Omega}_l^\top \sum_{j=1}^{s} n_j \nabla l_1^{(j)} (\boldsymbol{\beta}^*) / N_s.
\end{aligned} \tag{8}
$$

Based on (7) and (8), we have

$$
\hat{\beta}_{1,l}^{one} - \beta_l^* = (I) + (II) + (III) + (IV) + (V) + (VI), \tag{9}
$$

where

$$
\begin{aligned}
(I) =&\boldsymbol{\Omega}_l^\top \sum_{j=1}^{s} n_j(\boldsymbol{H}-\boldsymbol{H}_1^{(j)})(\hat{\boldsymbol{\beta}}_1^{(s)}-\boldsymbol{\beta}^*)/N_s,\\
(II) =&-(\hat{\boldsymbol{\Omega}}_{1,l}^{(s)}-\boldsymbol{\Omega}_l)^\top \left\{\sum_{j=1}^{s-1} n_j\boldsymbol{H}_1^{(j)}(\hat{\boldsymbol{\beta}}_1^{(s)}-\hat{\boldsymbol{\beta}}_2^{(s-1)})+n_s\triangledown l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)})\right\}/N_s,\\
(III) =&-\boldsymbol{\Omega}_l^\top\left\{\sum_{j=1}^{s} n_j\boldsymbol{H}_1^{(j)}(\boldsymbol{\beta}^*-\hat{\boldsymbol{\beta}}_2^{(j)})-\sum_{j=1}^{s} n_j\triangledown l_1^{(j)}(\boldsymbol{\beta}^*)+\sum_{j=1}^{s} n_j\triangledown l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)})\right\}/N_s,\\
(IV) =&-(\boldsymbol{\Omega}_l-\hat{\boldsymbol{\Omega}}_{1,l}^{(s)})^\top\left\{\sum_{j=1}^{s-1} n_jH_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}-\hat{\boldsymbol{\beta}}_2^{(s-1)})-\sum_{j=1}^{s} n_j\triangledown l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)})+n_s\triangledown l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)})\right.\\
&\left.+n_sH_1^{(s)}(\hat{\boldsymbol{\beta}}_2^{(s)}-\hat{\boldsymbol{\beta}}_1^{(s)})\right\}/N_s,\\
(V) =&-\hat{\boldsymbol{\Omega}}_{1,l}^{(s)\top}\left\{\sum_{j=1}^{s-1} n_jH_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}-\hat{\boldsymbol{\beta}}_2^{(s-1)})-\sum_{j=1}^{s} n_j\triangledown l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)})+n_s\triangledown l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)})\right.\\
&\left.+n_sH_1^{(s)}(\hat{\boldsymbol{\beta}}_2^{(s)}-\hat{\boldsymbol{\beta}}_1^{(s)})\right\}/N_s,\\
(VI) =&\boldsymbol{\Omega}_l^\top\sum_{j=1}^{s} n_j\triangledown l_1^{(j)}(\boldsymbol{\beta}^*)/N_s.
\end{aligned}
$$

According to the proof of Theorem 5 in the Appendix A, we have shown that (I)-(IV) are $o_p(N_s^{-1/2})$, and (VI) multiply by $N_s^{-1/2}$ converges weakly to a normal distribution under some mild conditions. In addition, the order of (V) may be larger than $N_s^{-1/2}$. The decomposition of $\hat{\beta}_{1,l}^{one}-\beta_l^*$ implies that we need to minus (V) from (9) to acquire a new estimator of β_l^* which converges weakly to a normal distribution. As a result, we propose the following estimator for β_l^*:

$$
\begin{aligned}
\hat{\beta}_{1,l}^{d(s)} =&\hat{\beta}_{1,l}^{one}+\hat{\boldsymbol{\Omega}}_{1,l}^{(s)\top}\left\{\sum_{j=1}^{s-1} n_jH_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}-\hat{\boldsymbol{\beta}}_2^{(s-1)})-\sum_{j=1}^{s} n_j\triangledown l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)})\right.\\
&\left.+n_s\triangledown l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)})+n_sH_1^{(s)}(\hat{\boldsymbol{\beta}}_2^{(s)}-\hat{\boldsymbol{\beta}}_1^{(s)})\right\}/N_s\\
=&\hat{\beta}_{1,l}^{(s)}+\hat{\boldsymbol{\Omega}}_{1,l}^{(s)\top}\left\{\sum_{j=1}^{s} n_j\boldsymbol{H}_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}-\hat{\boldsymbol{\beta}}_1^{(s)})-\sum_{j=1}^{s} n_j\triangledown l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)})\right\}/N_s. \qquad (10)
\end{aligned}
$$

Similarly, we propose the following estimator for β_l^* based on $\hat{\boldsymbol{\beta}}_2^{(s)}$:

$$
\hat{\beta}_{2,l}^{d(s)}=\hat{\beta}_{2,l}^{(s)}+\hat{\boldsymbol{\Omega}}_{2,l}^{(s)\top}\left\{\sum_{j=1}^{s} n_j\boldsymbol{H}_2^{(j)}(\hat{\boldsymbol{\beta}}_1^{(j)}-\hat{\boldsymbol{\beta}}_2^{(s)})-\sum_{j=1}^{s} n_j\triangledown l_2^{(j)}(\hat{\boldsymbol{\beta}}_1^{(j)})\right\}/N_s, \qquad (11)
$$

where $\hat{\beta}_{2,l}^{(s)}$ is the lth element of $\hat{\boldsymbol{\beta}}_2^{(s)}$, and $\hat{\boldsymbol{\Omega}}_{2,l}^{(s)}$ is the lth column of $\hat{\boldsymbol{\Omega}}_2^{(s)}$. Subsequently, we propose an averaged estimator to avoid efficiency loss due to sample splitting:

$$\hat{\beta}_l^{da(s)} = \frac{\hat{\beta}_{1,l}^{d(s)} + \hat{\beta}_{2,l}^{d(s)}}{2}.$$

For a matrix $\boldsymbol{M} \in R^{p_0 \times p_1}$, let

$$\|\boldsymbol{M}\|_1 = \sum_{j_1=1}^{p_0}\sum_{j_2=1}^{p_1} |M_{j_1,j_2}|, \quad \text{and} \quad \|\boldsymbol{M}\|_{\infty,\infty} = \max_{1\le j_2 \le p_1} \sum_{j_1=1}^{p_0} |M_{j_1,j_2}|,$$

where M_{j_1,j_2} is the (j_1, j_2)th element of $\boldsymbol{M}$. To derive upper bounds for $\|\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}_1^{(s)}\|_{\infty,\infty}$ and $\|\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}_2^{(s)}\|_{\infty,\infty}$ easily, we use the method of Cai et al. (2011) to obtain $\hat{\boldsymbol{\Omega}}_1^{(s)}$ and $\hat{\boldsymbol{\Omega}}_2^{(s)}$. For simplicity, we only present the construction of $\hat{\boldsymbol{\Omega}}_1^{(s)}$. Note that $\hat{\boldsymbol{\Omega}}_2^{(s)}$ can be obtained via a similar way based on $\sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)}$ with the corresponding tuning parameter κ_s. Let $\hat{\boldsymbol{\Omega}}$ be the solution of the following optimization problem:

$$\min \|\tilde{\boldsymbol{\Omega}}\|_1 \quad \text{subject to} \quad \left\| \sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)} \tilde{\boldsymbol{\Omega}}/N_s - \boldsymbol{I}_p \right\|_\infty \le h_s, \tag{12}$$

where h_s is a tuning parameter and $\boldsymbol{I}_p$ is a unit matrix of size p. Note that the solution of (12) is not symmetric in general. The final estimator $\hat{\boldsymbol{\Omega}}_1^{(s)}$ is obtained by symmetrizing $\hat{\boldsymbol{\Omega}}$ as follows:

$$\hat{\Omega}_{1,j_1,j_2}^{(s)} = \hat{\Omega}_{1,j_2,j_1}^{(s)} = \hat{\Omega}_{j_1,j_2} I(|\hat{\Omega}_{j_1,j_2}| \le |\hat{\Omega}_{j_2,j_1}|) + \hat{\Omega}_{j_2,j_1} I(|\hat{\Omega}_{j_2,j_1}| < |\hat{\Omega}_{j_1,j_2}|),$$

where $\hat{\Omega}_{1,j_1,j_2}^{(s)}$, and $\hat{\Omega}_{j_1,j_2}$ are the (j_1, j_2)th elements of $\hat{\boldsymbol{\Omega}}_1^{(s)}$ and $\hat{\boldsymbol{\Omega}}$, respectively, and $\hat{\Omega}_{1,j_2,j_1}^{(s)}$, and $\hat{\Omega}_{j_2,j_1}$ are the (j_2, j_1)th elements of $\hat{\boldsymbol{\Omega}}_1^{(s)}$, and $\hat{\boldsymbol{\Omega}}$, respectively. Both (10) and (11) imply that $\{\sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)} \hat{\boldsymbol{\beta}}_2^{(j)} - \sum_{j=1}^{s-1} n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)})\}$ and $\{\sum_{j=1}^{s-1} n_j \boldsymbol{H}_2^{(j)} \hat{\boldsymbol{\beta}}_1^{(j)} - \sum_{j=1}^{s-1} n_j \nabla l_2^{(j)}(\hat{\boldsymbol{\beta}}_1^{(j)})\}$ should be stored as historical summary statistics at the $(s-1)$th step to acquire $\hat{\beta}_{1,l}^{d(s)}$ and $\hat{\beta}_{2,l}^{d(s)}$. In addition, we should also store T_s, which is defined as

$$\begin{aligned} T_s = &\frac{1}{N_s}\Bigg\{ \sum_{j=1}^{s}\sum_{i=1}^{n_j/2} g_{\hat{\boldsymbol{\beta}}_2^{(j)}}(Y_i^{(j)}, \boldsymbol{X}_i^{(j)}) g_{\hat{\boldsymbol{\beta}}_2^{(j)}}^\top(Y_i^{(j)}, \boldsymbol{X}_i^{(j)}) \\ &+ \sum_{j=1}^{s}\sum_{i=n_j/2+1}^{n_j} g_{\hat{\boldsymbol{\beta}}_1^{(j)}}(Y_i^{(j)}, \boldsymbol{X}_i^{(j)}) g_{\hat{\boldsymbol{\beta}}_1^{(j)}}^\top(Y_i^{(j)}, \boldsymbol{X}_i^{(j)}) \Bigg\} \end{aligned}$$

to estimate the asymptotic variance of $\sqrt{N_s}(\hat{\beta}_l^{da(s)} - \beta_l^*)$. Denote $Q_1^{(s-1)} = \sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)} \hat{\boldsymbol{\beta}}_2^{(j)} - \sum_{j=1}^{s-1} n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)})$ and $Q_2^{(s-1)} = \sum_{j=1}^{s-1} n_j \boldsymbol{H}_2^{(j)} \hat{\boldsymbol{\beta}}_1^{(j)} - \sum_{j=1}^{s-1} n_j \nabla l_2^{(j)}(\hat{\boldsymbol{\beta}}_1^{(j)})$. The proposed debiasing procedure is presented in the following Algorithm 2.

Let $\sigma_l^2 = \boldsymbol{\Omega}_l^\top E(\boldsymbol{Z}\boldsymbol{Z}^\top)\boldsymbol{\Omega}_l$. Additional conditions are needed to prove Theorem 5.

Algorithm 2 Online pointwise inference for the SIMs.

Input: Streaming data sets $\mathcal{D}_1 \ldots \mathcal{D}_s \ldots$;

1. Calculate the offline lasso penalized estimators $\widehat{\boldsymbol{\beta}}_1^{(1)}$, $\widehat{\boldsymbol{\beta}}_2^{(1)}$ via (2) and (3) based on $\mathcal{D}_1$;

2. Update $n_1 H_1^{(1)}$, $n_1 H_2^{(1)}$, $Q_1^{(1)}$, $Q_2^{(1)}$ and T_1;

3. **for** $s = 2, 3, \ldots,$ **do**

(i). Read the current data set $\mathcal{D}_s$;

(ii). Update online lasso penalized estimators $\widehat{\boldsymbol{\beta}}_1^{(s)}$ and $\widehat{\boldsymbol{\beta}}_2^{(s)}$ via Algorithm 1;

(iii). Update and store the summary statistics $\{\sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)}, \sum_{j=1}^{s} n_j \boldsymbol{H}_2^{(j)}, Q_1^{(s)}, Q_2^{(s)}, T_s\}$;

(iv). Calculate $\hat{\boldsymbol{\Omega}}_1^{(s)}$ and $\hat{\boldsymbol{\Omega}}_2^{(s)}$ by using (12);

(v). Update the online debiasing estimators $\hat{\beta}_{1,l}^{d(s)}$ and $\hat{\beta}_{2,l}^{d(s)}$ via (10) and (11);

(vi). Compute $\hat{\boldsymbol{\beta}}_l^{da(s)} = \{\hat{\boldsymbol{\beta}}_{1,l}^{da(s)} + \hat{\boldsymbol{\beta}}_{2,l}^{da(s)}\}/2$ and $\hat{\sigma}_{l,s}^2$ by (13);

(vii). Release data set $\mathcal{D}_s$ from the memory;

end for

Output: $\hat{\boldsymbol{\beta}}_l^{da(s)}$ and $\hat{\sigma}_{l,s}^2$ for $s = 1, 2, \ldots$

(D1) For any $1 \leq l \leq p$,

$$\sigma_l^2 \geq G_1,$$

where G_1 is a positive constant.

(D2) There exists a positive number $v(p)$ depending on p, and a positive constant ω which belongs to $[0, 1)$ such that for any $1 \leq s \leq m$,

$$\max\{\|\hat{\boldsymbol{\Omega}}_1^{(s)} - \boldsymbol{\Omega}\|_{\infty,\infty}, \|\hat{\boldsymbol{\Omega}}_2^{(s)} - \boldsymbol{\Omega}\|_{\infty,\infty}\} = O_p((g(s, s_0)\|\boldsymbol{\Omega}\|_{\infty,\infty}^4 \log p/N_s)^{(1-\omega)/2} v(p)),$$

where $g(s, s_0)$ is a function of s and s_0.

(D3) For any $1 \leq s \leq m$,

$$\|\boldsymbol{\Omega}\|_{\infty,\infty} \left\| \left\{ \sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(j)}) + \sum_{j=1}^{s} n_j \nabla l_1^{(j)} (\hat{\boldsymbol{\beta}}_2^{(j)}) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)} (\boldsymbol{\beta}^*) \right\} / N_s^{1/2} \right\|_\infty$$
$$= o_p(1),$$

and

$$\|\boldsymbol{\Omega}\|_{\infty,\infty} \left\| \left\{ \sum_{j=1}^{s} n_j \boldsymbol{H}_2^{(j)} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_1^{(j)}) + \sum_{j=1}^{s} n_j \nabla l_2^{(j)} (\hat{\boldsymbol{\beta}}_1^{(j)}) - \sum_{j=1}^{s} n_j \nabla l_2^{(j)} (\boldsymbol{\beta}^*) \right\} / N_s^{1/2} \right\|_\infty$$
$$= o_p(1).$$

(D4) For any $1 \le s \le m$,

$$\{g(s,s_0)\}^{(1-\omega)/2}\|\mathbf{\Omega}\|_{\infty,\infty}^{2(1-\omega)}a_3^{2s-2}s_0(\log p)^{1-\omega/2}v(p)N_s^{\omega/2-1/2}=o(1),$$
$$\|\mathbf{\Omega}\|_{\infty,\infty}a_3^{2s-2}d_1^2N_s^{\alpha_1/2-1/2}s\sqrt{\log p}M_6^s \le A_1,$$

and

$$\{g(s,s_0)\}^{(1-\omega)/2}\|\mathbf{\Omega}\|_{\infty,\infty}^{2(1-\omega)}sv(p)(\log p)^{1-\omega/2}a_3^{2s-2}d_1^2N_s^{\alpha_1/2+\omega/2-1}M_6^s \le A_1.$$

Condition (D1) assumes that the asymptotic variance of $\sqrt{N_s}(\hat{\beta}_l^{da(s)}-\beta_l^*)$ is bounded away from zero. Condition (D2) provides an upper bound for $\max\{\|\hat{\mathbf{\Omega}}_1^{(s)}-\mathbf{\Omega}\|_{\infty,\infty},\|\hat{\mathbf{\Omega}}_2^{(s)}-\mathbf{\Omega}\|_{\infty,\infty}\}$. When $\nabla l_1^{(j)}(\boldsymbol{\beta})$ is differentiable with respect to $\boldsymbol{\beta}$ for $1 \le j \le s$, the expression $\nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)})+\boldsymbol{H}_1^{(j)}(\boldsymbol{\beta}^*-\hat{\boldsymbol{\beta}}_2^{(j)})$ is the first order Taylor expansion of $\nabla l_1^{(j)}(\boldsymbol{\beta}^*)$ at $\hat{\boldsymbol{\beta}}_2^{(j)}$. In particular, we do not impose stronger exact ℓ_0 sparsity conditions on the population inverse of the second-order derivative of the expected loss function, in contrast to the node-wise lasso method in Han et al. (2021) and Luo et al. (2023). As a result, condition (D3) presents an upper bound for the orders of the $\|\cdot\|_\infty$ norm between the difference of the weighted summations of these $\nabla l_1^{(j)}(\boldsymbol{\beta}^*)$ and that of the corresponding first order Taylor expansions. Under the setting that $s=1$ and p is fixed, this condition is equivalent to $\|\hat{\boldsymbol{\beta}}_2^{(1)}-\boldsymbol{\beta}^*\|_2=o_p(N_1^{-1/4})$, which is easily verified under some mild conditions. For the high-dimensional setting with streaming data, it is challenging to obtain explicit orders of these $\|\hat{\boldsymbol{\beta}}_2^{(j)}-\boldsymbol{\beta}^*\|_2$ under this condition. However, we have shown that conditions (D2) and (D3) are satisfied in Corollaries 8 and 12 for the Huber loss and the negative log-likelihood associated with the logistic regression model, respectively. In addition, when $\max\{\log\{g(s,s_0)\},\log(\|\mathbf{\Omega}\|_{\infty,\infty}),s,\log s_0,\log\log p,\log\{v(p)\}\}=o(\log(N_s))$, condition (D4) is fulfilled. Conditions (D2)-(D4) can ensure that the first four terms on the right side of (9) are $o_p(N_s^{-1/2})$ by the proof of Theorem 5 in the Appendix A. As described in Subsection 2.2, the distinct data structures and statistical problems addressed in our work and by Neykov et al. (2016) lead to a significant divergence in condition (D4) from the assumptions regarding n, p, and s_0 found in Propositions 2.2.1 and 2.2.3, and Theorem 2.3.4 of Neykov et al. (2016). The following theorem demonstrates the asymptotic properties of $\sqrt{N_s}(\hat{\beta}_l^{da(s)}-\beta_l^*)$.

Theorem 5 *Under the conditions of Theorem 4, suppose that conditions (D1)-(D4) are satisfied. Then for any $1 \le s \le m$ and $1 \le l \le p$, we have that $\sigma_l^{-1}\sqrt{N_s}(\hat{\beta}_l^{da(s)}-\beta_l^*)$ converges to a standard normal random variable in distribution as $p \to \infty$.*

The asymptotic variance of $\sqrt{N_s}(\hat{\beta}_l^{da(s)}-\beta_l^*)$ can be estimated by

$$\hat{\sigma}_{l,s}^2=(\hat{\mathbf{\Omega}}_{1,l}^{(s)}+\hat{\mathbf{\Omega}}_{2,l}^{(s)})^\top T_s(\hat{\mathbf{\Omega}}_{2,l}^{(s)}+\hat{\mathbf{\Omega}}_{2,l}^{(s)})/4. \tag{13}$$

Then for any given significant level $\alpha \in (0,1)$, a $(1-\alpha)$ confidence interval for β_l^* is

$$[\hat{\beta}_l^{da(s)}-N_s^{-1/2}\hat{\sigma}_{l,s}z_{\alpha/2},\hat{\beta}_l^{da(s)}+N_s^{-1/2}\hat{\sigma}_{l,s}z_{\alpha/2}],$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of the standard normal distribution.

3. Examples

In this section, we provide two concrete examples to illustrate the proposed method.

3.1 Huber Loss

Actually, we often encounter data subject to heavily-tailed errors in finance and economics (Fan et al., 2017, 2021). The Huber loss as an important way of robustification has been well studied recently (Fan et al., 2017; Sun et al., 2020; Loh, 2021; Wang et al., 2021). The Huber loss function is defined as follows:

$$l(Y, \boldsymbol{X}^\top \boldsymbol{\beta}) = \rho_\tau(Y - \boldsymbol{X}^\top \boldsymbol{\beta}),$$

where

$$\rho_\tau(x) = \frac{x^2}{2} I(|x| \le \tau) + (\tau |x| - \frac{\tau^2}{2}) I(|x| > \tau),$$

for some constant $\tau > 0$. We can observe that the Huber loss is robust to the heavy-tailed observation noise due to the fact that the linear part of the Huber loss penalizes the residuals. Let $\boldsymbol{\beta}_\tau^* = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} E\{\rho_\tau(Y - \boldsymbol{X}^\top \boldsymbol{\beta})\}$, and $\epsilon_\tau = Y - \boldsymbol{X}^\top \boldsymbol{\beta}_\tau^*$. If ϵ_τ is a continuous random variable, then we have

$$\boldsymbol{H}_\tau = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} E\{\rho_\tau(Y - \boldsymbol{X}^\top \boldsymbol{\beta})\}|_{\boldsymbol{\beta} = \boldsymbol{\beta}_\tau^*} = E\{\boldsymbol{X}\boldsymbol{X}^\top I(|\epsilon_\tau| \le \tau)\},$$

$$\boldsymbol{H}_1^{(s)} = \frac{2}{n_s} \sum_{i=1}^{n_s/2} \boldsymbol{X}_i^{(s)} \boldsymbol{X}_i^{(s)\top} I(|Y_i^{(s)} - \boldsymbol{X}_i^{(s)\top} \hat{\boldsymbol{\beta}}_2^{(s)}| \le \tau),$$

and

$$\boldsymbol{H}_2^{(s)} = \frac{2}{n_s} \sum_{i=n_s/2+1}^{n_s} \boldsymbol{X}_i^{(s)} \boldsymbol{X}_i^{(s)\top} I(|Y_i^{(s)} - \boldsymbol{X}_i^{(s)\top} \hat{\boldsymbol{\beta}}_1^{(s)}| \le \tau), \quad s = 1, \cdots, m.$$

We can obtain the estimators $\hat{\boldsymbol{\beta}}_1^{(s)}$, $\hat{\boldsymbol{\beta}}_2^{(s)}$ and $\hat{\boldsymbol{\beta}}_{ave}^{(s)}$ by using the estimation procedure in Algorithm 1, for $s = 1, \cdots, m$.

The following conditions are needed to establish the consistency of $\hat{\boldsymbol{\beta}}_1^{(s)}$, $\hat{\boldsymbol{\beta}}_2^{(s)}$ and $\hat{\boldsymbol{\beta}}_{ave}^{(s)}$.

- (E1) There exists a positive constant e_1 such that for any $\tau > e_1$, $E\{\boldsymbol{X}\boldsymbol{X}^\top I(|Y - \boldsymbol{X}^\top \boldsymbol{\beta}| \le \tau)\} > 0$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$, and 0 is not the minimizer of the function $\boldsymbol{\beta} \to E\{\rho_\tau(Y - \boldsymbol{X}^\top \boldsymbol{\beta})\}$.
- (E2) There exists a positive constant B_1 such that $\|\boldsymbol{X}\|_{\psi_2} \le B_1$.
- (E3) There exist two positive constants B_2 and B_3 such that for any $\tau > e_1$,

$$E|\epsilon_\tau| \le B_2, \text{ and } B_3 \le \inf_{\|\boldsymbol{\Delta}\|_2 = 1} \|\boldsymbol{H}_\tau^{1/2} \boldsymbol{\Delta}\|_2^2 \le \sup_{\|\boldsymbol{\Delta}\|_2 = 1} \|\boldsymbol{H}_\tau^{1/2} \boldsymbol{\Delta}\|_2^2 \le B_2.$$

- (E4) There exist two positive constants B_4 and $0 < \alpha_2 < 1$ such that for any $2 \le s \le m$,

$$\frac{\log p}{n_s} \le B_4 \quad \text{or} \quad \log p / n_s > (\log p)^{\alpha_2}.$$

(E5) For any given $\tau > e_1$, there exists a positive constant L_τ depending on τ such that $\sup_{x\in R} f_{\epsilon_\tau|X}(x) \le L_\tau$ almost surely, where $f_{\epsilon_\tau|\boldsymbol{X}}(\cdot)$ is the conditional density function of ϵ_τ given $\boldsymbol{X}$.

(E6) $2^s s_0\sqrt{\log p/N_s} = o(1)$ for $1 \le s \le m$. There exists a positive number a_0' such that $m = o(\min(p^{a_0'}, p^{g_1}))$, where g_1 is a positive number depending on e_1, B_1, B_2, B_3 and B_4.

The assumption $E\{\boldsymbol{X}\boldsymbol{X}^\top I(|Y - \boldsymbol{X}^\top\boldsymbol{\beta}| \le \tau)\} > 0$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$ in condition (E1) suggests that $E\{\rho_\tau(Y - \boldsymbol{X}^\top\boldsymbol{\beta})\}$ is a strictly convex function of $\boldsymbol{\beta}$. Both this assumption and 0 is not the minimizer $E\{\rho_\tau(Y - \boldsymbol{X}^\top\boldsymbol{\beta})\}$ imply that condition (C2) is satisfied. Condition (E2) implies condition (C3). Conditions (E2)-(E4) suggest condition (C6). Conditions (E2)-(E5) lead to condition (C7). According to Lemma 14 and the proof of Corollary 6 in the Appendix B, we can obtain $P_s(n_1, \ldots, n_s, p) = 4sp^{-a_0'} - \sum_{j=1}^{s}\{\exp(-g_4 n_j - g_1\log p) + 2ep^{-a_0' N_j/n_j}\}$ and $P(n_s, p) = \exp(-g_4 n_s - g_1 \log p)$. This implies that condition (C8) is satisfied under condition (E6). Condition (E4) indicates that p can be arbitrary large as $\log p/n_s > (\log p)^{\alpha_2}$ satisfies, which seems contrary to common sense of high-dimensional analysis. However, to derive the subsequent Corollary 6, condition (C4) (i.e., $s_0^3 \log p = o(n_1^{\alpha_1})$) is also required. When considered in conjunction, these two assumptions become coherent. The data structure in this work is notably more complex compared to that in Han et al. (2022). Consequently, the assumptions for $n_s(N_s)$, p, s and s_0 (i.e., conditions (C4), (E4) and (E6)) in our analysis are more complicated than the single condition (C4) presented in Han et al. (2022).

The following Corollary 6 provides the ℓ_1 and ℓ_2 bounds for $\hat{\boldsymbol{\beta}}_1^{(s)}$, $\hat{\boldsymbol{\beta}}_2^{(s)}$ and $\hat{\boldsymbol{\beta}}_{ave}^{(s)}$ with sub-Gaussion predictor scenario.

Corollary 6 *Suppose that conditions (C1), (C4) and (E1)-(E6) hold. For any $1 \le s \le m$, assume $\lambda_s = c_{1s}'\sqrt{\log p/N_s}$ and $\gamma_s = c_{2s}'\sqrt{\log p/N_s}$, where c_{1s}' and c_{2s}' could be any constants which belong to $[2\tau B_1\sqrt{2(a_0'+1)/a_1}, a_2']$, and a_2' could be any constant no less than $2\tau B_1\sqrt{2(a_0'+1)/a_1}$. If $\tau \ge g_2$ and*

$$\max_{1\le s\le m-1} d_1'^2 a_3'^{2s-2} N_s^{\alpha_1/2-1/2} s M_\tau^s \le A_1',$$

where

$$a_3' = \max\{(2B_2 + 3a_2'/2)/\min\{B_3/3, g_3/2\}, 8 + 2B_2/\{\tau B_1\sqrt{2(a_0'+1)/a_1}\}\},$$
$$M_\tau = [\max\{\sqrt{32B_1^4(a_0'+2)/a_4'}, 8B_1^2(a_0'+2)/a_4'\} + 4\sqrt{2}L_\tau B_1^3 + 1]a_3'd_1',$$

A_1' could be any constant, $d_1' = \max\{3a_2'/g_3, 4\}$, a_4' is a positive constant not depending on any parameter, and g_2 and g_3 are two positive constants depending on e_1, B_1, B_2, B_3 and B_4. Then for any $1 \le s \le m$, we have that with probability at least $1 - 4(s-1)p^{-a_0'} -$

$\sum_{j=1}^{s}\{\exp(-g_4 n_j - g_1\log p) + 2ep^{-a_0' N_j/n_j}\}$,

$$||\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}_\tau^*||_2 \le d_s'\sqrt{\frac{s_0\log p}{N_s}}, \quad ||\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}_\tau^*||_1 \le d_s'^2 s_0\sqrt{\frac{\log p}{N_s}},$$
$$||\hat{\boldsymbol{\beta}}_2^{(s)} - \boldsymbol{\beta}_\tau^*||_2 \le d_s'\sqrt{\frac{s_0\log p}{N_s}}, \quad ||\hat{\boldsymbol{\beta}}_2^{(s)} - \boldsymbol{\beta}_\tau^*||_1 \le d_s'^2 s_0\sqrt{\frac{\log p}{N_s}},$$
$$||\hat{\boldsymbol{\beta}}_{ave}^{(s)} - \boldsymbol{\beta}_\tau^*||_2 \le d_s'\sqrt{\frac{s_0\log p}{N_s}}, \quad \textit{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(s)} - \boldsymbol{\beta}_\tau^*||_1 \le d_s'^2 s_0\sqrt{\frac{\log p}{N_s}},$$

where g_4 is a positive constant depending on e_1, B_1, B_2, B_3 and B_4 and $d_s' = d_1' a_3'^{s-1}$.

Based on the condition $\max_{1\le s\le m-1} d_1'^2 a_3'^{2s-2} N_s^{\alpha_1/2-1/2} s M_\tau^s \le A_1'$ and Corollary 6, we can obtain that the ℓ_1 and ℓ_2 norms of the difference between the estimators $\hat{\boldsymbol{\beta}}_1^{(s)}, \hat{\boldsymbol{\beta}}_2^{(s)}$, and $\hat{\boldsymbol{\beta}}_{ave}^{(s)}$ and β_τ^* are of orders $\sqrt{s_0^2\log p/(M_\tau^s s N_s^{\alpha_1/2+1/2})}$ and $\sqrt{s_0\log p/(M_\tau^{s/2}\sqrt{s}N_s^{\alpha_1/4+3/4})}$, respectively. When $\boldsymbol{X}$ follows a Gaussian distribution, we can simplify the assumptions and obtain a similar result. The following conditions are required.

(E7) $\boldsymbol{X}$ follows a Gaussian distribution and $\sup_{\|\boldsymbol{\Delta}\|_2=1} \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}\|_2^2 \le B_5$.

(E8) There exist two positive constants B_2 and B_3 such that for any $\tau > e_1$,

$$E|\epsilon_\tau| \le B_2, \text{ and } \inf_{\|\boldsymbol{\Delta}\|_2=1} \|\boldsymbol{H}_\tau^{1/2}\boldsymbol{\Delta}\|_2^2 \ge B_3.$$

(E9) $2^s s_0\sqrt{\log p/N_s} = o(1)$ for $1 \le s \le m$. There exists a positive number a_0' such that $m = o(\min(p^{a_0'}, p^{g_5}))$, where g_5 is a positive number depending on e_1, B_2, B_3, B_4 and B_5.

Under condition (E7), we have $\sup_{\|\boldsymbol{\Delta}\|_2=1} \|\boldsymbol{H}_\tau^{1/2}\boldsymbol{\Delta}\|_2^2 \le B_5$ and $\|\boldsymbol{X}\|_{\psi_2} \le B_6$ by the proof of Corollaries 7, 9, 11 and 13 in the Appendix B, where B_6 is a positive number depending on B_5. As a result, conditions (E7) and (E8) imply conditions (E2) and (E3). Condition (E9), which is similar to condition (E6), leads to condition (C8). In particular, when the predictors $\boldsymbol{X}$ follows the Gaussian distribution, the next Corollary 7 develops the ℓ_1 and ℓ_2 bounds for $\hat{\boldsymbol{\beta}}_1^{(s)}$, $\hat{\boldsymbol{\beta}}_2^{(s)}$ and $\hat{\boldsymbol{\beta}}_{ave}^{(s)}$.

Corollary 7 *Suppose that conditions (C1), (C4), (E1), (E4), (E5), and (E7)-(E9) hold. For any $1 \le s \le m$, assume $\lambda_s = c_{3s}'\sqrt{\log p/N_s}$ and $\gamma_s = c_{4s}'\sqrt{\log p/N_s}$, where c_{3s}' and c_{4s}' could be any constants which belong to $[2\tau B_6\sqrt{2(a_0'+1)/a_1}, a_5']$, and a_5' could be any constant no less than $2\tau B_6\sqrt{2(a_0'+1)/a_1}$. If $\tau \ge g_6$ and*

$$\max_{1\le s\le m-1} \tilde{d}_1^2 a_6'^{2s-2} N_s^{\alpha_1/2-1/2} s M_\tau'^s \le A_1',$$

where

$$a_6' = \max\{(2B_5 + 3a_5'/2)/\min\{B_3/3, g_7/2\}, 8 + 2B_2/\{\tau B_6\sqrt{2(a_0'+1)/a_1}\}\},$$
$$M_\tau' = [\max\{\sqrt{32B_6^4(a_0'+2)/a_4'}, 8B_6^2(a_0'+2)/a_4'\} + 4\sqrt{2}L_\tau B_6^3 + 1]a_6'\tilde{d}_1,$$

A_1' could be any constant, $\tilde{d}_1 = \max\{3a_5'/g_7, 4\}$, and g_6 and g_7 are two positive constants depending on e_1, B_2, B_3, B_4 and B_5. Then for any $1 \le s \le m$, we have that with probability at least $1 - 4(s-1)p^{-a_0'} - \sum_{j=1}^{s}\{\exp(-g_8 n_j - g_5 \log p) + 2ep^{-a_0' N_j/n_j}\}$,

$$||\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}_\tau^*||_2 \le \tilde{d}_s\sqrt{\frac{s_0 \log p}{N_s}}, \quad ||\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}_\tau^*||_1 \le \tilde{d}_s^2 s_0\sqrt{\frac{\log p}{N_s}},$$
$$||\hat{\boldsymbol{\beta}}_2^{(s)} - \boldsymbol{\beta}_\tau^*||_2 \le \tilde{d}_s\sqrt{\frac{s_0 \log p}{N_s}}, \quad ||\hat{\boldsymbol{\beta}}_2^{(s)} - \boldsymbol{\beta}_\tau^*||_1 \le \tilde{d}_s^2 s_0\sqrt{\frac{\log p}{N_s}},$$
$$||\hat{\boldsymbol{\beta}}_{ave}^{(s)} - \boldsymbol{\beta}_\tau^*||_2 \le \tilde{d}_s\sqrt{\frac{s_0 \log p}{N_s}}, \quad \textit{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(s)} - \boldsymbol{\beta}_\tau^*||_1 \le \tilde{d}_s^2 s_0\sqrt{\frac{\log p}{N_s}},$$

where g_8 is a positive constant depending on e_1, B_2, B_3, B_4 and B_5 and $\tilde{d}_s = \tilde{d}_1 a_6'^{s-1}$.

The following conditions are required for the asymptotic normality of $\hat{\beta}_l^{da(s)}$ in the case of sub-Gaussian predictor.

(E10) There exist a constant G_1' such that for any $\tau \ge e_1$ and $1 \le l \le p$,

$$\sigma_{\tau,l}^2 \ge G_1'.$$

(E11) For any $\tau \ge e_1$,

$$\max_{1\le j\le p}\sum_{k=1}^{p}|\Omega_{\tau,k,j}|^{\omega} \le v(p),$$

where $\Omega_{\tau,k,j}$ is the (k,j)th element of $\boldsymbol{\Omega}_\tau$.

(E12) For any $\tau \ge e_1$ and $1 \le s \le m$,

$$\{s^2 M_\tau^{2s} s_0\}^{(1-\omega)/2}\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty}^{2(1-\omega)} a_3'^{2s-2} s_0 (\log p)^{1-\omega/2} v(p) N_s^{\omega/2-1/2} = o(1),$$
$$\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty} a_3'^{2s-2} d_1'^2 N_s^{\alpha_1/2-1/2} s\sqrt{\log p} M_\tau^s \le A_1',$$
$$\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty} a_3'^{s-1} s_0^{1/2} N_s^{-1/2}\log p = o(1),$$
$$\{s^2 M_\tau^{2s} s_0\}^{(1-\omega)/2}\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty}^{2(1-\omega)} s v(p)(\log p)^{1-\omega/2} a_3'^{2s-2} d_1'^2 N_s^{\alpha_1/2+\omega/2-1} M_\tau^s \le A_1',$$

and

$$\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty} a_3'^{2s-2} d_1'^2 s N_s^{\alpha_1-1/2} \le A_1'.$$

Condition (E10) implies condition (D1). Condition (E11) is analogous to the uniformity class of matrices assumption in Cai et al. (2011). This condition is for deriving the upper bound of $\max\{\|\hat{\boldsymbol{\Omega}}_1^{(s)} - \boldsymbol{\Omega}_\tau\|_{\infty,\infty}, \|\hat{\boldsymbol{\Omega}}_2^{(s)} - \boldsymbol{\Omega}_\tau\|_{\infty,\infty}\}$. When $\boldsymbol{H}_\tau = \left(\rho^{-|k_1-k_2|}\right)_{1\le k_1,k_2\le p}$, then $v(p) = O(1)$, where ρ could be any constant which belongs to $(0,1)$. Condition (E12) leads to conditions (D3) and (D4). In condition (E12), $g(s, s_0) = s^2 M_\tau^{2s} s_0$. Moreover, this condition is satisfied when $\max\{s, \log s_0, \log(\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty}), \log\log p, \log\{v(p)\}\} = o(\log(N_s))$. Given the more complex data structure in this study compared to that in Han et al. (2023), condition (E12) in our work is inherently more intricate than the second assumption in condition (C8) of Han et al. (2023). The following corollary provides the asymptotic distribution of $\sqrt{N_s}(\hat{\beta}_l^{da(s)} - \beta_{\tau,l}^*)$ with the sub-Gaussian predictor scenario, where $\beta_{\tau,l}^*$ is the lth element of $\boldsymbol{\beta}_\tau^*$.

Corollary 8 *Under the same conditions of Corollary 6, suppose in addition that conditions (E10)-(E12) are satisfied and for any* $1 \leq s \leq m$, $h_s = c'_{5s} s M_\tau^s s_0^{1/2} \|\boldsymbol{\Omega}_\tau\|_{\infty,\infty} \sqrt{\log p / N_s}$ *and* $\kappa_s = c'_{6s} s M_\tau^s s_0^{1/2} \|\boldsymbol{\Omega}_\tau\|_{\infty,\infty} \sqrt{\log p / N_s}$*, where* c'_{5s} *and* c'_{6s} *could be any constants no less than* 1*. Then for any* $1 \leq s \leq m$ *and* $1 \leq l \leq p$*, we have* $\sigma_{\tau,l}^{-1} \sqrt{N_s} (\hat{\beta}_l^{da(s)} - \beta^*_{\tau,l})$ *converges to a standard normal random variable in distribution as* $p \to \infty$*.*

By replacing the positive numbers M_τ, a'_3 and d'_1 with M'_τ, a'_6 and $\tilde{d}_1$ in condition (E12), we get the following condition (E13) for the asymptotic normality of $\hat{\beta}_l^{da(s)}$ under the Gaussian predictor case.

(E13) For any $\tau \geq e_1$ and $1 \leq s \leq m$,

$$\{s^2 M'^{2s}_\tau s_0\}^{(1-\omega)/2} \|\boldsymbol{\Omega}_\tau\|_{\infty,\infty}^{2(1-\omega)} a'^{2s-2}_6 s_0 (\log p)^{1-\omega/2} v(p) N_s^{\omega/2-1/2} = o(1),$$
$$\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty} a'^{2s-2}_6 \tilde{d}_1^2 N_s^{\alpha_1/2-1/2} s \sqrt{\log p} M'^s_\tau \leq A'_1,$$
$$\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty} a'^{s-1}_6 s_0^{1/2} N_s^{-1/2} \log p = o(1),$$
$$\{s^2 M'^{2s}_\tau s_0\}^{(1-\omega)/2} \|\boldsymbol{\Omega}_\tau\|_{\infty,\infty}^{2(1-\omega)} s v(p) (\log p)^{1-\omega/2} a'^{2s-2}_6 \tilde{d}_1^2 N_s^{\alpha_1/2+\omega/2-1} M'^s_\tau \leq A'_1,$$

and

$$\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty} a'^{2s-2}_6 \tilde{d}_1^2 s N_s^{\alpha_1-1/2} \leq A'_1.$$

Under the Gaussian predictor scenario, we also establish the corresponding asymptotic distribution of $\sqrt{N_s} (\hat{\beta}_l^{da(s)} - \beta^*_{\tau,l})$ in the next corollary.

Corollary 9 *Under the same conditions of Corollary 7, suppose in addition that conditions (E10), (E11) and (E13) hold and for any* $1 \leq s \leq m$, $h_s = c'_{7s} s M'^s_\tau s_0^{1/2} \|\boldsymbol{\Omega}_\tau\|_{\infty,\infty} \sqrt{\log p / N_s}$ *and* $\kappa_s = c'_{8s} s M'^s_\tau s_0^{1/2} \|\boldsymbol{\Omega}_\tau\|_{\infty,\infty} \sqrt{\log p / N_s}$*, where* c'_{7s} *and* c'_{8s} *could be any constants no less than* 1*. Then for any* $1 \leq s \leq m$ *and* $1 \leq l \leq p$*, we have* $\sigma_{\tau,l}^{-1} \sqrt{N_s} (\hat{\beta}_l^{da(s)} - \beta^*_{\tau,l})$ *converges to a standard normal random variable in distribution as* $p \to \infty$*.*

3.2 Logistic Loss

If Y is a binary outcomes that takes only the value 0 or 1, the logistic regression model is widely used in finance, business, computer science, and genetics (Hosmer Jr et al., 2013; Sur and Candès, 2019; Ma et al., 2021). In this example, we consider the following negative log-likelihood as the loss function:

$$l(Y, \boldsymbol{X}^\top \boldsymbol{\beta}) = \log\{1 + \exp(\boldsymbol{X}^\top \boldsymbol{\beta})\} - Y \boldsymbol{X}^\top \boldsymbol{\beta}.$$

We then have

$$\boldsymbol{H} = \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} E\{l(Y - \boldsymbol{X}^\top\boldsymbol{\beta})\}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = E[\boldsymbol{X}\boldsymbol{X}^\top \frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta})}{\{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta})\}^2}],$$

$$\boldsymbol{H}_1^{(s)} = \frac{2}{n_s}\sum_{i=1}^{n_s/2} \boldsymbol{X}_i^{(s)}\boldsymbol{X}_i^{(s)\top} \frac{\exp(\boldsymbol{X}_i^{(s)\top}\hat{\boldsymbol{\beta}}_2^{(s)})}{\{1+\exp(\boldsymbol{X}_i^{(s)\top}\hat{\boldsymbol{\beta}}_2^{(s)})\}^2},$$

and

$$\boldsymbol{H}_2^{(s)} = \frac{2}{n_s}\sum_{i=n_s/2+1}^{n_s} \boldsymbol{X}_i^{(s)}\boldsymbol{X}_i^{(s)\top} \frac{\exp(\boldsymbol{X}_i^{(s)\top}\hat{\boldsymbol{\beta}}_1^{(s)})}{\{1+\exp(\boldsymbol{X}_i^{(s)\top}\hat{\boldsymbol{\beta}}_1^{(s)})\}^2}, \quad s = 1, \cdots, m.$$

We first consider the sub-Gaussian predictor case. An additional condition is required for Corollary 10.

(E14) $2^s s_0\sqrt{\log p/N_s} = o(1)$ for $1 \le s \le m$. There exists a positive number a_0'' such that $m = o(\min(p^{a_0''}, p^{g_1'}))$, where g_1' is a positive number depending on M_2, B_1 and B_4.

Based on Lemma 15 and the proof of Corollary 10 below in the Appendix B, we can obtain $P_s(n_1, \ldots, n_s, p) = 4sp^{-a_0''} + \sum_{j=1}^{s}\{\exp(-g_3'n_j - g_1'\log p) + 2ep^{-a_0''N_j/n_j}\}$ and $P(n_s, p) = \exp(-g_3'n_s - g_1'\log p)$. This indicates that condition (C8) is satisfied under condition (E14). As outlined in Subsection 3.1, the data structure in this research is more complicated than that in Negahban et al. (2010). As a result, the assumptions for $n_s(N_s)$, p, s and s_0 (i.e., conditions (C4),(E4), and (E14)) related to the following Corollary 10 in the case of sub-Gaussian predictor. We then obtain that the consistency of $\hat{\boldsymbol{\beta}}_1^{(s)}$, $\hat{\boldsymbol{\beta}}_2^{(s)}$ and $\hat{\boldsymbol{\beta}}_{ave}^{(s)}$, is more complex than that in Corollary 5 of Negahban et al. (2010).

Corollary 10 *Assume that conditions (C1), (C4), (C5), (E2), (E4) and (E14) are satisfied. For any $1 \le s \le m$, assume $\lambda_s = c_{1s}''\sqrt{\log p/N_s}$ and $\gamma_s = c_{2s}''\sqrt{\log p/N_s}$, where c_{1s}'' and c_{2s}'' could be any constants which belong to $[2B_1\sqrt{2(a_0''+1)/a_1}, a_2'']$, and a_2'' could be any constant no less than $2B_1\sqrt{2(a_0''+1)/a_1}$. Suppose in addition that*

$$\max_{1\le s\le m-1} a_3''^{2s-2} d_1''^2 N_s^{\alpha_1/2-1/2} s\tilde{M}^s \le A_1'',$$

where

$$a_3'' = \max\{(2M_3 + 3a_2''/2)/\min\{M_2/3, g_2'/2\}, 8 + 2M_3/\{B_1\sqrt{2(a_0''+1)/a_1}\}\},$$
$$\tilde{M} = [\max\{\sqrt{32B_1^4(a_0''+2)/a_4'}, 8B_1^2(a_0''+2)/a_4'\} + 4\sqrt{2}B_1^3 + 1]a_3''d_1'',$$

A_1'' could be any constant, $d_1'' = \max\{3a_2''/g_2', 4\}$, and g_2' is a positive constant depending on M_2, M_3, B_1, and B_4. Then for any $1 \le s \le m$, we have that with probability at least

$1-4(s-1)p^{-a_0''}-\sum_{j=1}^{s}\{\exp(-g_3'n_j-g_1'\log p)+2ep^{-a_0''N_j/n_j}\}$,

$$||\hat{\boldsymbol{\beta}}_1^{(s)}-\boldsymbol{\beta}^*||_2\le d_s''\sqrt{\frac{s_0\log p}{N_s}},\quad ||\hat{\boldsymbol{\beta}}_1^{(s)}-\boldsymbol{\beta}^*||_1\le d_s''^2s_0\sqrt{\frac{\log p}{N_s}},$$
$$||\hat{\boldsymbol{\beta}}_2^{(s)}-\boldsymbol{\beta}^*||_2\le d_s''^2\sqrt{\frac{s_0\log p}{N_s}},\quad ||\hat{\boldsymbol{\beta}}_2^{(s)}-\boldsymbol{\beta}^*||_1\le d_s''^2s_0\sqrt{\frac{\log p}{N_s}},$$
$$||\hat{\boldsymbol{\beta}}_{ave}^{(s)}-\boldsymbol{\beta}_\tau^*||_2\le d_s''\sqrt{\frac{s_0\log p}{N_s}},\quad \textit{and}\quad ||\hat{\boldsymbol{\beta}}_{ave}^{(s)}-\boldsymbol{\beta}^*||_1\le d_s''^2s_0\sqrt{\frac{\log p}{N_s}},$$

where g_3' is a positive constant depending on M_2, M_3, B_1 and B_4, and $d_s''=a_3''^{s-1}d_1''$, .

By applying the condition $\max_{1\le s\le m-1}a_3''^{2s-2}d_1''^2N_s^{\alpha_1/2-1/2}s\tilde{M}^s\le A_1''$ and Corollary 10, we have that the ℓ_1 and ℓ_2 norms of the difference between the estimators in Corollary 10 and β^* are of orders $\sqrt{s_0^2\log p/(\tilde{M}^s sN_s^{\alpha_1/2+1/2})}$ and $\sqrt{s_0\log p/(\tilde{M}^{s/2}\sqrt{s}N_s^{\alpha_1/4+3/4})}$, respectively. Under the Gaussian predictor case, since condition (E7) implies $\sup_{\|\boldsymbol{\Delta}\|_2=1}\|\boldsymbol{H}^{1/2}\boldsymbol{\Delta}\|_2^2\le B_5$ and $\|\boldsymbol{X}\|_{\psi_2}\le B_6$, we can replace conditions (C5) and (E2) with (E7) and the following (E15).

(E15) There exists a positive constant M_2 such that

$$\inf_{\|\boldsymbol{\Delta}\|_2=1}\|\boldsymbol{H}^{1/2}\boldsymbol{\Delta}\|_2^2\ge M_2.$$

(E16) $2^ss_0\sqrt{\log p/N_s}=o(1)$ for $1\le s\le m$. There exists a positive number a_0' such that $m=o(\min(p^{a_0'},p^{g_4'}))$, where g_4' is a positive number depending on M_2, B_4 and B_5.

Condition (E16) is similar to condition (E14). The following Corollary 11 also establishes the consistency of $\hat{\boldsymbol{\beta}}_1^{(s)}$, $\hat{\boldsymbol{\beta}}_2^{(s)}$ and $\hat{\boldsymbol{\beta}}_{ave}^{(s)}$ with Gaussian predictor scenario.

Corollary 11 *Assume that conditions (C1), (C4), (E4), (E7), (E15) and (E16) are satisfied. For any $1\le s\le m$, assume $\lambda_s=c_{3s}''\sqrt{\log p/N_s}$ and $\gamma_s=c_{4s}''\sqrt{\log p/N_s}$, where c_{3s}'' and c_{4s}'' could be any constants which belong to $[2B_6\sqrt{2(a_0''+1)/a_1},a_4'']$, and a_4'' could be any constant no less than $2B_6\sqrt{2(a_0''+1)/a_1}$. Suppose in addition that*

$$\max_{1\le s\le m-1}a_5''^{2s-2}\tilde{d}_1''^2N_s^{\alpha_1/2-1/2}s\tilde{M}'^s\le A_1'',$$

where

$$a_5''=\max\{(2B_5+3a_4''/2)/\min\{M_2/3,g_5'/2\},8+2B_5/\{B_6\sqrt{2(a_0''+1)/a_1}\}\},$$
$$\tilde{M}'=[\max\{\sqrt{32B_6^4(a_0''+2)/a_4'},8B_6^2(a_0''+2)/a_4'\}+4\sqrt{2}B_6^3+1]a_5''\tilde{d}_1'',$$

A_1'' could be any constant, $\tilde{d}_1''=\max\{3a_4''/g_5',4\}$, and g_5' is a positive constant depending on M_2, B_4, and B_5. Then for any $1\le s\le m$, we have that with probability at least

$1 - 4(s-1)p^{-a_0''} - \sum_{j=1}^{s}\{\exp(-g_6' n_j - g_4' \log p) + 2ep^{-a_0'' N_j/n_j}\}$,

$$||\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}^*||_2 \le \tilde{d}_s''\sqrt{\frac{s_0 \log p}{N_s}}, \quad ||\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}^*||_1 \le \tilde{d}_s''^2 s_0 \sqrt{\frac{\log p}{N_s}},$$
$$||\hat{\boldsymbol{\beta}}_2^{(s)} - \boldsymbol{\beta}^*||_2 \le \tilde{d}_s''^2\sqrt{\frac{s_0 \log p}{N_s}}, \quad ||\hat{\boldsymbol{\beta}}_2^{(s)} - \boldsymbol{\beta}^*||_1 \le \tilde{d}_s''^2 s_0 \sqrt{\frac{\log p}{N_s}},$$
$$||\hat{\boldsymbol{\beta}}_{ave}^{(s)} - \boldsymbol{\beta}_\tau^*||_2 \le \tilde{d}_s''\sqrt{\frac{s_0 \log p}{N_s}}, \quad \textit{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(s)} - \boldsymbol{\beta}^*||_1 \le \tilde{d}_s''^2 s_0 \sqrt{\frac{\log p}{N_s}},$$

where g_6' is a positive constant depending on M_2, B_4, and B_5, and $\tilde{d}_s'' = a_5''^{s-1}\tilde{d}_1''$.

Two additional conditions are needed to prove the asymptotic normality of $\hat{\beta}_l^{da(s)}$ in the case of sub-Gaussian predictor.

(E17) $\max_{1\le j\le p}\sum_{k=1}^{p}|\Omega_{k,j}|^{\omega} \le v(p)$.
(E18) For any $1 \le s \le m$,

$$\{s^2\tilde{M}^{2s}s_0\}^{(1-\omega)/2}\|\boldsymbol{\Omega}\|_{\infty,\infty}^{2(1-\omega)}a_3''^{2s-2}s_0(\log p)^{1-\omega/2}v(p)N_s^{\omega/2-1/2} = o(1),$$
$$\|\boldsymbol{\Omega}\|_{\infty,\infty}a_3''^{2s-2}d_1''^2 N_s^{\alpha_1/2-1/2}s\sqrt{\log p}\tilde{M}^s \le A_1'',$$
$$\|\boldsymbol{\Omega}\|_{\infty,\infty}a_3''^{s-1}s_0^{1/2}N_s^{-1/2}\log p = o(1),$$
$$\{s^2\tilde{M}^{2s}s_0\}^{(1-\omega)/2}\|\boldsymbol{\Omega}\|_{\infty,\infty}^{2(1-\omega)}sv(p)(\log p)^{1-\omega/2}a_3''^{2s-2}d_1''^2 N_s^{\alpha_1/2+\omega/2-1}\tilde{M}^s \le A_1'',$$

and

$$\|\boldsymbol{\Omega}\|_{\infty,\infty}a_3''^{2s-2}d_1''^2 s N_s^{\alpha_1-1/2} \le A_1''.$$

Conditions (E17) and (E18) are similar to conditions (E11) and (E12). In the case of $\boldsymbol{H} = \left(\rho^{-|k_1-k_2|}\right)_{1\le k_1,k_2\le p}$, $v(p) = O(1)$, where ρ could be any constant which belongs to $(0,1)$. Furthermore, in condition (E18), $g(s,s_0) = s^2\tilde{M}^{2s}s_0$. This condition is met if $\max\{s, \log s_0, \log(\|\boldsymbol{\Omega}\|_{\infty,\infty}), \log\log p, \log\{v(p)\}\} = o(\log(N_s))$. Additionally, due to the complex data structure in our study, condition (E18) presents more intricacies compared to condition (C8) in van de Geer et al. (2014). The following corollary 12 demonstrates the asymptotic properties of $\sqrt{N_s}(\hat{\beta}_l^{da(s)} - \beta_l^*)$ with sub-Gaussian predictor scenario.

Corollary 12 *Under the conditions of Corollary 10, suppose that conditions (D1), (E17) and (E18) are satisfied and for any $1 \le s \le m$, $h_s = c_{5s}'' s\tilde{M}^s s_0^{1/2}\|\boldsymbol{\Omega}\|_{\infty,\infty}\sqrt{\log p/N_s}$ and $\kappa_s = c_{6s}'' s\tilde{M}^s s_0^{1/2}\|\boldsymbol{\Omega}\|_{\infty,\infty}\sqrt{\log p/N_s}$, where c_{5s}'' and c_{6s}'' could be any constants no less than 1. Then for any $1 \le s \le m$ and $1 \le l \le p$, we have that $\sigma_l^{-1}\sqrt{N_s}(\hat{\beta}_l^{da(s)} - \beta_l^*)$ converges to a standard normal random variable in distribution as $p \to \infty$.*

By replacing $\tilde{M}$, a_3'' and d_1'' with $\tilde{M}'$, a_5'' and $\tilde{d}_1''$ in condition (E18), we obtain the following condition (E19) for the asymptotic normality of $\hat{\beta}_l^{da(s)}$ in the case of Gaussian predictor.

(E19) For any $1 \leq s \leq m$,

$$\{s^2 \tilde{M}'^{2s} s_0\}^{(1-\omega)/2} \|\mathbf{\Omega}\|_{\infty,\infty}^{2(1-\omega)} a_5''^{2s-2} s_0 (\log p)^{1-\omega/2} v(p) N_s^{\omega/2-1/2} = o(1),$$
$$\|\mathbf{\Omega}\|_{\infty,\infty} a_5''^{2s-2} \tilde{d}_1''^2 N_s^{\alpha_1/2-1/2} s \sqrt{\log p} \tilde{M}'^s \leq A_1'',$$
$$\|\mathbf{\Omega}\|_{\infty,\infty} a_5''^{s-1} s_0^{1/2} N_s^{-1/2} \log p = o(1),$$
$$\{s^2 \tilde{M}'^{2s} s_0\}^{(1-\omega)/2} \|\mathbf{\Omega}\|_{\infty,\infty}^{2(1-\omega)} s v(p) (\log p)^{1-\omega/2} a_5''^{2s-2} \tilde{d}_1''^2 N_s^{\alpha_1/2+\omega/2-1} \tilde{M}'^s \leq A_1'',$$

and

$$\|\mathbf{\Omega}\|_{\infty,\infty} a_5''^{2s-2} \tilde{d}_1''^2 s N_s^{\alpha_1-1/2} \leq A_1''.$$

Similarly, the following corollary 13 also provides the asymptotic properties of $\sqrt{N_s}(\hat{\beta}_l^{da(s)} - \beta_l^*)$ with Gaussian predictor scenario.

Corollary 13 *Under the conditions of Corollary 11, suppose that conditions (D1), (E17) and (E19) are satisfied and for any $1 \leq s \leq m$, $h_s = c_{7s}'' s \tilde{M}'^s s_0^{1/2} \|\mathbf{\Omega}\|_{\infty,\infty} \sqrt{\log p / N_s}$ and $\kappa_s = c_{8s}'' s \tilde{M}'^s s_0^{1/2} \|\mathbf{\Omega}\|_{\infty,\infty} \sqrt{\log p / N_s}$, where c_{7s}'' and c_{8s}'' could be any constants no less than 1. Then for any $1 \leq s \leq m$ and $1 \leq l \leq p$, we have that $\sigma_l^{-1} \sqrt{N_s}(\hat{\beta}_l^{da(s)} - \beta_l^*)$ converges to a standard normal random variable in distribution as $p \to \infty$.*

4. Simulation Studies

In this section, we conduct extensive simulation studies to examine the finite-sample performance of the proposed online lasso and debiasing procedures.

4.1 Evaluation of the Online Consistent Estimation

In this subsection, we first investigate the performance of the proposed online lasso method and randomly generate a total of N_m samples that arrive in a sequence of m data batches, denoted by $\{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$, from the following two examples with the continuous and discrete outcome described in Section 3:

Model 1: $Y_i^{(j)} = 3\boldsymbol{X}_i^{(j)\top} \boldsymbol{\beta}_0 + 10 \sin(\boldsymbol{X}_i^{(j)\top} \boldsymbol{\beta}_0) + \epsilon_i^{(j)}, i = 1, \ldots, n_j,\ j = 1, \ldots, m,$ where $\boldsymbol{X}_i^{(j)}$ is generated from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with covariance matrix $\mathbf{\Sigma} = \left(2^{-|k_1-k_2|}\right)_{1 \leq k_1, k_2 \leq p}$, and the true parameter $\boldsymbol{\beta}_0 = \tilde{\boldsymbol{\beta}} / \|\mathbf{\Sigma}^{1/2} \tilde{\boldsymbol{\beta}}\|_2$ with

$$\tilde{\beta}_l = \begin{cases} s_0 + 1 - l, & \text{for } 1 \leq l \leq s_0, \\ 0, & \text{for } s_0 + 1 \leq l \leq p. \end{cases}$$

The random error $\epsilon_i^{(j)}$ is generated from four types of distributions: (i) standard normal distribution, denoted as $\mathcal{N}(0,1)$; (ii) log-normal distribution with the log location parameter 0 and log shape parameter 1, denoted as $\mathrm{LN}(0,1)$; (iii) Student's t-distribution with 3 degrees of freedom, denoted as $t(3)$; (iv) Weibull distribution with shape parameter 0.5 and scale parameter 0.5, denoted as Weibull(0.5; 0.5).

Model 2: $\Pr(Y_i^{(j)} \mid \boldsymbol{X}_i^{(j)}) = \frac{\exp\{\boldsymbol{X}_i^{(j)\top} \boldsymbol{\beta}_0 + \sin(\boldsymbol{X}_i^{(j)\top}) \boldsymbol{\beta}_0\}}{1 + \exp\{\boldsymbol{X}_i^{(j)\top} \boldsymbol{\beta}_0 + \sin(\boldsymbol{X}_i^{(j)\top} \boldsymbol{\beta}_0)\}}, i = 1, \ldots, n_j,\ j = 1, \ldots, m,$ where $\boldsymbol{X}_i^{(j)}$ is generated from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with the same true parameter $\boldsymbol{\beta}_0$ as in Model 1. For the design matrix, we consider two scenarios: (i) $\boldsymbol{\Sigma}$ is Toeplitz with $\Sigma_{k_1,k_2} = 0.5^{|k_1-k_2|}$; (ii) $\boldsymbol{\Sigma} = \boldsymbol{I}$. For each type of models, we consider the following combinations of (N_m, m, n_j, p, s_0), $j = 1, \ldots, m$: (i) $(N_m, m, n_j, p, s_0) = (1600, 16, 100, 200, 5)$; (ii) $(N_m, m, n_j, p, s_0) = (3200, 16, 200, 400, 10)$.

For comparison, we also consider the following methods: (i) the proposed online lasso estimator at several intermediate points for $s = 1, \ldots, m$, denoted by "online"; (ii) the offline lasso estimator at the terminal time point m, denoted by "offline"; (iii) the offline lasso estimator with final data batch $\mathcal{D}_m$, denoted by "final". To measure the estimation accuracy, we calculate the sine distance between the estimator $\hat{\beta}_\tau$ and true parameter β_0 defined as follows:

$$\sin\theta\left(\hat{\beta}_\tau, \beta_0\right) = 1 - \frac{< \hat{\beta}_\tau, \beta_0 >}{\|\hat{\beta}_\tau\|_2 \, \|\beta_0\|_2},$$

where $< a, b >$ is the inner product of vectors a and b. Here, we report the sine distance instead of $\|\hat{\beta}_\tau - c_\tau\beta_0\|_2$ for all simulation configurations. As c_τ may take different values under different models and different settings, the sine distance is free of c_τ.

The tuning parameters λ_s and γ_s, $s = 1, \ldots, m$, are chosen by the modified BIC (Wang et al., 2007). For example, we obtain λ_s by minimizing

$$\begin{aligned}\mathrm{BIC}(\lambda_s) = &\log\Bigg[(\hat{\boldsymbol{\beta}}(\lambda_s) - \hat{\boldsymbol{\beta}}_2^{(s-1)})^\top \sum_{j=1}^{s-1} \frac{n_j}{2N_s} \boldsymbol{H}_1^{(j)} (\hat{\boldsymbol{\beta}}(\lambda_s) - \hat{\boldsymbol{\beta}}_2^{(s-1)}) \\ &+ \frac{2}{N_s} \sum_{i=1}^{n_s/2} l(Y_i^{(s)}, \boldsymbol{X}_i^{(s)\top} \hat{\boldsymbol{\beta}}(\lambda_s))\Bigg] + C_{N_s} \frac{\log(N_s/2)}{N_s/2} \big\|\hat{\boldsymbol{\beta}}(\lambda_s)\big\|_0,\end{aligned}$$

where $\hat{\boldsymbol{\beta}}(\lambda_s)$ is obtained from (5), $C_{N_s} = c \log\log(p)$, c is a constant, and $\|\cdot\|_0$ denotes the number of nonzero elements in a vector. Furthermore, we choose the robustification parameter τ in the Huber loss such that 80% of the prediction errors are in $[-\tau, \tau]$.

Table 1 summarizes the results for Models 1 and 2 averaged over 200 replications. We can see that, as the number of data batches s increases, i.e. the sample size grows, the sine distance associated with the proposed online lasso estimator decreases rapidly. To illustrate this, for the continuous response in Model 1 with $(N_m, m, n_j, p, s_0) = (1600, 16, 100, 200, 5)$ and the random error following the standard normal distribution $N(0, 1)$, the sine distance drops from 0.031 to 0.002 as the batch index s increases from 4 to 16. The analogous results are observed for the binary response in Model 2. As expected, these findings validate the estimation consistency of our proposed online lasso method. Meanwhile, the sine distance of the proposed online estimator closely matches that of the offline benchmark, which uses the full data set. This suggests that the proposed online method effectively captures key information despite relying primarily on summary statistics from historical batches. Moreover, the performance of the proposed online method employing the Huber loss is comparable to that using the least squares (LS) loss with continuous responses across various types of error term. In particular, when the error terms follow heavy-tailed distributions, the Huber loss is proved to be considerably more robust and is thus preferred. In comparison to the lasso estimator, which utilizes only the data from the final batch without retaining information from earlier batches, our proposed method achieves a significantly reduced sine distance. This reduction underscores the superior effectiveness of the proposed online approach. More generally, the proposed method consistently demonstrates a notably low sine distance across all scenarios, affirming its strong and reliable performance.

Table 1: The sine distance under different settings in Section 4.1 are summarized.

		online				offline	final
Model	Batch index s	4	8	12	16		
	$(N_m, m, n_j, p, s_0) = (1600, 16, 100, 200, 5)$						
	$\mathcal{N}$(0,1)	0.031	0.010	0.004	0.002	0.002	0.025
Model 1	LN(0,1)	0.056	0.020	0.008	0.004	0.004	0.044
Huber	$t(3)$	0.045	0.015	0.006	0.003	0.003	0.037
	Weibull(0.5,0.5)	0.057	0.021	0.008	0.004	0.004	0.042
	$\mathcal{N}$(0,1)	0.030	0.013	0.006	0.004	0.004	0.041
Model 1	LN(0,1)	0.057	0.026	0.013	0.007	0.008	0.071
LS	$t(3)$	0.048	0.022	0.011	0.006	0.008	0.060
	Weibull(0.5,0.5)	0.062	0.029	0.014	0.008	0.009	0.074
	$(N_m, m, n_j, p, s_0) = (3200, 16, 200, 400, 10)$						
	$\mathcal{N}$(0,1)	0.036	0.012	0.005	0.003	0.003	0.029
Model 1	LN(0,1)	0.064	0.023	0.009	0.004	0.005	0.051
Huber	$t(3)$	0.048	0.016	0.006	0.003	0.004	0.040
	Weibull(0.5,0.5)	0.073	0.026	0.009	0.005	0.006	0.057
	$\mathcal{N}$(0,1)	0.035	0.015	0.007	0.004	0.005	0.048
Model 1	LN(0,1)	0.066	0.030	0.015	0.009	0.010	0.081
LS	$t(3)$	0.049	0.021	0.010	0.006	0.007	0.065
	Weibull(0.5,0.5)	0.079	0.037	0.018	0.010	0.012	0.092
	$(N_m, m, n_j, p, s_0) = (1600, 16, 100, 200, 5)$						
Model 2	$\Sigma = I$	0.183	0.083	0.060	0.052	0.038	0.371
logistic	$\Sigma = (0.5^{\lvert k_1-k_2\rvert})$	0.113	0.064	0.052	0.049	0.038	0.340
	$(N_m, m, n_j, p, s_0) = (3200, 16, 200, 400, 10)$						
Model 2	$\Sigma = I$	0.165	0.078	0.057	0.049	0.035	0.339
logistic	$\Sigma = (0.5^{\lvert k_1-k_2\rvert})$	0.117	0.070	0.055	0.048	0.040	0.339

To gain deeper insights into how the upper bounds of the proposed estimator are affected by the number of data batches m, in contrast to the traditional offline lasso estimator, we conduct a series of simulation studies. These studies follow the same setting and data-generating process as described in Model 1, but with different sample sizes. Specifically, we fix the full data sample size $N_m = 2100$ and vary different batch sizes, i.e., $m = 21, 41, 51, 101, 201$. The sample size for the first batch is set to $n_1 = 100$ to ensure a sufficiently large initial sample, while the sample sizes for the remaining batches are evenly distributed according to the total number of batches. Correspondingly, (i) Case 1: $(N_m, m, n_1, n_j, p, s_0) = (2100, 21, 100, 100, 200, 5)$; (ii) Case 2: $(N_m, m, n_1, n_j, p, s_0) = (2100, 41, 100, 50, 200, 5)$; (iii) Case 3: $(N_m, m, n_1, n_j, p, s_0) = (2100, 51, 100, 40, 200, 5)$; (iv) Case 4: $(N_m, m, n_1, n_j, p, s_0) = (2100, 101, 100, 20, 200, 5)$; (v) Case 5: $(N_m, m, n_1, n_j, p, s_0) = (2100, 201, 100, 10, 200, 5)$.

Table 2: The sine distance ($\times 10^{-1}$) under different settings in Section 4.1 for Model 1 with Huber loss are summarized. Note that Q1, Q2, Q3 and Q4 represent the $(1 + m^*/4)$th, $(1 + m^*/2)$th, $(1 + m^*3/4)$th ($m^* = m - 1$) and mth batch, respectively.

			online			
Model	cases	$(m-1, n_j)$	Q1	Q2	Q3	Q4
	1	(20, 100)	0.247	0.061	0.024	0.012
	2	(40, 50)	0.119	0.034	0.019	0.016
$\mathcal{N}$(0,1)	3	(50, 40)	0.109	0.039	0.031	0.032
	4	(100, 20)	0.130	0.094	0.100	0.118
	5	(200, 10)	0.497	0.419	0.514	0.605
	1	(20, 100)	0.452	0.116	0.043	0.020
	2	(40, 50)	0.225	0.059	0.030	0.022
LN(0,1)	3	(50, 40)	0.185	0.051	0.035	0.032
	4	(100, 20)	0.167	0.103	0.103	0.118
	5	(200, 10)	0.513	0.428	0.520	0.610

The detailed simulation results for the sine distance across different quantile batches over 200 replications are presented in Table 2. The following conclusions can be drawn: (1) When the batch size is not large, with an increase in the number of data batches s, i.e., as the sample size grows, the sine distance linked to the proposed online lasso estimator decreases, and consistency is achieved. For example, in Case 1 with normal errors, as s increases from 6 to 21, and the sine distance decreases from 0.0247 to 0.0012. (2) For larger batch sizes, the sine distance initially decreases as s increases but subsequently increases, indicating that while consistency is achieved in the initial batches, it is not consistently maintained in later batches. For instance, in Case 4 with normal errors, as s increases from 26 to 51, the sine distance decreases from 0.0130 to 0.0094. However, as s increases further from 76 to 101, the sine distance rises from 0.0100 to 0.0118. (3) When the full sample size is held constant, increasing the number of data batches m leads to an increase in the sine distance of the proposed online lasso estimator, suggesting that consistency is not maintained when the batch size becomes too large. For example, with normal errors, as

m increases from 21 to 201, the sine distance rises from 0.0012 to 0.0605. In summary, the proposed online lasso estimators remain consistent as long as the number of data batches m does not increase too rapidly.

4.2 Evaluation of the Online Pointwise Inference

In this subsection, we conduct simulations to check the performance of the online debiasing estimator via the null hypothesis $H_{0,l} : \beta_l^* = 0,\ l \in \{1, \ldots, p\}$, which is equivalent to the null hypothesis $H_{0,l} : \beta_{0,l} = 0$. We consider two types of example under the same settings as in Section 4.1 except for the different combinations of (N_m, m, n_j, p, s_0), $j = 1, \ldots, m$: (i) $(N_m, m, n_j, p, s_0) = (1600, 16, 100, 200, 5)$; (ii) $(N_m, m, n_j, p, s_0) = (2400, 12, 200, 400, 10)$.

For comparison, we consider the following methods: (i) the proposed online debiasing estimator at several intermediate points for $s = 1, \ldots, m$, denoted by "online-deb"; (ii) the offline debiasing estimator at the terminal time point m, denoted by "offline-deb"; (iii) the offline debiasing estimator with final data batch $\mathcal{D}_m$, denoted by "final-deb". To evaluate the performance of different methods, we compute the following measurements:
(a) FPR: the average False Positive Rate corresponding to zero coefficients $\beta_l, s_0+1 \leq l \leq p$;
(b) TPR(l): the True Positive Rate corresponding to $\beta_l, 1 \leq l \leq s_0$.

The detailed calculations for the sth batch are given by

$$\begin{aligned}
\text{FPR} =& \text{Average}\Big\{\frac{1}{p - s_0} \sum_{l=s_0+1}^{p} I\big(\sqrt{N_s}|\hat{\beta}_l^{da(s)}|/\hat{\sigma}_{l,s} \geq z_{\alpha/2}\big)\Big\}, \\
\text{TPR}(l) =& \text{Average}\Big\{I\big(\sqrt{N_s}|\hat{\beta}_l^{da(s)}|/\hat{\sigma}_{l,s} \geq z_{\alpha/2}\big)\Big\},
\end{aligned}$$

where "Average" represents the average rate over 200 replications.

The tuning parameters h_s and κ_s, $s = 1, \cdots, m$, are determined as follows. Following Cai et al. (2011), we can use the offline cross-validation scheme to select the tuning parameters h_1 and κ_1 in (12) with only the first data batch $\mathcal{D}_1$. However, it is infeasible for streaming data since we can not access the entire raw data at the same time. Motivated by Tashman (2000) and Han et al. (2021), we adopt the following "rolling-original-recalibration" scheme to select the tuning parameters h_s, κ_s, $s = 1, \ldots, m$. Here, we just present the selection of h_s, the similar idea can be used for κ_s. For $s \geq 2$, we regard the previous cumulative data set $\{\mathcal{D}_1, \ldots, \mathcal{D}_{s-1}\}$ as the training set that trains the estimator $\widehat{\mathbf{\Omega}}_1^{(s-1)}(h)$ for a sequence of h in a candidate set $\mathcal{S}_h$ while the current data batch $\mathcal{D}_s$ is the validation set. Thus, when the data batch $\mathcal{D}_s$ arrives, we select h_s by choosing the smallest likelihood loss on the validation sample as follows:

$$h_s = \underset{h \in \mathcal{S}_h}{\arg\min} \left(\operatorname{tr}\left\{2\boldsymbol{H}_1^{(s)}\widehat{\mathbf{\Omega}}_1^{(s-1)}(h)/n_s\right\} - \log[\det\{\widehat{\mathbf{\Omega}}_1^{(s-1)}(h)\}]\right).$$

For Models 1 and 2 with $(N_m, m, n_j, p, s_0) = (1600, 16, 100, 200, 5)$, Tables 3 and 4 present the FPRs and TPRs the proposed online pointwise tests at a significance level of 0.05 across 200 replications. Similarly, for $(N_m, m, n_j, p, s_0) = (2400, 12, 200, 400, 10)$, Table 6 presents the results for Model 2, while the results for Model 1 with Huber and least squares (LS) losses are summarized in Tables 5 and 7, respectively. The results show that the average FPRs for all zero coefficients consistently remain around 0.05, indicating that the proposed method successfully maintains the nominal level for these coefficients, suggesting the asymptotic normality of the proposed online debiased lasso estimator. For nonzero coefficients, as the number of data batches s increases (and thus the sample size grows), the TPR of the proposed estimator approaches 1. For example, in Model 1 with continuous response and $(N_m, m, n_j, p, s_0) = (2400, 12, 20, 400, 10)$ and random error $N(0, 1)$, the TPR(9) in Table 5 increases from 0.76 to 1 as the batch index s grows from 3 to 12. Similar patterns are observed for the binary response in Model 2. Moreover, the FPRs and TPRs of the proposed online estimator closely align with those of the offline benchmark method, illustrating the effectiveness of our approach in preserving essential information, even when primarily relying on summary statistics from historical batches. Furthermore, the TPRs (or empirical power) of the proposed online method surpass those of the final-deb method, highlighting its superior performance. Overall, the simulation results across various settings confirm the robustness and effectiveness of the proposed method.

5. Real Data Example

5.1 Nasdaq Stock Data

In this subsection, we illustrate the proposed method with the Nasdaq stock data set, which is collected from January 1, 2008 to November 2, 2018. For this data set, the response variable is the return of the Nasdaq 100 index for every three days, and the covariates are $p = 226$ stock returns for every three days during this period. Similar to Lan et al. (2016), our goal in this study is to find the most relevant stocks that can be used to construct a small portfolio, which tracks the return of the Nasdaq 100 index.

To apply our proposed procedure, the data are split into $m = 10$ batches. We take the first two-year data set as the first data batch ($n_1 = 164$) to guarantee a sufficiently large sample size at the initial stage and the next one-year data set as the subsequent data batch ($n_j = 82, j = 2, \cdots, m-1$). In addition, the sample size of the final batch is $n_m = 72$. Hence, the streaming data consists of $m = 10$ data batches with a total sample size $N_m = 892$. Before applying the proposed procedure, we carry out two elliptical tests, i.e., Pseudo-Gaussian test (Cassart et al., 2008) and Skew Optimal test (Babić et al., 2021), for every two principal components of covariates to test roughly the assumption of the linearity of expectation in condition (C1). For the resulting p-values, we consider their mean, standard deviation, and the frequency of p-values that are larger than 0.05. In addition, when the assumption of elliptical distribution is violated and the performance of the elliptical test is unsatisfactory, we apply the coordinatewise Gaussianization (Mai et al., 2023) to transform the original covariates into normal distributions. The associated p-values of the elliptical test for original and transformed covariates are summarized in Table 8. From Table 8, we observe that both tests of the frequency of p-values for transformed covariates of Nasdaq stock data are above 0.7, which is notably higher than 0.4, the frequency of original covariates. This suggests that applying the coordinatewise Gaussianization transformation is more reasonable in this example.

To identify important stocks that are associated with the Nasdaq 100 index, we apply the proposed online procedure to sequentially test the significance of each regression coefficient at a prespecified level $\alpha = 0.05$, i.e., testing $H_{0,l} : \beta_{0,l} = 0$ for $l = 1, \dots, p$. The selection methods of the tuning parameters λ_s, γ_s, h_s, and κ_s, $s = 1, \dots, m$ are the same as those

Table 3: The average True/False positive rates under different settings for Model 1 with $(N_m, m, n_j, p, s_0) = (1600, 16, 100, 200, 5)$ in Section 4.2 are summarized.

		online-deb				offline-deb	final-deb
	Batch index s	4	8	12	16		
	FPR	0.045	0.044	0.050	0.050	0.053	0.050
	TPR(1)	1.000	1.000	1.000	1.000	1.000	1.000
$\mathcal{N}$(0,1)	TPR(2)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(3)	1.000	1.000	1.000	1.000	1.000	1.000
Huber	TPR(4)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(5)	0.965	1.000	1.000	1.000	1.000	0.910
	FPR	0.046	0.045	0.050	0.053	0.053	0.052
	TPR(1)	1.000	1.000	1.000	1.000	1.000	1.000
LN(0,1)	TPR(2)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(3)	1.000	1.000	1.000	1.000	1.000	1.000
Huber	TPR(4)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(5)	0.955	1.000	1.000	1.000	1.000	0.880
	FPR	0.046	0.053	0.054	0.056	0.052	0.052
	TPR(1)	1.000	1.000	1.000	1.000	1.000	1.000
$\mathcal{N}$(0,1)	TPR(2)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(3)	1.000	1.000	1.000	1.000	1.000	0.995
LS	TPR(4)	0.975	1.000	1.000	1.000	1.000	0.995
	TPR(5)	0.475	1.000	1.000	1.000	1.000	0.785
	FPR	0.041	0.047	0.048	0.052	0.053	0.051
	TPR(1)	1.000	1.000	1.000	1.000	1.000	1.000
LN(0,1)	TPR(2)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(3)	0.985	1.000	1.000	1.000	1.000	1.000
LS	TPR(4)	0.885	1.000	1.000	1.000	1.000	0.985
	TPR(5)	0.340	1.000	1.000	1.000	1.000	0.715

in the simulation studies. To ensure the stability of selection in this online framework, the identified stocks are required to be significant at the level of 0.1 for the $m-1$ batch. It is reasonable for financial managers to track the stocks for more time and establish a portfolio cautiously, especially for risk-averse investors. We find that 22 stocks are identified as important stocks at the significance level of 0.05. Correspondingly, the p-values of these regression coefficients over the 10 batches are plotted in Figure 1. From this figure, as we

Table 4: The average True/False positive rates under different settings for Model 2 with $(N_m, m, n_j, p, s_0) = (1600, 16, 100, 200, 5)$ in Section 4.2 are summarized.

		online-deb				offline-deb	final-deb
$\boldsymbol{\Sigma}$	Batch index s	4	8	12	16		
$\boldsymbol{I}$	FPR	0.038	0.047	0.050	0.048	0.048	0.043
	TPR(1)	1.000	1.000	1.000	1.000	1.000	0.990
	TPR(2)	1.000	1.000	1.000	1.000	1.000	0.845
	TPR(3)	0.980	1.000	1.000	1.000	1.000	0.595
	TPR(4)	0.720	0.970	1.000	1.000	1.000	0.315
	TPR(5)	0.225	0.555	0.760	0.850	0.930	0.105
$(0.5^{\lvert k_1-k_2\rvert})$	FPR	0.044	0.047	0.049	0.052	0.048	0.045
	TPR(1)	1.000	1.000	1.000	1.000	1.000	0.990
	TPR(2)	1.000	1.000	1.000	1.000	1.000	0.995
	TPR(3)	0.955	1.000	1.000	1.000	1.000	0.985
	TPR(4)	0.670	0.910	0.985	1.000	0.995	0.700
	TPR(5)	0.250	0.510	0.685	0.780	0.635	0.315

collect data more and more, the most selected stocks are more significant and relatively stable. In addition, we use a Kolmogorov-Smirnov test for the residuals obtained from the proposed SIMs, where the nonparametric function is estimated by the nonparametric local linear kernel method. We also consider the residuals obtained from the linear model based on least squares (LM-LS) and Huber (LM-Huber) losses. The detailed results are presented in Table 9. The p-values of ten batches based on SIMs are all larger than 0.05. Therefore, this example demonstrates that our proposed method can be effectively applied to analyze the stock data set and performs reasonably well.

5.2 Financial Distress Data

In this section, we illustrate our method with the financial distress data set, which is available from https://www.kaggle.com/datasets/shebrahimi/financial-distress. This data set is collected from a sample of companies. Time series varies between 1 to 10 for each company. For this data set, the financial distress index can be regarded as the response variable and other 82 variables are covariates that consist of some financial and non-financial characteristics of the sampled companies. In addition, this data set consists of a total of $N_m = 1008$ observations, and the response and the covariates have been standardized to have zero mean and unit variance. Our goal of this study is to select the variables that significantly affect the companies' financial distress.

In this example, the covariates include 190 interaction terms (products of 20 pairs of the original covariates). As a result, the dimension of the feature vector is $p = 272$. Before

Table 5: The average True/False positive rates under the Huber loss for Model 1 with $(N_m, m, n_j, p, s_0) = (2400, 12, 200, 400, 10)$ in Section 4.2 are summarized.

		online-deb				offline-deb	final-deb
	Batch index s	3	6	9	12		
	FPR	0.046	0.046	0.049	0.050	0.050	0.049
	TPR(1)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(2)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(3)	1.000	1.000	1.000	1.000	1.000	1.000
$\mathcal{N}$(0,1)	TPR(4)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(5)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(6)	1.000	1.000	1.000	1.000	1.000	1.000
Huber	TPR(7)	1.000	1.000	1.000	1.000	1.000	0.990
	TPR(8)	0.975	1.000	1.000	1.000	1.000	0.920
	TPR(9)	0.760	0.925	0.965	1.000	1.000	0.680
	TPR(10)	0.350	0.480	0.680	0.800	0.800	0.370
	FPR	0.046	0.047	0.050	0.050	0.050	0.049
	TPR(1)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(2)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(3)	1.000	1.000	1.000	1.000	1.000	1.000
LN(0,1)	TPR(4)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(5)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(6)	1.000	1.000	1.000	1.000	1.000	1.000
Huber	TPR(7)	1.000	1.000	1.000	1.000	1.000	0.995
	TPR(8)	0.965	1.000	1.000	1.000	1.000	0.900
	TPR(9)	0.700	0.880	0.965	0.995	1.000	0.640
	TPR(10)	0.305	0.480	0.625	0.745	0.750	0.330

applying the proposed procedure, we conduct the same elliptical tests as in Section 5.1. From Table 8, we can see that both tests of the frequency of p-values for the financial distress data are above 0.6. Therefore, we assume that the covariates approximately follow an elliptical distribution. Subsequently, we split the data into $m = 10$ batches randomly, take the $n_1 = 108$ observations as the first batch, and set each of the remaining 9 batches containing $n_j = 100$ observations. To identify the influential variables, we aim to test: $H_{0,l} : \beta_{0,l} = 0$ for $l = 1, \ldots, p$. The tuning parameters λ_s, γ_s, h_s and κ_s, $s = 1, \ldots, m$ are determined by the same methods as described in the simulation studies. Given a prespecified level $\alpha = 0.05$, we observe that 37 variables are significant in the online framework, and the associated p-values of the 10 batches are presented in Figure 2. From this figure, we can find that the most variables are more significant and reach relative stability as more and more data are collected. This example indicates that our proposed method can be applied to analyze the data set with binary outcomes and perform reasonably well.

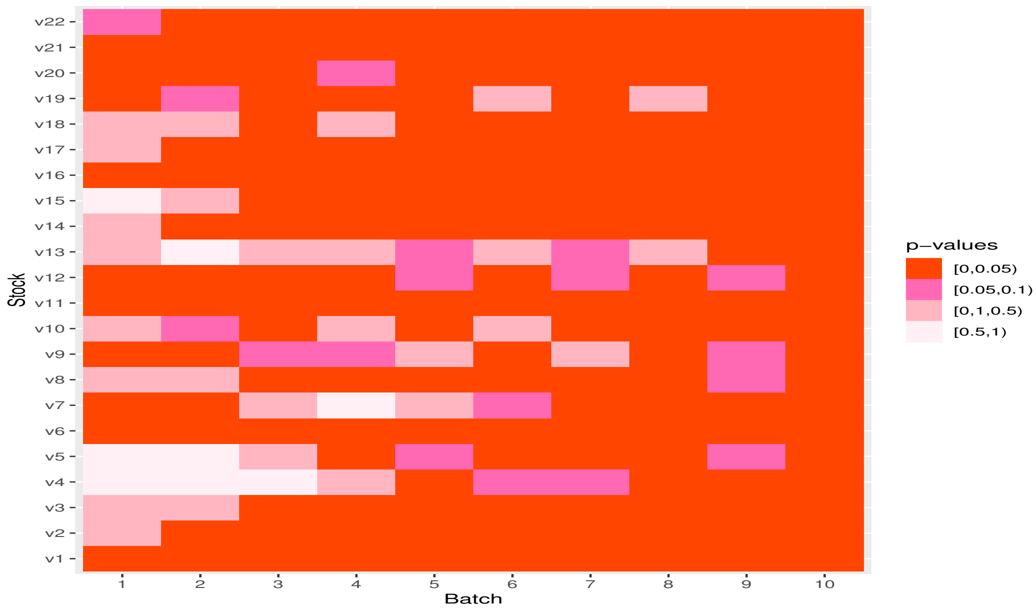

Figure 1: p-values for Nasdaq stock data

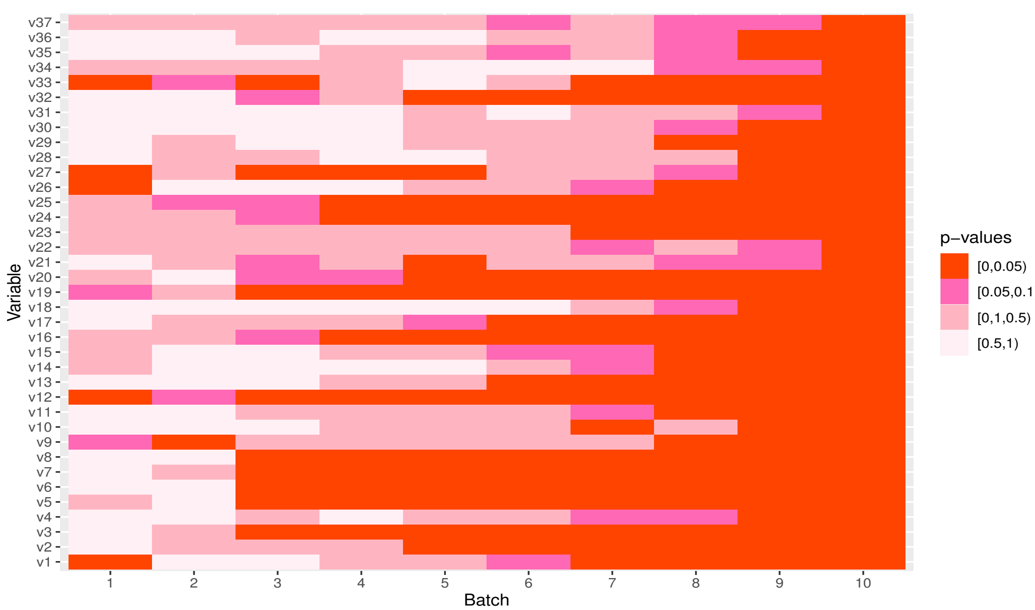

Figure 2: p-values for financial distress data

Table 6: The average True/False positive rates under different settings for Model 2 with $(N_m, m, n_j, p, s_0) = (2400, 12, 200, 400, 10)$ in Section 4.2 are summarized.

$\boldsymbol{\Sigma}$	Batch index s	online-deb 3	6	9	12	offline-deb	final-deb
	FPR	0.041	0.049	0.050	0.050	0.050	0.043
	TPR(1)	1.000	1.000	1.000	1.000	1.000	0.990
	TPR(2)	1.000	1.000	1.000	1.000	1.000	0.975
	TPR(3)	1.000	1.000	1.000	1.000	1.000	0.900
	TPR(4)	0.985	1.000	1.000	1.000	1.000	0.785
$\boldsymbol{I}$	TPR(5)	0.970	1.000	1.000	1.000	1.000	0.745
	TPR(6)	0.850	0.990	1.000	1.000	1.000	0.485
	TPR(7)	0.625	0.975	1.000	1.000	1.000	0.385
	TPR(8)	0.390	0.875	1.000	0.990	1.000	0.190
	TPR(9)	0.250	0.510	0.670	0.710	0.830	0.160
	TPR(10)	0.085	0.016	0.200	0.290	0.375	0.080
	FPR	0.046	0.047	0.048	0.046	0.049	0.046
	TPR(1)	0.995	1.000	1.000	1.000	1.000	0.935
	TPR(2)	0.995	1.000	1.000	1.000	1.000	0.980
	TPR(3)	0.960	0.995	1.000	1.000	1.000	0.980
	TPR(4)	0.945	0.995	1.000	1.000	1.000	0.940
$(0.5^{\lvert k_1-k_2 \rvert})$	TPR(5)	0.875	1.000	1.000	1.000	1.000	0.865
	TPR(6)	0.715	0.945	1.000	1.000	1.000	0.775
	TPR(7)	0.570	0.935	0.995	1.000	1.000	0.605
	TPR(8)	0.395	0.640	0.820	0.930	0.925	0.350
	TPR(9)	0.185	0.345	0.490	0.650	0.675	0.180
	TPR(10)	0.090	0.225	0.305	0.370	0.305	0.125

6. Discussion

In this paper, we studied the statistical inference of SIMs with streaming data under the high-dimensional regime. The proposed procedure was applicable to the streaming data, that is, only depended on the current batch of the data stream with summary statistics from the historical data. In addition, our method was developed for general convex loss functions, which could be effectively used to handle heavy-tailed errors or discrete responses. Meanwhile, we established the ℓ_1 and ℓ_2 bounds of the proposed online lasso estimators and the asymptotic normality of the proposed online debiased lasso estimators. Simulation studies were conducted to show the effectiveness of the proposed method and applications to two real data examples were provided to illustrate our method.

There are several other interesting avenues for the future work. First, the current work relies on the assumption of homogeneous data, that is, the streaming data is assumed to be i.i.d. sampled. It would be an interesting topic to address the problem of non-homogeneous data. Second, we require that the data is completely observed in our framework. It is unclear how to extend the proposed method in the presence of incomplete data such as missing

Table 7: The average True/False positive rates under the least squares (LS) loss for Model 1 with $(N_m, m, n_j, p, s_0) = (2400, 12, 200, 400, 10)$ in Section 4.2 are summarized.

		online-deb				offline-deb	final-deb
	Batch index s	3	6	9	12		
	FPR	0.038	0.046	0.047	0.048	0.051	0.054
	TPR(1)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(2)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(3)	1.000	1.000	1.000	1.000	1.000	1.000
$\mathcal{N}$(0,1)	TPR(4)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(5)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(6)	1.000	1.000	1.000	1.000	1.000	1.000
LS	TPR(7)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(8)	1.000	1.000	1.000	1.000	1.000	0.940
	TPR(9)	0.995	1.000	1.000	1.000	1.000	0.735
	TPR(10)	0.960	1.000	1.000	1.000	1.000	0.310
	FPR	0.043	0.051	0.052	0.052	0.056	0.053
	TPR(1)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(2)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(3)	1.000	1.000	1.000	1.000	1.000	1.000
LN(0,1)	TPR(4)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(5)	1.000	1.000	1.000	1.000	1.000	1.000
	TPR(6)	1.000	1.000	1.000	1.000	1.000	1.000
LS	TPR(7)	1.000	1.000	1.000	1.000	1.000	0.995
	TPR(8)	1.000	1.000	1.000	1.000	1.000	0.890
	TPR(9)	0.990	1.000	1.000	1.000	1.000	0.630
	TPR(10)	0.870	0.970	1.000	1.000	0.980	0.240

Table 8: The elliptical tests for two real data examples. The mean and standard deviation of p-values, and averaged frequency of p-values larger than 0.05 are summarized.

		Original Data			Coordinatewise Gaussianization		
X	Test	mean	sd	Freq	mean	sd	Freq
Nasdaq	Pseudo-Gaussian	0.10301	0.15494	0.45763	0.52709	0.41507	0.71186
stock	SkewOptimal	0.11425	0.17902	0.38983	0.45782	0.33594	0.75424
Financial	Pseudo-Gaussian	0.28603	0.30536	0.69118			
distress	SkewOptimal	0.33737	0.37347	0.60294			

data or censored data. Third, the selection of the parameter τ is crucial for the Huber loss function in real implementation. It is challenging to provide a data-driven selector to determine τ in a streaming manner with theoretical guarantees. We leave space here for

Table 9: The residual test for Nasdaq stock data. The p-values based on single index model (SIM), linear model with the huber loss (LM-Huber) and least squared loss (LM-LS) are summarized for Nasdaq stock data.

	online									
model	1	2	3	4	5	6	7	8	9	10
SIM	0.422	0.853	0.653	0.543	0.785	0.059	0.901	0.671	0.103	0.769
LM-Huber	0.198	0.000	0.232	0.673	0.005	0.002	0.101	0.098	0.184	0.257
LM-LS	0.000	0.019	0.504	0.483	0.286	0.546	0.004	0.645	0.915	0.000

future research. Fourth, we neither prove nor guarantee that the estimators $\hat{\beta}_1^{(s)}, \hat{\beta}_2^{(s)}$ and $\hat{\beta}_{\text{ave}}^{(s)}$ attain the optimal convergence rate. The development of estimators that achieve the optimal convergence rate presents a significant challenge and warrants further investigation. Lastly, very few methods have been developed on the goodness of fit test for the high dimensional SIMs. To the best of our knowledge, the most relevant works are Tan and Zhu (2019) and Tan and Zhu (2022), which accommodate the goodness of fit test for parametric single and multiple index models with continuous responses, respectively. However, both studies focus on parametric models and scenarios with diverging dimensional predictors. It remains an open and challenging problem to conduct the goodness of fit test for high-dimensional SIMs, especially in the online setting. Investigating this would be an interesting and important research problem for a separate study in the future.

Acknowledgments

The authors thank the Editor, Professor Debdeep Pati and three reviewers for their insightful comments and suggestions that greatly improved the quality of the paper. Dongxiao Han is partially supported by the National Natural Science Foundation of China (No.12101330, No.12231011), Tianjin Municipal Natural Science Foundation (No.23JCYBJC01270), and the Fundamental Research Funds for the Central Universities, Nankai University (No.632415 63). Jinhan Xie is supported by the National Key R&D Program of China (No.102022YFA10 03701). Jin Liu is supported by the National Natural Science Foundation of China (No.12201 316). Dongxiao Han and Jin Liu are partially supported by Shenzhen Wukong Investment Management Co. Ltd. Liuquan Sun is supported by the National Natural Science Foundation of China (No.12171463). Jian Huang is partially supported by the research grants (1-BDCC, 4-ZZ4B, 1-WZ3P, and 4-ZZPN) from The Hong Kong Polytechnic University and the National Natural Science Foundation of China (Grant No.72331005). Bei Jiang and Linglong Kong are partially supported by grants from the Canada CIFAR AI Chairs program, the Alberta Machine Intelligence Institute (AMII), and Natural Sciences and Engineering Council of Canada (NSERC), and Linglong Kong is also partially supported by grants from the Canada Research Chair program from NSERC.

Appendix A. Proofs of Proposition and Theorems

This Appendix contains technical proofs for Proposition 2 and Theorems 4-5 in Section 2.

A.1 Proof of Proposition 2

Proof By conditions (C1) and (C2), and the Jensen's inequality, we have

$$E\{l(Y, \boldsymbol{X}^\top \boldsymbol{\beta})\} = E[E\{l(Y, \boldsymbol{X}^\top \boldsymbol{\beta})|\boldsymbol{X}^\top \boldsymbol{\beta}_0, \epsilon\}] \geq E\{l(Y, c_{\boldsymbol{\beta}} \boldsymbol{X}^\top \boldsymbol{\beta}_0)\}, \tag{14}$$

where $c_{\boldsymbol{\beta}}$ is a constant depending on $\boldsymbol{\beta}$. Condition (C2) and (14) imply that there exists some constant $k_1 \neq 0$ such that $\boldsymbol{\beta}^* = k_1 \boldsymbol{\beta}_0$. We finish the proof of Proposition 2. ■

A.2 Proof of Theorem 4

Proof We will prove the theorem by mathematical induction. In what follows, we assume that n_1 is sufficient large. Using condition (C3), a Hoeffding-type inequality (Vershynin, 2012, Proposition 5.10) and the union inequality, we can show

$$P(\|\frac{2}{n_1} \sum_{i=1}^{n_1/2} \boldsymbol{Z}_i^{(1)}\|_\infty \geq \frac{\lambda_1}{2}) \leq ep \exp(-\frac{a_1 \lambda_1^2 n_1}{8M_1^2}) \leq ep^{-a_0}, \tag{15}$$

where a_1 is a positive constant not depending on any parameter, $\lambda_1 = c_{11}\sqrt{\log p / n_1}$, c_{11} could be any constant which belongs to $[2M_1\sqrt{2(a_0+1)/a_1}, a_2]$, and a_2 could be any constant no less than $2M_1\sqrt{2(a_0+1)/a_1}$. For any $\boldsymbol{\Delta} \in \mathbb{R}^p$, define $\boldsymbol{\Delta}_S = \{\Delta_j | \beta_j^* \neq 0\}$, and $\boldsymbol{\Delta}_{S^c} = \{\Delta_j | \beta_j^* = 0\}$, where Δ_j is the jth element of $\boldsymbol{\Delta}$. Let $\hat{\boldsymbol{\Delta}}_1^{(1)} = \hat{\boldsymbol{\beta}}_1^{(1)} - \boldsymbol{\beta}^*$, and $\hat{\boldsymbol{\Delta}}_2^{(1)} = \hat{\boldsymbol{\beta}}_2^{(1)} - \boldsymbol{\beta}^*$. According to the fact that $\hat{\boldsymbol{\beta}}_1^{(1)}$ is the minimizer of (2) and the convexity of $l_1^{(1)}(\boldsymbol{\beta})$, one can show

$$\hat{\boldsymbol{\Delta}}_1^{(1)\top} \nabla l_1^{(1)}(\boldsymbol{\beta}^*) \leq l_1^{(1)}(\hat{\boldsymbol{\beta}}_1^{(1)}) - l_1^{(1)}(\boldsymbol{\beta}^*) \leq \lambda_1 \|\boldsymbol{\beta}^*\|_1 - \lambda_1 \|\hat{\boldsymbol{\beta}}_1^{(1)}\|_1 \leq \lambda_1 \|\hat{\boldsymbol{\Delta}}_{1S}^{(1)}\|_1 - \lambda_1 \|\hat{\boldsymbol{\Delta}}_{1S^c}^{(1)}\|_1. \tag{16}$$

In light of the Hölder's inequality, (15) and (16), we can prove that with probability at least $1 - ep^{-a_0}$,

$$-\frac{\lambda_1}{2} \|\hat{\boldsymbol{\Delta}}_1^{(1)}\|_1 \leq -\|\nabla l_1^{(1)}(\boldsymbol{\beta}^*)\|_\infty \|\hat{\boldsymbol{\Delta}}_1^{(1)}\|_1 \leq \lambda_1 \|\hat{\boldsymbol{\Delta}}_{1S}^{(1)}\|_1 - \lambda_1 \|\hat{\boldsymbol{\Delta}}_{1S^c}^{(1)}\|_1.$$

This implies that with probability at least $1 - ep^{-a_0}$,

$$\|\hat{\boldsymbol{\Delta}}_{1S^c}^{(1)}\|_1 \leq 3 \|\hat{\boldsymbol{\Delta}}_{1S}^{(1)}\|_1,$$

which indicates that with probability at least $1 - ep^{-a_0}$,

$$\hat{\boldsymbol{\Delta}}_1^{(1)} \in C_1 \equiv \{\boldsymbol{\Delta} | \|\boldsymbol{\Delta}_{S^c}\|_1 \leq 3 \|\boldsymbol{\Delta}_S\|_1\}. \tag{17}$$

Let $C_2 \equiv \{\boldsymbol{\Delta} | \|\boldsymbol{\Delta}\|_1 \leq 1\}$. Based on conditions (C4), (C6), (15), the triangle inequality, the Hölder's inequality and the Cauchy-Schwarz inequality, we can show that with probability at least $1 - P(n_1, p) - ep^{-a_0}$,

$$\begin{aligned}
& l_1^{(1)}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) + \lambda_1 \|\boldsymbol{\beta}^* + \boldsymbol{\Delta}\|_1 - l_1^{(1)}(\boldsymbol{\beta}^*) - \lambda_1 \|\boldsymbol{\beta}^*\|_1 \\
\geq & \boldsymbol{\Delta}^\top \nabla l_1^{(1)}(\boldsymbol{\beta}^*) + M_4 ||\boldsymbol{\Delta}||_2^2 - M_5 \sqrt{\frac{\log p}{n_1}} ||\boldsymbol{\Delta}||_1 ||\boldsymbol{\Delta}||_2 + \lambda_1 \|\boldsymbol{\Delta}_{S^c}\|_1 - \lambda_1 \|\boldsymbol{\Delta}_S\|_1 \\
\geq & - \|\boldsymbol{\Delta}\|_1 \|\nabla l_1^{(1)}(\boldsymbol{\beta}^*)\|_\infty + M_4 ||\boldsymbol{\Delta}||_2^2 - M_5 \sqrt{\frac{\log p}{n_1}} ||\boldsymbol{\Delta}||_1 ||\boldsymbol{\Delta}||_2 + \lambda_1 \|\boldsymbol{\Delta}_{S^c}\|_1 - \lambda_1 \|\boldsymbol{\Delta}_S\|_1 \\
\geq & M_4 ||\boldsymbol{\Delta}||_2^2 - M_5 \sqrt{\frac{\log p}{n_1}} ||\boldsymbol{\Delta}||_1 ||\boldsymbol{\Delta}||_2 - \frac{3\lambda_1}{2} \|\boldsymbol{\Delta}_S\|_1 \\
\geq & M_4 ||\boldsymbol{\Delta}||_2^2 - 4M_5 \sqrt{\frac{\log p}{n_1}} ||\boldsymbol{\Delta}||_2 ||\boldsymbol{\Delta}_S||_1 - \frac{3\lambda_1}{2} \|\boldsymbol{\Delta}_S\|_1 \\
\geq & (M_4 - 4M_5 \sqrt{\frac{s_0 \log p}{n_1}}) ||\boldsymbol{\Delta}||_2^2 - \frac{3\sqrt{s_0}\lambda_1}{2} \|\boldsymbol{\Delta}\|_2 \\
\geq & \frac{M_4}{2} ||\boldsymbol{\Delta}||_2^2 - \frac{3\sqrt{s_0}\lambda_1}{2} \|\boldsymbol{\Delta}\|_2, \qquad (18)
\end{aligned}$$

for all $\boldsymbol{\Delta} \in C_1 \cap C_2$. Some algebra shows that the right side of (18) is positive as long as $||\boldsymbol{\Delta}||_2 > 3\sqrt{s_0}\lambda_1/M_4$. It follows from Lemma 4 of Negahban et al. (2012) that with probability at least $1 - P(n_1, p) - ep^{-a_0}$,

$$||\hat{\boldsymbol{\Delta}}_1^{(1)}||_2 \leq 3\sqrt{s_0}\lambda_1/M_4. \qquad (19)$$

Thus, by the Cauchy-Schwarz inequality and (17), we have that with probability at least $1 - P(n_1, p) - ep^{-a_0}$,

$$||\hat{\boldsymbol{\Delta}}_1^{(1)}||_1 \leq 4||\hat{\boldsymbol{\Delta}}_{1S}^{(1)}||_1 \leq 4\sqrt{s_0}||\hat{\boldsymbol{\Delta}}_{1S}^{(1)}||_2 \leq 4\sqrt{s_0}||\hat{\boldsymbol{\Delta}}_1^{(1)}||_2 \leq 12 s_0 \lambda_1/M_4. \qquad (20)$$

Let $\gamma_1 = c_{21}\sqrt{\log p/n_1}$, where c_{21} could be any constant which belongs to $[2M_1\sqrt{2(a_0+1)/a_1}, a_2]$. Similar to (19) and (20), we can obtain that with probability at least $1-P(n_1,p)-ep^{-a_0}$,

$$||\hat{\boldsymbol{\Delta}}_2^{(1)}||_2 \leq 3\sqrt{s_0}\gamma_1/M_4, \quad \text{and} \quad ||\hat{\boldsymbol{\Delta}}_2^{(1)}||_1 \leq 12 s_0 \gamma_1/M_4. \qquad (21)$$

Using (19)-(21), and the triangle inequality, one can prove that with probability at least $1 - P(n_1, p) - 2ep^{-a_0}$,

$$||\hat{\boldsymbol{\beta}}_{ave}^{(1)} - \boldsymbol{\beta}^*||_2 \leq 3\sqrt{s_0}(\lambda_1 + \gamma_1)/(2M_4), \quad \text{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(1)} - \boldsymbol{\beta}^*||_1 \leq 6 s_0 (\lambda_1 + \gamma_1)/M_4.$$

Let $d_1 = \max\{3a_2/M_4, 4\}$. Then we can show that with probability at least $1 - P(n_1, p) - 2ep^{-a_0}$,

$$\begin{aligned}
& ||\hat{\boldsymbol{\Delta}}_1^{(1)}||_2 \leq d_1 \sqrt{\frac{s_0 \log p}{n_1}}, \quad ||\hat{\boldsymbol{\Delta}}_1^{(1)}||_1 \leq d_1^2 s_0 \sqrt{\frac{\log p}{n_1}}, \\
& ||\hat{\boldsymbol{\Delta}}_2^{(1)}||_2 \leq d_1 \sqrt{\frac{s_0 \log p}{n_1}}, \quad ||\hat{\boldsymbol{\Delta}}_2^{(1)}||_1 \leq d_1^2 s_0 \sqrt{\frac{\log p}{n_1}}, \\
& ||\hat{\boldsymbol{\beta}}_{ave}^{(1)} - \boldsymbol{\beta}^*||_2 \leq d_1 \sqrt{\frac{s_0 \log p}{n_1}}, \quad \text{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(1)} - \boldsymbol{\beta}^*||_1 \leq d_1^2 s_0 \sqrt{\frac{\log p}{n_1}}. \qquad (22)
\end{aligned}$$

Similar to (15), we have

$$P(\|\frac{2}{N_2}\sum_{i=1}^{n_2/2}\boldsymbol{Z}_i^{(2)}\|_\infty \geq \frac{\lambda_2}{2}) \leq ep^{-a_0N_2/n_2}, \tag{23}$$

where $\lambda_2 = c_{12}\sqrt{\log p/N_2}$, and c_{12} can be any constant which belongs to $[2M_1\sqrt{2(a_0+1)/a_1}$, $a_2]$. Define $\hat{\boldsymbol{\Delta}}_1^{(2)} = \hat{\boldsymbol{\beta}}_1^{(2)} - \boldsymbol{\beta}^*$, and $\hat{\boldsymbol{\Delta}}_2^{(2)} = \hat{\boldsymbol{\beta}}_2^{(2)} - \boldsymbol{\beta}^*$. Using the fact that $\hat{\boldsymbol{\beta}}_1^{(2)}$ is the minimizer of (5) in the main manuscript and the triangle inequality, one can prove

$$L_{12}(\hat{\boldsymbol{\beta}}_1^{(2)}) - L_{12}(\boldsymbol{\beta}^*) \leq \lambda_2\|\boldsymbol{\beta}^*\|_1 - \lambda_2\|\hat{\boldsymbol{\beta}}_1^{(2)}\|_1 \leq \lambda_2\|\hat{\boldsymbol{\Delta}}_{1S}^{(2)}\|_1 - \lambda_2\|\hat{\boldsymbol{\Delta}}_{1S^c}^{(2)}\|_1. \tag{24}$$

By the convexity of $L_{12}(\boldsymbol{\beta})$, the Cauchy-Schwarz inequality, the Hölder's inequality, conditions (C4), (C5), (C7), (22) and (23), we can show that with probability at least $1 - P_1(n_1, p) - ep^{-a_0N_2/n_2}$,

$$\begin{aligned}
&L_{12}(\hat{\boldsymbol{\beta}}_1^{(2)}) - L_{12}(\boldsymbol{\beta}^*) &(25)\\
\geq& -\frac{n_1}{N_2}\{\hat{\boldsymbol{\Delta}}_1^{(2)\top}\boldsymbol{H}_1^{(1)}\hat{\boldsymbol{\Delta}}_2^{(1)}\} + \hat{\boldsymbol{\Delta}}_1^{(2)\top}\frac{2}{N_2}\sum_{i=1}^{n_2/2}\boldsymbol{Z}_i^{(2)}\\
\geq& -\frac{n_1}{N_2}\{\hat{\boldsymbol{\Delta}}_1^{(2)\top}(\boldsymbol{H}_1^{(1)} - \boldsymbol{H})\hat{\boldsymbol{\Delta}}_2^{(1)} + \hat{\boldsymbol{\Delta}}_1^{(2)\top}\boldsymbol{H}\hat{\boldsymbol{\Delta}}_2^{(1)}\} - \|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_1\|\frac{2}{N_2}\sum_{i=1}^{n_2/2}\boldsymbol{Z}_i^{(2)}\|_\infty\\
\geq& -\frac{n_1}{N_2}\{\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_1\|\boldsymbol{H}_1^{(1)} - \boldsymbol{H}\|_\infty\|\hat{\boldsymbol{\Delta}}_2^{(1)}\|_1 + M_3\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_2\|\hat{\boldsymbol{\Delta}}_2^{(1)}\|_2\} - \frac{\lambda_2}{2}\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_1\\
\geq& -d_1^2M_6\sqrt{\frac{s_0^3\log p}{N_2}}\sqrt{\frac{\log p}{N_2}}\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_1 - M_3d_1\sqrt{\frac{s_0\log p}{N_2}}\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_2 - \frac{\lambda_2}{2}\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_1\\
\geq& -\frac{\lambda_2}{4}\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_1 - M_3d_1\sqrt{\frac{s_0\log p}{N_2}}\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_2 - \frac{\lambda_2}{2}\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_1\\
=& -\frac{3\lambda_2}{4}\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_1 - M_3d_1\sqrt{\frac{s_0\log p}{N_2}}\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_2. &(26)
\end{aligned}$$

Both (24) and (25) imply that with probability at least $1 - P_1(n_1, p) - ep^{-a_0N_2/n_2}$,

$$\|\hat{\boldsymbol{\Delta}}_{1S^c}^{(2)}\|_1 \leq 7\|\hat{\boldsymbol{\Delta}}_{1S}^{(2)}\|_1 + \frac{2M_3d_1}{M_1\sqrt{(a_0+1)/a_1}}\sqrt{s_0}\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_2. \tag{27}$$

It is straightforward to verify

$$\begin{aligned}
&L_{12}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) + \lambda_2\|\boldsymbol{\beta}^* + \boldsymbol{\Delta}\|_1 - L_{12}(\boldsymbol{\beta}^*) - \lambda_2\|\boldsymbol{\beta}^*\|_1\\
=&\frac{n_1}{N_2}\{\boldsymbol{\Delta}^\top\boldsymbol{H}_1^{(1)}\boldsymbol{\Delta}/2 + \boldsymbol{\Delta}^\top\boldsymbol{H}_1^{(1)}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(1)})\} + \frac{n_2}{N_2}\{l_1^{(2)}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - l_1^{(2)}(\boldsymbol{\beta}^*)\}\\
&+ \lambda_2\|\boldsymbol{\beta}^* + \boldsymbol{\Delta}\|_1 - \lambda_2\|\boldsymbol{\beta}^*\|_1. &(28)
\end{aligned}$$

Let $b_0 = 2M_3d_1/\{M_1\sqrt{2(a_0+1)/a_1}\}$, and $D_2 \equiv \{\boldsymbol{\Delta}|\|\boldsymbol{\Delta}_{S^c}\|_1 \leq 7\|\boldsymbol{\Delta}_S\|_1 + b_0\sqrt{s_0}\|\boldsymbol{\Delta}\|_2\}$. By conditions (C4), (C5), (C7), (22), the Hölder's inequality and the Cauchy-Schwarz inequality, we can show that with probability at least $1 - P_1(n_1, p)$,

$$
\begin{aligned}
&\boldsymbol{\Delta}^\top \boldsymbol{H}_1^{(1)} \boldsymbol{\Delta}/2 + \boldsymbol{\Delta}^\top \boldsymbol{H}_1^{(1)} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(1)}) \\
=&\boldsymbol{\Delta}^\top \boldsymbol{H} \boldsymbol{\Delta}/2 + \boldsymbol{\Delta}^\top (\boldsymbol{H}_1^{(1)} - \boldsymbol{H}) \boldsymbol{\Delta}/2 + \boldsymbol{\Delta}^\top \boldsymbol{H} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(1)}) + \boldsymbol{\Delta}^\top (\boldsymbol{H}_1^{(1)} - \boldsymbol{H})(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(1)}) \\
\geq& M_2 \|\boldsymbol{\Delta}\|_2^2/2 - \|\boldsymbol{H}_1^{(1)} - \boldsymbol{H}\|_\infty (\|\boldsymbol{\Delta}\|_1^2/2 + \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(1)}\|_1 \|\boldsymbol{\Delta}\|_1) - M_3 \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(1)}\|_2 \|\boldsymbol{\Delta}\|_2 \\
\geq& M_2 \|\boldsymbol{\Delta}\|_2^2/2 - M_6 \sqrt{\frac{s_0 \log p}{n_1}} (64 \|\boldsymbol{\Delta}_S\|_1^2 + b_0^2 s_0 \|\boldsymbol{\Delta}\|_2^2) - M_3 d_1 \sqrt{\frac{s_0 \log p}{n_1}} \|\boldsymbol{\Delta}\|_2 \\
&- M_6 d_1^2 s_0 \sqrt{s_0} \frac{\log p}{n_1} (8 \|\boldsymbol{\Delta}_S\|_1 + b_0 \sqrt{s_0} \|\boldsymbol{\Delta}\|_2) \\
\geq& M_2 \|\boldsymbol{\Delta}\|_2^2/2 - M_6 \sqrt{\frac{s_0^3 \log p}{n_1}} (64 \|\boldsymbol{\Delta}_S\|_2^2 + b_0^2 \|\boldsymbol{\Delta}\|_2^2) - M_3 d_1 \sqrt{\frac{s_0 \log p}{n_1}} \|\boldsymbol{\Delta}\|_2 \\
&- M_6 d_1^2 s_0^2 \frac{\log p}{n_1} (8 + b_0) \|\boldsymbol{\Delta}\|_2 \\
\geq& M_2 \|\boldsymbol{\Delta}\|_2^2/2 - M_6 \sqrt{\frac{s_0^3 \log p}{n_1}} (64 \|\boldsymbol{\Delta}_S\|_2^2 + b_0^2 \|\boldsymbol{\Delta}\|_2^2) - 2 M_3 d_1 \sqrt{\frac{s_0 \log p}{n_1}} \|\boldsymbol{\Delta}\|_2, \qquad (29)
\end{aligned}
$$

for all $\boldsymbol{\Delta} \in D_2$. Using conditions (C6), (23), (27), the Hölder's inequality and the Cauchy-Schwarz inequality, one can prove that with probability at least $1 - P(n_2, p) - ep^{-a_0 N_2/n_2}$,

$$
\begin{aligned}
\frac{n_2}{N_2} \{ l_1^{(2)}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - l_1^{(2)}(\boldsymbol{\beta}^*) \} \geq& \frac{n_2}{N_2} \{ \boldsymbol{\Delta}^\top \triangledown l_1^{(2)}(\boldsymbol{\beta}^*) + M_4 ||\boldsymbol{\Delta}||_2^2 - M_5 \sqrt{\frac{\log p}{n_2}} ||\boldsymbol{\Delta}||_1 ||\boldsymbol{\Delta}||_2 \} \\
\geq& -\frac{\lambda_2}{2} \|\boldsymbol{\Delta}\|_1 + \frac{n_2}{N_2} (M_4 ||\boldsymbol{\Delta}||_2^2 - 8 M_5 \sqrt{\frac{\log p}{n_2}} ||\boldsymbol{\Delta}_S||_1 ||\boldsymbol{\Delta}||_2) \\
&- \sqrt{\frac{n_2}{N_2}} M_5 b_0 \sqrt{\frac{s_0 \log p}{N_2}} \|\boldsymbol{\Delta}\|_2^2 \\
\geq& -\frac{\lambda_2}{2} \|\boldsymbol{\Delta}\|_1 + \frac{n_2}{N_2} M_4 ||\boldsymbol{\Delta}||_2^2 - 8 M_5 \sqrt{\frac{s_0 \log p}{N_2}} ||\boldsymbol{\Delta}||_2^2 \\
&- M_5 b_0 \sqrt{\frac{s_0 \log p}{N_2}} \|\boldsymbol{\Delta}\|_2^2, \qquad (30)
\end{aligned}
$$

for all $\boldsymbol{\Delta} \in C_2 \cap D_2$. Based on (28)-(30), condition (C4) and the Cauchy-Schwarz inequality, we have that with probability at least $1 - P(n_2, p) - P_1(n_1, p) - ep^{-a_0 N_2/n_2}$,

$$
\begin{aligned}
&L_{12}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) + \lambda_2 \|\boldsymbol{\beta}^* + \boldsymbol{\Delta}\|_1 - L_{12}(\boldsymbol{\beta}^*) - \lambda_2 \|\boldsymbol{\beta}^*\|_1 \\
\geq& \min\{\frac{M_2}{2}, M_4\} \|\boldsymbol{\Delta}\|_2^2 - 8 M_5 \sqrt{\frac{s_0 \log p}{N_2}} ||\boldsymbol{\Delta}||_2^2 - M_5 b_0 \sqrt{\frac{s_0 \log p}{N_2}} \|\boldsymbol{\Delta}\|_2^2 \\
&- M_6 \sqrt{\frac{s_0^3 \log p}{n_1}} (64 \|\boldsymbol{\Delta}_S\|_2^2 + b_0^2 \|\boldsymbol{\Delta}\|_2^2) - 2 M_3 d_1 \sqrt{\frac{s_0 \log p}{N_2}} \|\boldsymbol{\Delta}\|_2 \\
&- \frac{\lambda_2}{2} \|\boldsymbol{\Delta}\|_1 + \lambda_2 \|\boldsymbol{\beta}^* + \boldsymbol{\Delta}\|_1 - \lambda_2 \|\boldsymbol{\beta}^*\|_1
\end{aligned}
$$

$$
\begin{aligned}
&\geq \min\{\frac{M_2}{3},\frac{M_4}{2}\}\|\boldsymbol{\Delta}\|_2^2 - 2M_3 d_1\sqrt{\frac{s_0\log p}{N_2}}\|\boldsymbol{\Delta}\|_2 - \frac{3\lambda_2}{2}\|\boldsymbol{\Delta}_S\|_1 \\
&\geq \min\{\frac{M_2}{3},\frac{M_4}{2}\}\|\boldsymbol{\Delta}\|_2^2 - (2M_3 d_1 + \frac{3}{2}a_2)\sqrt{\frac{s_0\log p}{N_2}}\|\boldsymbol{\Delta}\|_2. \qquad (31)
\end{aligned}
$$

Some algebra shows that the right hand side of (31) is positive when $||\boldsymbol{\Delta}||_2 > d_2\sqrt{s_0\log p/N_2}$, where $d_2 = a_3 d_1$ and $a_3 = \max\{(2M_3{+}3a_2/2)/\min\{M_2/3, M_4/2\}, 8{+}2M_3/\{M_1\sqrt{2(a_0+1)}/\sqrt{a_1}\}\}$. It follows from Lemma 4 of Negahban et al. (2012) that with probability at least $1 - P(n_2,p) - P_1(n_1,p) - ep^{-a_0N_2/n_2}$,

$$
||\hat{\boldsymbol{\Delta}}_1^{(2)}||_2 \leq d_2\sqrt{\frac{s_0\log p}{N_2}}. \qquad (32)
$$

Then by the Cauchy-Schwarz inequality, we have that with probability at least $1-P(n_2,p)-P_1(n_1,p) - ep^{-a_0N_2/n_2}$,

$$
||\hat{\boldsymbol{\Delta}}_1^{(2)}||_1 \leq 8\|\hat{\boldsymbol{\Delta}}_{1S}^{(2)}\|_1 + b_0\sqrt{s_0}\|\hat{\boldsymbol{\Delta}}_1^{(2)}\|_2 \leq (8+b_0)\sqrt{s_0}||\hat{\boldsymbol{\Delta}}_1^{(1)}||_2 \leq d_2^2 s_0\sqrt{\frac{\log p}{N_2}}. \qquad (33)
$$

Let $\gamma_2 = c_{22}\sqrt{\log p/N_2}$, where c_{22} could be any constant that belongs to $[2M_1\sqrt{2(a_0+1)/a_1}, a_2]$. Similar to (32) and (33), we have that with probability at least $1-P(n_2,p)-P_1(n_1,p) - e/p^{a_0N_2/n_2}$,

$$
||\hat{\boldsymbol{\Delta}}_2^{(2)}||_2 \leq d_2\sqrt{\frac{s_0\log p}{N_2}}, \quad \text{and} \quad ||\hat{\boldsymbol{\Delta}}_2^{(2)}||_1 \leq d_2^2 s_0\sqrt{\frac{\log p}{N_2}}. \qquad (34)
$$

In light of (33), and (34) and the triangle inequality, we can obtain that with probability at least $1 - P(n_2,p) - P_1(n_1,p) - 2ep^{-a_0N_2/n_2}$,

$$
\begin{aligned}
&||\hat{\boldsymbol{\Delta}}_1^{(2)}||_2 \leq d_2\sqrt{\frac{s_0\log p}{N_2}}, \quad ||\hat{\boldsymbol{\Delta}}_1^{(2)}||_1 \leq d_2^2 s_0\sqrt{\frac{\log p}{N_2}}, \\
&||\hat{\boldsymbol{\Delta}}_2^{(2)}||_2 \leq d_2\sqrt{\frac{s_0\log p}{N_2}}, \quad ||\hat{\boldsymbol{\Delta}}_2^{(2)}||_1 \leq d_2^2 s_0\sqrt{\frac{\log p}{N_2}}, \\
&||\hat{\boldsymbol{\beta}}_{ave}^{(2)} - \boldsymbol{\beta}^*||_2 \leq d_2\sqrt{\frac{s_0\log p}{N_2}}, \quad \text{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(2)} - \boldsymbol{\beta}^*||_1 \leq d_2^2 s_0\sqrt{\frac{\log p}{N_2}}. \qquad (35)
\end{aligned}
$$

Assume that with probability at least $1-P(n_{s-1},p)-P_{s-2}(n_1,\ldots,n_{s-2},p)-2ep^{-a_0N_{s-1}/n_{s-1}}$,

$$
\begin{aligned}
&||\hat{\boldsymbol{\Delta}}_1^{(s-1)}||_2 \leq d_{s-1}\sqrt{\frac{s_0\log p}{N_{s-1}}}, \quad ||\hat{\boldsymbol{\Delta}}_1^{(s-1)}||_1 \leq d_{s-1}^2 s_0\sqrt{\frac{\log p}{N_{s-1}}}, \\
&||\hat{\boldsymbol{\Delta}}_2^{(s-1)}||_2 \leq d_{s-1}\sqrt{\frac{s_0\log p}{N_{s-1}}}, \quad ||\hat{\boldsymbol{\Delta}}_2^{(s-1)}||_1 \leq d_{s-1}^2 s_0\sqrt{\frac{\log p}{N_{s-1}}}, \\
&||\hat{\boldsymbol{\beta}}_{ave}^{(s-1)} - \boldsymbol{\beta}^*||_2 \leq d_{s-1}\sqrt{\frac{s_0\log p}{N_{s-1}}}, \quad \text{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(s-1)} - \boldsymbol{\beta}^*||_1 \leq d_{s-1}^2 s_0\sqrt{\frac{\log p}{N_{s-1}}}, \qquad (36)
\end{aligned}
$$

where $d_{s-1} = d_1 a_3^{s-2}$. Let $\lambda_s = c_{1s}\sqrt{\log p/N_s}$, where c_{1s} could be any constant which belongs to $[2M_1\sqrt{2(a_0+1)/a_1}, a_2]$. Similar to (15) and (24), we have

$$P(\|\frac{2}{N_s}\sum_{i=1}^{n_s/2}\boldsymbol{Z}_i^{(s)}\|_\infty \geq \frac{\lambda_s}{2}) \leq ep^{-a_0N_s/n_s}, \tag{37}$$

and

$$L_{1s}(\hat{\boldsymbol{\beta}}_1^{(s)}) - L_{1s}(\boldsymbol{\beta}^*) \leq \lambda_s\|\hat{\boldsymbol{\Delta}}_{1S}^{(s)}\|_1 - \lambda_s\|\hat{\boldsymbol{\Delta}}_{1S^c}^{(s)}\|_1. \tag{38}$$

Based on the convexity of $L_{1s}(\boldsymbol{\beta})$, the Cauchy-Schwarz inequality, the Hölder's inequality, conditions (C4)-(C7), (36) and (37), one can prove that with probability at least $1 - ep^{-a_0N_s/n_s} - P_{s-1}(n_1,\ldots,n_{s-1},p)$,

$$\begin{aligned}
L_{1s}(\hat{\boldsymbol{\beta}}_1^{(s)}) - L_{1s}(\boldsymbol{\beta}^*) \geq &-\frac{N_{s-1}}{N_s}\{\hat{\boldsymbol{\Delta}}_1^{(s)\top}\frac{1}{N_{s-1}}\sum_{j=1}^{s-1}n_j\boldsymbol{H}_1^{(j)}\hat{\boldsymbol{\Delta}}_2^{(s-1)}\} + \hat{\boldsymbol{\Delta}}_1^{(s)\top}\frac{2}{N_s}\sum_{i=1}^{n_s/2}\boldsymbol{Z}_i^{(s)} \\
\geq &-\frac{N_{s-1}}{N_s}\{\hat{\boldsymbol{\Delta}}_1^{(s)\top}\frac{1}{N_{s-1}}\sum_{j=1}^{s-1}n_j(\boldsymbol{H}_1^{(j)}-\boldsymbol{H})\hat{\boldsymbol{\Delta}}_2^{(s-1)} + \hat{\boldsymbol{\Delta}}_1^{(s)\top}\boldsymbol{H}\hat{\boldsymbol{\Delta}}_2^{(s-1)}\} \\
&-\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_1\|\frac{2}{N_s}\sum_{i=1}^{n_s/2}\boldsymbol{Z}_i^{(s)}\|_\infty \\
\geq &-\frac{N_{s-1}}{N_s}\{\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_1\|\frac{1}{N_{s-1}}\sum_{j=1}^{s-1}n_j(\boldsymbol{H}_1^{(j)}-\boldsymbol{H})\|_\infty\|\hat{\boldsymbol{\Delta}}_2^{(s-1)}\|_1 \\
&+ M_3\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_2\|\hat{\boldsymbol{\Delta}}_2^{(s-1)}\|_2\} - \frac{\lambda_s}{2}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_1 \\
\geq &-d_{s-1}^2(\frac{1}{N_{s-1}}\sum_{j=1}^{s-1}n_jM_6^j\max\{\sqrt{\frac{s_0^3\log p}{n_j}}, \sqrt{s_0^3}\frac{\log p}{n_j}\})\sqrt{\frac{\log p}{N_s}}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_1 \\
&-\frac{\lambda_s}{2}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_1 - M_3d_{s-1}\sqrt{\frac{s_0\log p}{N_s}}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_2 \\
= &-d_{s-1}^2(\frac{1}{N_{s-1}}\sum_{j=1}^{s-1}M_6^j\max\{n_j^{1/2}n_1^{\alpha_1/2}\sqrt{\frac{s_0^3\log p}{n_1^{\alpha_1}}}, n_1^{\alpha_1}\sqrt{s_0^3}\frac{\log p}{n_1^{\alpha_1}}\})\sqrt{\frac{\log p}{N_s}} \\
&\times\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_1 - \frac{\lambda_s}{2}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_1 - M_3d_{s-1}\sqrt{\frac{s_0\log p}{N_s}}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_2 \\
\geq &-d_{s-1}^2N_{s-1}^{\alpha_1/2-1/2}(s-1)M_6^{s-1}\sqrt{\frac{s_0^3\log p}{n_1^{\alpha_1}}}\sqrt{\frac{\log p}{N_s}}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_1 - \frac{\lambda_s}{2}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_1 \\
&-M_3d_{s-1}\sqrt{\frac{s_0\log p}{N_s}}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_2 \\
\geq &-\frac{\lambda_s}{4}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_1 - M_3d_{s-1}\sqrt{\frac{s_0\log p}{N_s}}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_2 - \frac{\lambda_s}{2}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_1 \\
= &-\frac{3\lambda_s}{4}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_1 - M_3d_{s-1}\sqrt{\frac{s_0\log p}{N_s}}\|\hat{\boldsymbol{\Delta}}_1^{(s)}\|_2.
\end{aligned} \tag{39}$$

Both (38) and (39) indicate that with probability at least $1-ep^{-a_0N_s/n_s}-P_{s-1}(n_1,\ldots,n_{s-1},p)$,

$$\|\hat{\mathbf{\Delta}}_{1S^c}^{(s)}\|_1 \le 7\|\hat{\mathbf{\Delta}}_{1S}^{(s)}\|_1 + \frac{2M_3d_{s-1}}{M_1\sqrt{2(a_0+1)/a_1}}\sqrt{s_0}\|\hat{\mathbf{\Delta}}_1^{(s)}\|_2. \tag{40}$$

Using (36), (37), (40) and conditions (C4)-(C7), similar to (31), we can obtain that with probability at least $1-P(n_s,p)-P_{s-1}(n_1,\ldots,n_{s-1},p)-ep^{-a_0N_s/n_s}$,

$$\begin{aligned}&L_{1s}(\boldsymbol{\beta}^*+\mathbf{\Delta})+\lambda_s\|\boldsymbol{\beta}^*+\mathbf{\Delta}\|_1-L_{1s}(\boldsymbol{\beta}^*)-\lambda_s\|\boldsymbol{\beta}^*\|_1\\ &\ge\min\{\frac{M_2}{3},\frac{M_4}{2}\}\|\mathbf{\Delta}\|_2^2-(2M_3d_{s-1}+\frac{3}{2}a_2)\sqrt{\frac{s_0\log p}{N_s}}\|\mathbf{\Delta}\|_2,\end{aligned} \tag{41}$$

for all $\mathbf{\Delta}\in C_2\cap D_s$, where $D_s\equiv\{\mathbf{\Delta}|\|\mathbf{\Delta}_{S^c}\|_1\le 7\|\mathbf{\Delta}_S\|_1+b_1\sqrt{s_0}\|\mathbf{\Delta}\|_2\}$, and $b_1=2M_3d_{s-1}/\{M_1\sqrt{2(a_0+1)/a_1}\}$. Some algebra shows that the right hand side of (41) is positive as long as $||\mathbf{\Delta}||_2>d_s\sqrt{s_0\log p/N_2}$, where $d_s=d_1a_3^{s-1}$. Then it follows from Lemma 4 of Negahban et al. (2012) that with probability at least $1-P(n_s,p)-P_{s-1}(n_1,\ldots,n_{s-1},p)-ep^{-a_0N_s/n_s}$,

$$||\hat{\mathbf{\Delta}}_1^{(s)}||_2\le d_s\sqrt{\frac{s_0\log p}{N_s}}. \tag{42}$$

Let $\gamma_s=c_{2s}\sqrt{\log p/N_s}$, where c_{2s} could be any constant which belongs to $[2M_1\sqrt{2(a_0+1)/a_1}, a_2]$. Similar to (42), we have that with probability at least $1-P(n_s,p)-P_{s-1}(n_1,\ldots,n_{s-1},p)-ep^{-a_0N_s/n_s}$,

$$||\hat{\mathbf{\Delta}}_2^{(s)}||_2\le d_s\sqrt{\frac{s_0\log p}{N_s}}. \tag{43}$$

In light of (40), (42), (43) and the Cauchy-Schwarz inequality, we can show that with probability at least $1-P(n_s,p)-P_{s-1}(n_1,\ldots,n_{s-1},p)-2ep^{-a_0N_s/n_s}$,

$$||\hat{\mathbf{\Delta}}_1^{(s)}||_1\le 8\|\hat{\mathbf{\Delta}}_{1S}^{(s)}\|_1+b_1\sqrt{s_0}\|\hat{\mathbf{\Delta}}_1^{(s)}\|_2\le(8+b_1)\sqrt{s_0}||\hat{\mathbf{\Delta}}_1^{(s)}||_2\le d_s^2s_0\sqrt{\frac{\log p}{N_s}},$$

and

$$||\hat{\mathbf{\Delta}}_2^{(s)}||_1\le 8\|\hat{\mathbf{\Delta}}_{2S}^{(s)}\|_1+b_1\sqrt{s_0}\|\hat{\mathbf{\Delta}}_2^{(s)}\|_2\le(8+b_1)\sqrt{s_0}||\hat{\mathbf{\Delta}}_2^{(s)}||_2\le d_s^2s_0\sqrt{\frac{\log p}{N_s}}. \tag{44}$$

Using (42)-(44) and the triangle inequality, we can show that with probability at least $1-P(n_s,p)-P_{s-1}(n_1,\ldots,n_{s-1},p)-2ep^{-a_0N_s/n_s}$,

$$\begin{aligned}&||\hat{\mathbf{\Delta}}_1^{(s)}||_2\le d_s\sqrt{\frac{s_0\log p}{N_s}},\quad ||\hat{\mathbf{\Delta}}_1^{(s)}||_1\le d_s^2s_0\sqrt{\frac{\log p}{N_s}},\\ &||\hat{\mathbf{\Delta}}_2^{(s)}||_2\le d_s\sqrt{\frac{s_0\log p}{N_s}},\quad ||\hat{\mathbf{\Delta}}_2^{(s)}||_1\le d_s^2s_0\sqrt{\frac{\log p}{N_s}},\\ &||\hat{\boldsymbol{\beta}}_{ave}^{(s)}-\boldsymbol{\beta}^*||_2\le d_s\sqrt{\frac{s_0\log p}{N_s}},\quad\text{and}\quad ||\hat{\boldsymbol{\beta}}_{ave}^{(s)}-\boldsymbol{\beta}^*||_1\le d_s^2s_0\sqrt{\frac{\log p}{N_s}}.\end{aligned} \tag{45}$$

We complete the proof of Theorem 4. ■

A.3 Proof of Theorem 5

Proof Recall that

$$\hat{\beta}_{1,l}^{d(s)} - \beta_l^* = (I) + (II) + (III) + (IV) + (VI), \tag{46}$$

$$\begin{aligned}
(I) =& \boldsymbol{\Omega}_l^\top \sum_{j=1}^{s} n_j(\boldsymbol{H} - \boldsymbol{H}_1^{(j)})(\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}^*)/N_s, \\
(II) =& -(\hat{\boldsymbol{\Omega}}_{1,l}^{(s)} - \boldsymbol{\Omega}_l)^\top \left\{ \sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)}(\hat{\boldsymbol{\beta}}_1^{(s)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) + n_s \nabla l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)}) \right\} /N_s, \\
(III) =& -\boldsymbol{\Omega}_l^\top \left\{ \sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(j)}) + \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\boldsymbol{\beta}^*) \right\}/N_s, \\
(IV) =& -(\boldsymbol{\Omega}_l - \hat{\boldsymbol{\Omega}}_{1,l}^{(s)})^\top \Bigg\{ \sum_{j=1}^{s-1} n_j H_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}) + n_s \nabla l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)}) \\
&+ n_s H_1^{(s)}(\hat{\boldsymbol{\beta}}_2^{(s)} - \hat{\boldsymbol{\beta}}_1^{(s)}) \Bigg\}/N_s, \\
(VI) =& -\boldsymbol{\Omega}_l^\top \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\boldsymbol{\beta}^*)/N_s.
\end{aligned}$$

For (I), by the Hölder's inequality, conditions (C4), (C7), (D4) and Theorem 4, we can show

$$\begin{aligned}
&|\boldsymbol{\Omega}_l^\top \sum_{j=1}^{s} n_j(\boldsymbol{H} - \boldsymbol{H}_1^{(j)})(\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}^*)/N_s| \\
\le& \|\boldsymbol{\Omega}\|_{\infty,\infty} \|\hat{\boldsymbol{\beta}}_1^{(s)} - \boldsymbol{\beta}^*\|_1 \|\sum_{j=1}^{s} n_j(\boldsymbol{H} - \boldsymbol{H}_1^{(j)})/N_s\|_\infty \\
=& O_p(\|\boldsymbol{\Omega}\|_{\infty,\infty} d_s^2 \sqrt{\frac{s_0^2 \log p}{N_s}} \frac{1}{N_s} \sum_{j=1}^{s} n_j M_6^j \max\{\sqrt{\frac{s_0 \log p}{n_j}}, \sqrt{s_0}\frac{\log p}{n_j}\}) \\
=& o_p(\|\boldsymbol{\Omega}\|_{\infty,\infty} d_s^2 N_s^{\alpha_1/2-1} s\sqrt{\log p} M_6^s) \\
=& o_p(N_s^{-1/2}).
\end{aligned} \tag{47}$$

For (II), in light of the Hölder's inequality, conditions (D2), (D4) and the KKT conditions for $\hat{\boldsymbol{\beta}}_1^{(s)}$, one can show

$$\begin{aligned}
&|(\hat{\boldsymbol{\Omega}}_{1,l}^{(s)} - \boldsymbol{\Omega}_l)^\top \{\sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)}(\hat{\boldsymbol{\beta}}_1^{(s)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) + n_s \nabla l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)})\}/N_s| \\
\le& \lambda_s \|\hat{\boldsymbol{\Omega}}_1^{(s)} - \boldsymbol{\Omega})\|_{\infty,\infty}
\end{aligned}$$

$$
\begin{aligned}
&= O_p(\{g(s,s_0)\}^{(1-\omega)/2}\|\boldsymbol{\Omega}\|_{\infty,\infty}^{2(1-\omega)}\sqrt{\frac{\log p}{N_s}}(\frac{\log p}{N_s})^{(1-\omega)/2}v(p)) \\
&= o_p(N_s^{-1/2}). \qquad (48)
\end{aligned}
$$

Based on the Hölder's inequality and condition (D3), for (III), we can prove

$$
\begin{aligned}
&|\boldsymbol{\Omega}_l^\top\{\sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(j)}) + \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\boldsymbol{\beta}^*)\}/N_s| \\
\le& \|\boldsymbol{\Omega}\|_{\infty,\infty}\|\{\sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_2^{(j)}) + \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\boldsymbol{\beta}^*)\}/N_s\|_\infty \\
=& o_p(N_s^{-1/2}). \qquad (49)
\end{aligned}
$$

For (IV), according to the Hölder's inequality, the triangle inequality, Theorem 4, and conditions (C4), (C5), (C7), and (D2)-(D4), we have

$$
\begin{aligned}
&|(\boldsymbol{\Omega}_l - \hat{\boldsymbol{\Omega}}_{1,l}^{(s)})^\top\{\sum_{j=1}^{s-1} n_j H_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}) + n_s \nabla l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)}) \\
&\quad + n_s H_1^{(s)}(\hat{\boldsymbol{\beta}}_2^{(s)} - \hat{\boldsymbol{\beta}}_1^{(s)})\}/N_s| \\
\le& |(\boldsymbol{\Omega}_l - \hat{\boldsymbol{\Omega}}_{1,l}^{(s)})^\top\{\sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)} - \boldsymbol{\beta}^*) + \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\boldsymbol{\beta}^*) - \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)})\}/N_s| \\
&\quad + |(\boldsymbol{\Omega}_l - \hat{\boldsymbol{\Omega}}_{1,l}^{(s)})^\top \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\boldsymbol{\beta}^*)/N_s| \\
&\quad + |(\hat{\boldsymbol{\Omega}}_{1,l}^{(s)} - \boldsymbol{\Omega}_l)^\top\{\sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)}(\hat{\boldsymbol{\beta}}_1^{(s)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) + n_s \nabla l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)})\}/N_s| \\
&\quad + |(\hat{\boldsymbol{\Omega}}_{1,l}^{(s)} - \boldsymbol{\Omega}_l)^\top \sum_{j=1}^{s} n_j(\boldsymbol{H}_1^{(j)} - \boldsymbol{H})(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_1^{(s)})/N_s| \\
&\quad + |(\hat{\boldsymbol{\Omega}}_{1,l}^{(s)} - \boldsymbol{\Omega}_l)^\top \boldsymbol{H}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_1^{(s)})|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \|\hat{\boldsymbol{\Omega}}_1^{(s)} - \boldsymbol{\Omega})\|_{\infty,\infty} \|\{\sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)} - \boldsymbol{\beta}^*) + \sum_{j=1}^{s-1} n_j \nabla l_1^{(j)}(\boldsymbol{\beta}^*) - \sum_{j=1}^{s-1} n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)})\}/N_s\|_\infty \\
&\quad + \|\hat{\boldsymbol{\Omega}}_1^{(s)} - \boldsymbol{\Omega})\|_{\infty,\infty} \|\sum_{j=1}^{s-1} n_j \nabla l_1^{(j)}(\boldsymbol{\beta}_2^*)/N_s\|_\infty \\
&\quad + \|\hat{\boldsymbol{\Omega}}_1^{(s)} - \boldsymbol{\Omega})\|_{\infty,\infty} \|\{\sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)}(\hat{\boldsymbol{\beta}}_1^{(s)} - \hat{\boldsymbol{\beta}}_2^{(s-1)}) + n_s \nabla l_1^{(s)}(\hat{\boldsymbol{\beta}}_1^{(s)})\}/N_s\|_\infty \\
&\quad + \|\hat{\boldsymbol{\Omega}}_1^{(s)} - \boldsymbol{\Omega})\|_{\infty,\infty} \|\sum_{j=1}^{s} n_j (\boldsymbol{H}_1^{(j)} - \boldsymbol{H})/N_s\|_\infty \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_1^{(s)}\|_1 \\
&\quad + \|\hat{\boldsymbol{\Omega}}_1^{(s)} - \boldsymbol{\Omega})\|_{\infty,\infty} \|\boldsymbol{H}\|_\infty \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_1^{(s)}\|_1 \\
&= O_p(\|\boldsymbol{\Omega}\|_{\infty,\infty}^{2(1-\omega)} \{g(s,s_0)\frac{\log p}{N_s}\}^{(1-\omega)/2} v(p) N_s^{-1/2}) \\
&\quad + O_p(\|\boldsymbol{\Omega}\|_{\infty,\infty}^{2(1-\omega)} \{g(s,s_0)\frac{\log p}{N_s}\}^{(1-\omega)/2} v(p) \sqrt{\frac{\log p}{N_{s-1}}} \frac{N_{s-1}}{N_s}) \\
&\quad + O_p(\|\boldsymbol{\Omega}\|_{\infty,\infty}^{2(1-\omega)} \{g(s,s_0)\frac{\log p}{N_s}\}^{(1-\omega)/2} v(p) \sqrt{\frac{\log p}{N_s}}) \\
&\quad + O_p(\|\boldsymbol{\Omega}\|_{\infty,\infty}^{2(1-\omega)} \{g(s,s_0)\frac{\log p}{N_s}\}^{(1-\omega)/2} v(p) d_s^2 \sqrt{\frac{s_0^2 \log p}{N_s}} \frac{1}{N_s} \sum_{j=1}^{s} n_j M_6^j \max\{\sqrt{\frac{s_0 \log p}{n_j}}, \\
&\quad \sqrt{s_0} \frac{\log p}{n_j}\}) + O_p(\|\boldsymbol{\Omega}\|_{\infty,\infty}^{2(1-\omega)} \{g(s,s_0)\frac{\log p}{N_s}\}^{(1-\omega)/2} v(p) d_s^2 \sqrt{\frac{s_0^2 \log p}{N_s}}) \\
&= o_p(N_s^{-1/2}) + O_p(\|\boldsymbol{\Omega}\|_{\infty,\infty}^{2(1-\omega)} \{g(s,s_0)\frac{\log p}{N_s}\}^{(1-\omega)/2} v(p) \sqrt{\log p} N_s^{-1/2}) \\
&\quad + O_p(\|\boldsymbol{\Omega}\|_{\infty,\infty}^{2(1-\omega)} \{g(s,s_0)\frac{\log p}{N_s}\}^{(1-\omega)/2} v(p) \sqrt{\log p} N_s^{-1/2}) \\
&\quad + o_p(\|\boldsymbol{\Omega}\|_{\infty,\infty}^{2(1-\omega)} \{g(s,s_0)\frac{\log p}{N_s}\}^{(1-\omega)/2} v(p) d_s^2 \sqrt{\log p} s M_6^s N_s^{\alpha_1/2 - 1/2} N_s^{-1/2}) \\
&\quad + O_p(\|\boldsymbol{\Omega}\|_{\infty,\infty}^{2(1-\omega)} \{g(s,s_0)\frac{\log p}{N_s}\}^{(1-\omega)/2} v(p) d_s^2 \sqrt{s_0^2 \log p} N_s^{-1/2}) \\
&= o_p(N_s^{-1/2}). \qquad (50)
\end{aligned}
$$

Combining (46)-(50), one can show

$$
\hat{\beta}_{1,l}^{d(s)} - \beta_l^* = -\boldsymbol{\Omega}_l^\top \sum_{j=1}^{s} n_j \nabla l_1^{(j)}(\boldsymbol{\beta}^*)/N_s + o_p(N_s^{-1/2}). \tag{51}
$$

Similarly, we have

$$
\hat{\beta}_{2,l}^{d(s)} - \beta_l^* = -\boldsymbol{\Omega}_l^\top \sum_{j=1}^{s} n_j \nabla l_2^{(j)}(\boldsymbol{\beta}^*)/N_s + o_p(N_s^{-1/2}). \tag{52}
$$

Both (51) and (52) imply

$$\hat{\beta}_l^{da(s)} - \beta_l^* = -\mathbf{\Omega}_l^\top \sum_{j=1}^s n_j \nabla l_1^{(j)}(\boldsymbol{\beta}^*)/(2N_s) - \mathbf{\Omega}_l^\top \sum_{j=1}^s n_j \nabla l_2^{(j)}(\boldsymbol{\beta}^*)/(2N_s) + o_p(N_s^{-1/2}).$$

It follows from condition (D1), slutsky's theorem and the central limit theorem that $\sigma_l^{-1}\sqrt{N_s}$ $(\hat{\beta}_l^{da(s)} - \beta_l^*)$ copnverges to a standard normal random variable in distribution. We accomplish the proof of Theorem 5. ■

Appendix B. Proofs of Corollaries

This Appendix contains technical proofs for Corollaries 6-13 in Section 3. The following Lemmas 14 and 15 are used to prove these corollaries.

Lemma 14 *Suppose that conditions (C1) and (E1)-(E4) are satisfied. Then there exist five positive constants g_1, g_2, g_3, g_4 and g_9 depending on e_1, B_1, B_2, B_3 and B_4 such that for any $\tau \geq g_2$ and $1 \leq j \leq m$, with probability at least* $1 - \exp(-g_4 n_j - g_1 \log p)$,

$$l_1^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_1^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top \nabla l_1^{(j)}(\boldsymbol{\beta}^*) \geq g_3 ||\mathbf{\Delta}||_2^2 - g_9 \sqrt{\frac{\log p}{n_j}} ||\mathbf{\Delta}||_1 ||\mathbf{\Delta}||_2,$$

and

$$l_2^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_2^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top \nabla l_2^{(j)}(\boldsymbol{\beta}^*) \geq g_3 ||\mathbf{\Delta}||_2^2 - g_9 \sqrt{\frac{\log p}{n_j}} ||\mathbf{\Delta}||_1 ||\mathbf{\Delta}||_2,$$

for all $||\mathbf{\Delta}||_2 \leq 1$.

Proof Let

$$Q_q(x) \begin{cases} x^2 \quad \text{if} \quad |x| \leq \frac{q}{2}, \\ (q - |x|)^2 \quad \text{if} \quad \frac{q}{2} \leq |x| \leq q, \\ 0 \quad \text{otherwise.} \end{cases}$$

Let q_1 and q_2 be two positive numbers which will be specified later. Define $g_2 = \max\{q_1 + q_2, e_1\}$. Now we show that for any $\tau \geq g_2$,

$$l_1^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_1^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top \nabla l_1^{(j)}(\boldsymbol{\beta}^*) \geq \frac{1}{n_j} \sum_{i=1}^{\frac{n_j}{2}} Q_{q_2 \|\mathbf{\Delta}\|_2} \{\boldsymbol{X}_i^{(j)\top} \mathbf{\Delta} I(|y_i^{(j)} - \boldsymbol{X}_i^{(j)\top} \boldsymbol{\beta}_\tau^*| \leq q_1)\}, \tag{53}$$

for all $\|\mathbf{\Delta}\|_2 \leq 1$. If $|\boldsymbol{X}_i^{(j)\top}\mathbf{\Delta}| > q_2 \|\mathbf{\Delta}\|_2$ or $|y_i^{(j)} - \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}_\tau^*| > q_1$, the right hand side of (53) is 0. According to the convexity of the Huber loss, (53) holds. When $|\boldsymbol{X}_i^{(j)\top}\mathbf{\Delta}| \leq q_2 \|\mathbf{\Delta}\|_2$

and $|Y_i^{(j)} - \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}_\tau^*| \le q_1$, we can obtain

$$
\begin{aligned}
&\rho_\tau\{Y_i^{(j)} - \boldsymbol{X}_i^{(j)\top}(\boldsymbol{\beta}_\tau^* + \boldsymbol{\Delta})\} - \rho_\tau(Y_i^{(j)} - \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}_\tau^*) - \boldsymbol{\Delta}^\top \nabla\rho_\tau(Y_i^{(j)} - \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}_\tau^*)\\
=&\frac{(X_i^{(j)\top}\boldsymbol{\Delta})^2}{2}\\
\ge&\frac{Q_{q_2\|\boldsymbol{\Delta}\|_2}\{\boldsymbol{X}_i^{(j)\top}\boldsymbol{\Delta} I(|Y_i^{(j)} - \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}_\tau^*| \le q_1)\}}{2},
\end{aligned}
$$

implying (53) is also satisfied. By (53), to prove Lemma 14, it suffices to show that with probability at least $1 - \exp(-g_4 n_j - g_1 \log p)$ (g_4 and g_1 are positive constants and will be specified later),

$$
\frac{2}{n_j}\sum_{i=1}^{\frac{n_j}{2}} Q_{q_2\|\boldsymbol{\Delta}\|_2}\{\boldsymbol{X}_i^{(j)\top}\boldsymbol{\Delta} I(|Y_i^{(j)} - \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}_\tau^*| \le q_1)\} \ge 2g_3||\boldsymbol{\Delta}||_2^2 - 2g_9\sqrt{\frac{\log p}{n_j}}\|\boldsymbol{\Delta}\|_1\|\boldsymbol{\Delta}\|_2,
$$

for all $\|\boldsymbol{\Delta}\|_2 \le 1$. Since $Q_{q_2\|\boldsymbol{\Delta}\|_2}(x\|\boldsymbol{\Delta}\|_2) = \|\boldsymbol{\Delta}\|_2^2 Q_{q_2}(x)$, it is equivalent to show that with probability at least $1 - \exp(-g_4 n_j - g_1 \log p)$,

$$
\frac{2}{n_j}\sum_{i=1}^{\frac{n_j}{2}} Q_{q_2}\{\boldsymbol{X}_i^{(j)\top}\boldsymbol{\Delta} I(|Y_i^{(j)} - \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}_\tau^*| \le q_1)\} \ge 2g_3 - 2g_9\sqrt{\frac{\log p}{n_j}}||\boldsymbol{\Delta}||_1,
$$

for all $\|\boldsymbol{\Delta}\|_2 = 1$. Define $Q_{1,\boldsymbol{\Delta}}(\boldsymbol{X}, Y) = \boldsymbol{X}^\top\boldsymbol{\Delta} I(|Y - \boldsymbol{X}^\top\boldsymbol{\beta}_\tau^*| \le q_1)$ and $Q_{2,\boldsymbol{\Delta}}(\boldsymbol{X}, Y) = Q_{q_2}\{Q_{1,\boldsymbol{\Delta}}(\boldsymbol{X}, Y)\}$. We first show that for any $\|\boldsymbol{\Delta}\|_2 = 1$,

$$
E[Q_{2,\boldsymbol{\Delta}}(\boldsymbol{X}, Y)] \ge \frac{B_3}{2}. \tag{54}
$$

According to condition (E3), one can prove $E(\boldsymbol{X}^\top\boldsymbol{\Delta})^2 \ge B_3$, so that it suffices to prove

$$
E\{(\boldsymbol{X}^\top\boldsymbol{\Delta})^2 - Q_{2,\boldsymbol{\Delta}}(\boldsymbol{X}, Y)\} \le \frac{B_3}{2}. \tag{55}
$$

Note that when $|Y - \boldsymbol{X}^\top\boldsymbol{\beta}_\tau^*| \le q_1$ and $|\boldsymbol{X}^\top\boldsymbol{\Delta}| \le q_2/2$, $Q_{2,\boldsymbol{\Delta}}(\boldsymbol{X}, Y) = (\boldsymbol{X}^\top\boldsymbol{\Delta})^2$. As a result, we can obtain

$$
\begin{aligned}
&E\{(\boldsymbol{X}^\top\boldsymbol{\Delta})^2 - Q_{2,\boldsymbol{\Delta}}(\boldsymbol{X}, Y)\}\\
\le&E\{(\boldsymbol{X}^\top\boldsymbol{\Delta})^2 I(|Y - \boldsymbol{X}^\top\boldsymbol{\beta}_\tau^*| > q_1)\} + E\{(\boldsymbol{X}^\top\boldsymbol{\Delta})^2 I(|\boldsymbol{X}^\top\boldsymbol{\Delta}| > q_2/2)\}.
\end{aligned} \tag{56}
$$

In light of the Cauchy-Schwarz inequality, the Chebyshev inequality and conditions (E2) and (E3), we have

$$
\begin{aligned}
E\{(\boldsymbol{X}^\top\boldsymbol{\Delta})^2 I(|Y - \boldsymbol{X}^\top\boldsymbol{\beta}_\tau^*| > q_1)\} \le&E\{(\boldsymbol{X}^\top\boldsymbol{\Delta})^4\}^{1/2}\{P(|Y - \boldsymbol{X}^\top\boldsymbol{\beta}_\tau^*| > q_1)\}^{1/2}\\
\le&4B_1^2\{E(|\epsilon_\tau|)/q_1\}^{1/2}\\
\le&4B_1^2\sqrt{B_2/q_1}.
\end{aligned} \tag{57}
$$

By the Cauchy-Schwarz inequality, Lemma 5.5 of Vershynin (2012) and condition (E2), we can show

$$\begin{aligned} E\{(\boldsymbol{X}^\top\boldsymbol{\Delta})^2 I(|\boldsymbol{X}^\top\boldsymbol{\Delta}|>q_2/2)\} \leq & E\{(\boldsymbol{X}^\top\boldsymbol{\Delta})^4\}^{1/2}\{P(|\boldsymbol{X}^\top\boldsymbol{\Delta}|>q_2/2)\}^{1/2} \\ \leq & 4B_1^2\{P(|\boldsymbol{X}^\top\boldsymbol{\Delta}|>q_2/2)\}^{1/2} \\ \leq & 4\sqrt{e}B_1^2\exp(-q_2^2 q_3/4), \end{aligned} \tag{58}$$

where q_3 is a positive number which depends on B_1. Let

$$q_1 = 256B_1^4B_2/B_3^2, \quad \text{and} \quad q_2 = \max\{\sqrt{4\max\{\log(16\sqrt{e}B_1^2/B_3),1\}/q_3},1\}.$$

Then (56)-(58) indicate that (55) is satisfied. Define

$$Q_3(t) = \sup_{\{\|\boldsymbol{\Delta}\|_2=1\}\cap\{\|\boldsymbol{\Delta}\|_1\leq t\}} |\frac{2}{n_j}\sum_{i=1}^{\frac{n_j}{2}} Q_{2,\boldsymbol{\Delta}}(\boldsymbol{X}_i,Y_i) - E\{Q_{2,\boldsymbol{\Delta}}(\boldsymbol{X}_i,Y_i)\}|.$$

Now we show that there exist two positive numbers q_4 and q_5 which depends on B_1 and B_3 such that with probability at most $\exp(-q_4n_j - t^2\log p)$,

$$Q_3(t) \geq \frac{B_3}{8} + 40q_2^2q_5\sqrt{\frac{\log p}{n_j}}t. \tag{59}$$

For any positive number $z^*(t)$, based on Theorem 14.2 of Bühlmann and Van De Geer (2011), one can prove

$$P(Q_3(t) \geq E\{Q_3(t)\} + z^*(t)) \leq \exp\{-\frac{n_j z^{*2}(t)}{64q_2^4}\}.$$

Setting $z^*(t) = B_3/8 + 8q_2^2\sqrt{\log p/n_j}t$ and $q_4 = B_3^2/(4096q_2^4)$, we have

$$P(Q_3(t) \geq E\{Q_3(t)\} + z^*(t)) \leq \exp\{-\frac{n_j(B_3/8 + 8q_2^2\sqrt{\log p/n_j}t)^2}{64q_2^4}\} \leq \exp(-q_4n_j - t^2\log p). \tag{60}$$

Let $\{\omega_i\}_{i=1}^{n_j/2}$ be an independent and identically distributed sequence of Rademacher variables. By Theorem 14.3 of Bühlmann and Van De Geer (2011) and the Ledoux-Talagrand contraction theorem (Ledoux-Talagrand, 1991, page 112), we have

$$\begin{aligned} E\{Q_3(t)\} \leq & 2E \sup_{\{\|\boldsymbol{\Delta}\|_2=1\}\cap\{\|\boldsymbol{\Delta}\|_1\leq t\}} \frac{2}{n_j}|\sum_{i=1}^{\frac{n_j}{2}} \omega_i Q_{2,\boldsymbol{\Delta}}(\boldsymbol{X}_i,Y_i)| \\ \leq & 8q_2^2E \sup_{\{\|\boldsymbol{\Delta}\|_2=1\}\cap\{\|\boldsymbol{\Delta}\|_1\leq t\}} \frac{2}{n_j}|\sum_{i=1}^{\frac{n_j}{2}} \omega_i\boldsymbol{X}_i^\top\boldsymbol{\Delta}I(|Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta}_\tau^*|\leq q_1)| \\ \leq & 8q_2^2tE\{\|\frac{2}{n_j}\sum_{i=1}^{\frac{n_j}{2}}\omega_i\boldsymbol{X}_iI(|Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta}_\tau^*|\leq q_1)\|_\infty\}. \end{aligned} \tag{61}$$

For any positive number a, in light of the Jensen's inequality, we can obtain

$$\begin{aligned}
&E\{a\|\frac{2}{n_j}\sum_{i=1}^{\frac{n_j}{2}}\omega_i\boldsymbol{X}_i I(|Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta}_\tau^*|\leq q_1)\|_\infty\}\\
\leq&\log E[\exp\{a\|\frac{2}{n_j}\sum_{i=1}^{\frac{n_j}{2}}\omega_i\boldsymbol{X}_i I(|Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta}_\tau^*|\leq q_1)\|_\infty\}]. \qquad (62)
\end{aligned}$$

Define

$$U_{i,l}=\begin{cases}\omega_i\boldsymbol{X}_{i,l}I(|Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta}_\tau^*|\leq q_1) & \text{if} \quad 1\leq l\leq p,\\ -\omega_i\boldsymbol{X}_{i,l}I(|Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta}_\tau^*|\leq q_1) & \text{if} \quad p+1\leq l\leq 2p.\end{cases}$$

According to Lemma 5.5 of Vershynin (2012), one can prove that there exists a positive number q_5 which is no less than 1 and depends on B_1 such that

$$\begin{aligned}
E[\exp\{a\|\frac{2}{n_j}\sum_{i=1}^{\frac{n_j}{2}}\omega_i\boldsymbol{X}_i I(|Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta}_\tau^*|\leq q_1)\|_\infty\}] =&E\{\max_{1\leq l\leq 2p}\exp(\frac{2a}{n_j}\sum_{i=1}^{\frac{n_j}{2}}U_{i,l})\}\\
\leq&2p\max_{1\leq l\leq 2p}E\{\exp(\frac{2a}{n_j}\sum_{i=1}^{\frac{n_j}{2}}U_{i,l})\}\\
=&2p\max_{1\leq l\leq 2p}[E\{\exp(\frac{2a}{n_j}U_{i,l})\}]^{n_j/2}\\
\leq&2p\exp(\frac{2a^2q_5^2}{n_j}). \qquad (63)
\end{aligned}$$

Let $a=\sqrt{n_j\log(2p)/(2q_5^2)}$. Both (62) and (63) imply

$$\begin{aligned}
&E\{\|\frac{2}{n_j}\sum_{i=1}^{\frac{n_j}{2}}\omega_i\boldsymbol{X}_i I(|Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta}_\tau^*|\leq q_1)\|_\infty\}\\
\leq&\big(\log E[\exp\{a\|\frac{2}{n_j}\sum_{i=1}^{\frac{n_j}{2}}\omega_i\boldsymbol{X}_i I(|Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta}_\tau^*|\leq q_1)\|_\infty\}]\big)/a\\
\leq&\frac{\log(2p)}{a}+\frac{2aq_5^2}{n_j}\\
=&2\sqrt{2q_5^2\log(2p)/n_j}\\
\leq&4q_5\sqrt{\frac{\log(p)}{n_j}}.
\end{aligned}$$

It follows from (61) that

$$E\{Q_3(t)\}\leq 32q_2^2q_5\sqrt{\frac{\log(p)}{n_j}}t. \qquad (64)$$

Both (60) and (64) indicate that (59) is satisfied. For any positive integer i, define $t_i = (2^{i-1}B_3 - B_3/4)/(80q_2^2 q_5\sqrt{\log p/n_j})$. According to (59), one can prove that there exist two positive numbers q_6 and q_7 which depend on B_1 and B_3 such that

$$
\begin{aligned}
&P\big(\text{there exists a } \boldsymbol{\Delta} \text{ such that } \|\boldsymbol{\Delta}\|_2 = 1 \text{ and } Q_3(\|\boldsymbol{\Delta}\|_1) \geq \frac{B_3}{4} + 80q_2^2 q_5\sqrt{\frac{\log p}{n_j}}\|\boldsymbol{\Delta}\|_1\big)\\
\leq&\sum_{i=1}^{\infty} P\big(\text{there exists a } \boldsymbol{\Delta} \text{ such that } \|\boldsymbol{\Delta}\|_2 = 1,\ Q_3(\|\boldsymbol{\Delta}\|_1) \geq \frac{B_3}{4} + 80q_2^2 q_5\sqrt{\frac{\log p}{n_j}}\|\boldsymbol{\Delta}\|_1,\\
&\text{and } 2^{i-3}B_3 \leq \frac{B_3}{4} + 80q_2^2 q_5\sqrt{\frac{\log p}{n_j}}\|\boldsymbol{\Delta}\|_1 \leq 2^{i-2}B_3\big)\\
\leq&\sum_{i=1}^{\infty} P\big(Q_3(t_i) \geq 2^{i-3}B_3\big)\\
=&\sum_{i=1}^{\infty} P\big(Q_3(t_i) \geq \frac{B_3}{8} + 40q_2^2 q_5\sqrt{\frac{\log p}{n_j}}t_i\big)\\
\leq&\sum_{i=1}^{\infty} \exp(-q_4 n_j - t_i^2 \log p)\\
\leq& q_6\exp(-q_7 n_j),
\end{aligned}
\tag{65}
$$

where the last inequality follows from sum of geometric series. Define $g_3 = B_3/8$ and $g_9 = 40q_2^2 q_5$. Then by (54), (65) and the triangle inequality, we can obtain that with probability at least $1 - q_6\exp(-q_7 n_j)$,

$$
\frac{2}{n_j}\sum_{i=1}^{\frac{n_j}{2}} Q_{q_2}\{\boldsymbol{X}_i^{(j)\top}\boldsymbol{\Delta} I(|Y_i^{(j)} - \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}_\tau^*| \leq q_1)\} \geq 2g_3 - 2g_9\sqrt{\frac{\log p}{n_j}}||\boldsymbol{\Delta}||_1,
$$

for all $\|\boldsymbol{\Delta}\|_2 = 1$. Let $g_1 = q_7/(2B_4)$ and $g_4 = q_7/3$. When $\log p/n_j \leq B_4$, it is easy to show $q_6\exp(-q_7 n_j) \leq \exp(-g_4 n_j - g_1\log p)/2$. Then we have that with probability at least $1 - \exp(-g_4 n_j - g_1 \log p)/2$,

$$
\frac{2}{n_j}\sum_{i=1}^{\frac{n_j}{2}} Q_{q_2}\{\boldsymbol{X}_i^{(j)\top}\boldsymbol{\Delta} I(|Y_i^{(j)} - \boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}_\tau^*| \leq q_1)\} \geq 2g_3 - 2g_9\sqrt{\frac{\log p}{n_j}}||\boldsymbol{\Delta}||_1, \tag{66}
$$

for all $\|\boldsymbol{\Delta}\|_2 = 1$. If $\log p/n_j > (\log p)^{\alpha_2}$, then $2g_3 - 2g_9\sqrt{\log p/n_j}||\boldsymbol{\Delta}||_1 < 0$. This implies that (66) is also satisfied. By (66), we have that with probability at least $1 - \exp(-g_4 n_j - g_1\log p)/2$,

$$
l_1^{(j)}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - l_1^{(j)}(\boldsymbol{\beta}^*) - \boldsymbol{\Delta}^\top \triangledown l_1^{(j)}(\boldsymbol{\beta}^*) \geq g_3||\boldsymbol{\Delta}||_2^2 - g_9\sqrt{\frac{\log p}{n_j}}||\boldsymbol{\Delta}||_1||\boldsymbol{\Delta}||_2, \tag{67}
$$

for all $||\mathbf{\Delta}||_2 \leq 1$. Similarly, we can prove that with probability at least $1 - \exp(-g_4 n_j - g_1 \log p)/2$,

$$l_2^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_2^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top \nabla l_2^{(j)}(\boldsymbol{\beta}^*) \geq g_3 ||\mathbf{\Delta}||_2^2 - g_9 \sqrt{\frac{\log p}{n_j}} ||\mathbf{\Delta}||_1 ||\mathbf{\Delta}||_2, \tag{68}$$

for all $||\mathbf{\Delta}||_2 \leq 1$. Both (67) and (68) suggest that with probability at least $1 - \exp(-g_4 n_j - g_1 \log p)$,

$$l_1^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_1^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top \nabla l_1^{(j)}(\boldsymbol{\beta}^*) \geq g_3 ||\mathbf{\Delta}||_2^2 - g_9 \sqrt{\frac{\log p}{n_j}} ||\mathbf{\Delta}||_1 ||\mathbf{\Delta}||_2,$$

and

$$l_2^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_2^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top \nabla l_2^{(j)}(\boldsymbol{\beta}^*) \geq g_3 ||\mathbf{\Delta}||_2^2 - g_9 \sqrt{\frac{\log p}{n_j}} ||\mathbf{\Delta}||_1 ||\mathbf{\Delta}||_2,$$

for all $||\mathbf{\Delta}||_2 \leq 1$. We complete the proof of Lemma 14. ■

Lemma 15 *Assume that conditions (C1), (C5), (E2) and (E4) hold. Then there exist four positive constants g_1', g_2', g_3' and g_7' depending on M_2, B_1 and B_4 such that for any $1 \leq j \leq m$, with probability at least $1 - \exp(-g_3' n_j - g_1' \log p)$,*

$$l_1^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_1^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top \nabla l_1^{(j)}(\boldsymbol{\beta}^*) \geq g_2' ||\mathbf{\Delta}||_2^2 - g_7' \sqrt{\frac{\log p}{n_j}} ||\mathbf{\Delta}||_1 ||\mathbf{\Delta}||_2,$$

and

$$l_2^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_2^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top \nabla l_2^{(j)}(\boldsymbol{\beta}^*) \geq g_2' ||\mathbf{\Delta}||_2^2 - g_7' \sqrt{\frac{\log p}{n_j}} ||\mathbf{\Delta}||_1 ||\mathbf{\Delta}||_2,$$

for all $||\mathbf{\Delta}||_2 \leq 1$.

Proof Applying the second-order Taylor expansion, we can show that there exists a number $x_0 \in [0, 1]$ such that

$$l_1^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_1^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top \nabla l_1^{(j)}(\boldsymbol{\beta}^*) = \frac{2}{n_j} \sum_{i=1}^{\frac{n_j}{2}} Q_1'(\boldsymbol{X}_i^{(j)\top} \boldsymbol{\beta}^* + x_0 \mathbf{\Delta}^\top \boldsymbol{X}_i^{(j)})(\mathbf{\Delta}^\top \boldsymbol{X}_i^{(j)})^2, \tag{69}$$

where $Q_1'(x) = e^x/(1+e^x)^2$. Let $q_1' \geq q_2'$ be two positive numbers and $q_3' = \min_{|x| \leq 2q_1'} Q_1'(x)$. Now we show

$$Q_1'(\boldsymbol{X}_i^{(j)\top} \boldsymbol{\beta}^* + x_0 \mathbf{\Delta}^\top \boldsymbol{X}_i^{(j)})(\mathbf{\Delta}^\top \boldsymbol{X}_i^{(j)})^2 \geq q_3' Q_{q_2' \|\mathbf{\Delta}\|_2} \{\boldsymbol{X}_i^{(j)\top} \mathbf{\Delta} I(|\boldsymbol{X}_i^{(j)\top} \boldsymbol{\beta}^*| \leq q_1')\}, \tag{70}$$

for all $||\mathbf{\Delta}||_2 \leq 1$. When $|\boldsymbol{X}_i^{(j)\top}\mathbf{\Delta}| > q_2'\|\mathbf{\Delta}\|_2$ or $|\boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}^*| > q_1'$, the right hand side of (70) is 0. Since the left hand side of (70) is nonnegative, (70) is satisfied. If $|\boldsymbol{X}_i^{(j)\top}\mathbf{\Delta}| \leq q_2'\|\mathbf{\Delta}\|_2$ and $|\boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}^*| \leq q_1'$, we have

$$|\boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}^* + x_0\mathbf{\Delta}^\top\boldsymbol{X}_i^{(j)}| \leq |\boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}^*| + |\mathbf{\Delta}^\top\boldsymbol{X}_i^{(j)}| \leq q_1' + q_2' \leq 2q_1', \tag{71}$$

for all $||\mathbf{\Delta}||_2 \leq 1$. It can be shown that

$$Q_{q_2'\|\mathbf{\Delta}\|_2}\{\boldsymbol{X}_i^{(j)\top}\mathbf{\Delta}I(|\boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}^*| \leq q_1')\} \leq (\mathbf{\Delta}^\top\boldsymbol{X}_i^{(j)})^2. \tag{72}$$

Both (71) and (72) imply that (70) is also satisfied. Using (69) and (70), we can obtain

$$l_1^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_1^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top\triangledown l_1^{(j)}(\boldsymbol{\beta}^*) \geq \frac{2q_3'}{n_j}\sum_{i=1}^{\frac{n_j}{2}} Q_{q_2'\|\mathbf{\Delta}\|_2}\{\boldsymbol{X}_i^{(j)\top}\mathbf{\Delta}I(|\boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}^*| \leq q_1')\}, \tag{73}$$

for all $||\mathbf{\Delta}||_2 \leq 1$. Similarly, one can prove

$$l_2^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_2^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top\triangledown l_2^{(j)}(\boldsymbol{\beta}^*) \geq \frac{2q_3'}{n_j}\sum_{i=\frac{n_j}{2}+1}^{n_j} Q_{q_2'\|\mathbf{\Delta}\|_2}\{\boldsymbol{X}_i^{(j)\top}\mathbf{\Delta}I(|\boldsymbol{X}_i^{(j)\top}\boldsymbol{\beta}^*| \leq q_1')\}, \tag{74}$$

for all $||\mathbf{\Delta}||_2 \leq 1$. In light of (73) and (74), similar to the proof of Lemma 14, we can show that there exist four positive constants g_1', g_2', g_3' and g_7' depending on M_2, B_1 and B_4 such that for any $1 \leq j \leq m$, with probability at least $1 - \exp(-g_3'n_j - g_1'\log p)$,

$$l_1^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_1^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top\triangledown l_1^{(j)}(\boldsymbol{\beta}^*) \geq g_2'||\mathbf{\Delta}||_2^2 - g_7'\sqrt{\frac{\log p}{n_j}}||\mathbf{\Delta}||_1||\mathbf{\Delta}||_2,$$

and

$$l_2^{(j)}(\boldsymbol{\beta}^* + \mathbf{\Delta}) - l_2^{(j)}(\boldsymbol{\beta}^*) - \mathbf{\Delta}^\top\triangledown l_2^{(j)}(\boldsymbol{\beta}^*) \geq g_2'||\mathbf{\Delta}||_2^2 - g_7'\sqrt{\frac{\log p}{n_j}}||\mathbf{\Delta}||_1||\mathbf{\Delta}||_2,$$

for all $||\mathbf{\Delta}||_2 \leq 1$. The proof of Lemma 15 is completed. ■

B.1 Proof of Corollary 6

Proof Under conditions (C1) and (E1), similar to the proof of Proposition 2, we can show that for any $\tau > e_1$, there exists some non-zero constant k_τ depending on $\rho_\tau(Y, \boldsymbol{X}^\top\boldsymbol{\beta})$ such that $\boldsymbol{\beta}_\tau^* = k_\tau\boldsymbol{\beta}_0$. Let $\zeta_1 = \epsilon_\tau I(|\epsilon_\tau| \leq \tau) + \tau\mathrm{sgn}(\epsilon_\tau)I(|\epsilon_\tau| > \tau)$. It is straightforward to show

$$\boldsymbol{Z} = -\boldsymbol{X}\zeta_1.$$

Based on the fact that $|\zeta_1| \leq \tau$ and condition (E2), we can prove

$$\|\boldsymbol{Z}\|_{\psi_2} = \|\boldsymbol{X}\zeta_1\|_{\psi_2} \leq \tau B_1. \tag{75}$$

By Lemma 14, we have that for any $\tau \geq g_2$ and $1 \leq j \leq m$, with probability at least $1 - \exp(-g_4 n_j - g_1 \log p)$,

$$l_1^{(j)}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - l_1^{(j)}(\boldsymbol{\beta}^*) - \boldsymbol{\Delta}^\top \nabla l_1^{(j)}(\boldsymbol{\beta}^*) \geq g_3||\boldsymbol{\Delta}||_2^2 - g_9\sqrt{\frac{\log p}{n_j}}||\boldsymbol{\Delta}||_1||\boldsymbol{\Delta}||_2,$$

and

$$l_2^{(j)}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - l_2^{(j)}(\boldsymbol{\beta}^*) - \boldsymbol{\Delta}^\top \nabla l_2^{(j)}(\boldsymbol{\beta}^*) \geq g_3||\boldsymbol{\Delta}||_2^2 - g_9\sqrt{\frac{\log p}{n_j}}||\boldsymbol{\Delta}||_1||\boldsymbol{\Delta}||_2, \tag{76}$$

for all $||\boldsymbol{\Delta}||_2 \leq 1$. It is sufficient to show that condition (C7) is satisfied by mathematical induction. Let $\lambda_1 = c'_{11}\sqrt{\log p/n_1}$, $\gamma_1 = c'_{21}\sqrt{\log p/n_1}$ and $d'_1 = \max\{3a'_2/g_3, 4\}$, where c'_{11} and c'_{21} could be any constants which belong to $[2\tau B_1\sqrt{2(a'_0+1)/a_1}, a'_2]$. Using (75) and (76), similar to (22), we can show that with probability at least $1 - \exp(-g_4 n_1 - g_1 \log p) - 2ep^{-a'_0}$,

$$\begin{aligned}
&||\hat{\boldsymbol{\Delta}}_1^{(1)}||_2 \leq d'_1\sqrt{\frac{s_0 \log p}{n_1}}, \quad ||\hat{\boldsymbol{\Delta}}_1^{(1)}||_1 \leq d_1'^2 s_0\sqrt{\frac{\log p}{n_1}},\\
&||\hat{\boldsymbol{\Delta}}_2^{(1)}||_2 \leq d'_1\sqrt{\frac{s_0 \log p}{n_1}}, \quad ||\hat{\boldsymbol{\Delta}}_2^{(1)}||_1 \leq d_1'^2 s_0\sqrt{\frac{\log p}{n_1}},\\
&||\hat{\boldsymbol{\beta}}_{ave}^{(1)} - \boldsymbol{\beta}^*||_2 \leq d'_1\sqrt{\frac{s_0 \log p}{n_1}}, \quad \text{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(1)} - \boldsymbol{\beta}^*||_1 \leq d_1'^2 s_0\sqrt{\frac{\log p}{n_1}}.
\end{aligned} \tag{77}$$

By conditions (E2) and (E5), the Cauchy-Schwarz inequality and (77), one can prove that with probability at least $1 - \exp(-g_4 n_1 - g_1 \log p) - ep^{-a'_0}$,

$$\begin{aligned}
&\|E(\boldsymbol{H}_1^{(1)}|\hat{\boldsymbol{\beta}}_2^{(1)}) - \boldsymbol{H}_\tau\|_\infty\\
=&\|E\{\boldsymbol{X}_1^{(1)}\boldsymbol{X}_1^{(1)\top}I(|Y_1^{(1)} - \boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)}| \leq \tau)|\hat{\boldsymbol{\beta}}_2^{(1)}\}\\
&- E\{\boldsymbol{X}_1^{(1)}\boldsymbol{X}_1^{(1)\top}I(|Y_1^{(1)} - \boldsymbol{X}_1^{(1)\top}\boldsymbol{\beta}_\tau^*| \leq \tau)\}\|_\infty\\
\leq& \max_{\substack{1\leq j\leq p\\1\leq k\leq p}} E\{L_\tau|X_{1,j}^{(1)}X_{1,k}^{(1)}||\boldsymbol{X}_1^{(1)\top}(\hat{\boldsymbol{\beta}}_2^{(1)} - \boldsymbol{\beta}_\tau^*)||\hat{\boldsymbol{\beta}}_2^{(1)}\}\\
\leq& \max_{\substack{1\leq j\leq p\\1\leq k\leq p}} L_\tau\{EX_{1,j}^{(1)2}X_{1,k}^{(1)2}\}^{1/2}(E[\{\boldsymbol{X}_1^{(1)\top}(\hat{\boldsymbol{\beta}}_2^{(1)} - \boldsymbol{\beta}_\tau^*)\}^2|\hat{\boldsymbol{\beta}}_2^{(1)}])^{1/2}\\
\leq& \max_{\substack{1\leq j\leq p\\1\leq k\leq p}} L_\tau\{EX_{1,j}^{(1)4}\}^{1/4}\{EX_{1,k}^{(1)4}\}^{1/4}(E[\{\boldsymbol{X}_1^{(1)\top}(\hat{\boldsymbol{\beta}}_2^{(1)} - \boldsymbol{\beta}_\tau^*)\}^2|\hat{\boldsymbol{\beta}}_2^{(1)}])^{1/2}\\
\leq& 4\sqrt{2}L_\tau B_1^3\|\hat{\boldsymbol{\beta}}_2^{(1)} - \boldsymbol{\beta}_\tau^*\|_2\\
\leq& 4\sqrt{2}L_\tau B_1^3 d'_1\sqrt{\frac{s_0 \log p}{n_1}},
\end{aligned} \tag{78}$$

where $X_{1,j}^{(1)}$ and $X_{1,k}^{(1)}$ are the jth and kth elements of $\boldsymbol{X}_1^{(1)}$, respectively. For any random variable ξ', let $\|\xi'|\hat{\boldsymbol{\beta}}_2^{(1)}\|_{\psi_1} = \sup_{l\geq 1}(E|\xi'|^l|\hat{\boldsymbol{\beta}}_2^{(1)})^{1/l}/l$. For any $1 \leq j, k \leq p$ and $l \geq 1$, using the Cauchy-Schwarz inequality and condition (E2), we have

$$[E\{|X_{1,j}^{(1)l}X_{1,k}^{(1)l}|I(|Y_1^{(1)} - \boldsymbol{X}_1^{(1)^\top}\hat{\boldsymbol{\beta}}_2^{(1)}| \leq \tau)|\hat{\boldsymbol{\beta}}_2^{(1)}\}]^{1/l}/l \leq (EX_{1,j}^{(1)2l})^{1/2l}(EX_{1,k}^{(1)2l})^{1/2l}/l \leq 2B_1^2.$$

This implies

$$\|X_{1,j}^{(1)}X_{1,k}^{(1)}I(|Y_1^{(1)} - \boldsymbol{X}_1^{(1)^\top}\hat{\boldsymbol{\beta}}_2^{(1)}| \leq \tau)|\hat{\boldsymbol{\beta}}_2^{(1)}\|_{\psi_1} \leq 2B_1^2.$$

Then by the triangle inequality, the Cauchy-Schwarz inequality and condition (E2), we have

$$\begin{aligned}
&\|X_{1,j}^{(1)}X_{1,k}^{(1)}I(|Y_1^{(1)} - \boldsymbol{X}_1^{(1)^\top}\hat{\boldsymbol{\beta}}_2^{(1)}| \leq \tau) - E\{X_{1,j}^{(1)}X_{1,k}^{(1)}I(|Y_1^{(1)} - \boldsymbol{X}_1^{(1)^\top}\hat{\boldsymbol{\beta}}_2^{(1)}| \leq \tau)|\hat{\boldsymbol{\beta}}_2^{(1)}\}|\hat{\boldsymbol{\beta}}_2^{(1)}\|_{\psi_1}\\
\leq&\|X_{1,j}^{(1)}X_{1,k}^{(1)}I(|Y_1^{(1)} - \boldsymbol{X}_1^{(1)^\top}\hat{\boldsymbol{\beta}}_2^{(1)}| \leq \tau)|\hat{\boldsymbol{\beta}}_2^{(1)}\|_{\psi_1}\\
&+|E\{X_{1,j}^{(1)}X_{1,k}^{(1)}I(|Y_1^{(1)} - \boldsymbol{X}_1^{(1)^\top}\hat{\boldsymbol{\beta}}_2^{(1)}| \leq \tau)\}|\hat{\boldsymbol{\beta}}_2^{(1)}|\\
\leq&2B_1^2 + (EX_{1,j}^{(1)2})^{1/2}(EX_{1,k}^{(1)2})^{1/2}\\
\leq&4B_1^2. \qquad (79)
\end{aligned}$$

Let

$$\begin{aligned}
&\tilde{\zeta}_{j,k}\\
=&\frac{2}{n_1}\sum_{i=1}^{n_1/2}X_{i,j}^{(1)}X_{i,k}^{(1)}I(|Y_1^{(1)} - \boldsymbol{X}_1^{(1)^\top}\hat{\boldsymbol{\beta}}_2^{(1)}| \leq \tau) - E\{X_{1,j}^{(1)}X_{1,k}^{(1)}I(|Y_1^{(1)} - \boldsymbol{X}_1^{(1)^\top}\hat{\boldsymbol{\beta}}_2^{(1)}| \leq \tau)|\hat{\boldsymbol{\beta}}_2^{(1)}\}.
\end{aligned}$$

For any $x > 0$, according to (79), a Bernstein-type inequality (Vershynin, 2012, Proposition 5.16) and the union inequality, we can show

$$\begin{aligned}
P(\|\boldsymbol{H}_1^{(1)} - E(\boldsymbol{H}_1^{(1)}|\hat{\boldsymbol{\beta}}_2^{(1)})\|_\infty \geq x|\hat{\boldsymbol{\beta}}_2^{(1)}) \leq& p^2\max_{\substack{1\leq j\leq p\\1\leq k\leq p}}P(|\tilde{\zeta}_{j,k}| \geq x|\hat{\boldsymbol{\beta}}_2^{(1)})\\
\leq&2p^2\exp\{-a_4'\min(\frac{x^2n_1}{32B_1^4}, \frac{xn_1}{8B_1^2})\}, \qquad (80)
\end{aligned}$$

where a_4' is a positive constant not depending on any parameter. Let

$$x = \max\{\sqrt{32B_1^4(a_0'+2)/a_4'}, 8B_1^2(a_0'+2)/a_4'\}\sqrt{\log p/n_1}.$$

Then we have

$$P(\|\boldsymbol{H}_1^{(1)} - E(\boldsymbol{H}_1^{(1)}|\hat{\boldsymbol{\beta}}_2^{(1)})\|_\infty \geq x|\hat{\boldsymbol{\beta}}_2^{(1)}) \leq 2p^2\exp\{-a_4'\min(\frac{x^2n_1}{32B_1^4}, \frac{xn_1}{8B_1^2})\} \leq 2p^{-a_0'}.$$

It follows from the Law of Total Probability that

$$\begin{aligned}
&P\big(\|\boldsymbol{H}_1^{(1)} - E(\boldsymbol{H}_1^{(1)}|\hat{\boldsymbol{\beta}}_2^{(1)})\|_\infty \geq \max\{\sqrt{32B_1^4(a_0'+2)/a_4'}, 8B_1^2(a_0'+2)/a_4'\}\sqrt{\log p/n_1}\big)\\
\leq&2p^{-a_0'}. \qquad (81)
\end{aligned}$$

Using (78), (81) and the triangle inequality, we can obtain that with probability at least $1 - \exp(-g_4 n_1 - g_1 \log p) - (2 + e)p^{-a_0'}$,

$$\begin{aligned}\|\boldsymbol{H}_1^{(1)} - \boldsymbol{H}_\tau\|_\infty \leq & \max\{\sqrt{32B_1^4(a_0'+2)a_4'}, 8B_1^2(a_0'+2)a_4'\}\sqrt{\log p/n_1} \\ & + 4\sqrt{2}L_\tau B_1^3 d_1' \sqrt{\frac{s_0 \log p}{n_1}} \\ \leq & M_\tau \sqrt{\frac{s_0 \log p}{n_1}},\end{aligned} \tag{82}$$

where

$$M_\tau = [\max\{\sqrt{32B_1^4(a_0'+2)/a_4'}, 8B_1^2(a_0'+2)/a_4'\} + 4\sqrt{2}L_\tau B_1^3 + 1]a_3' d_1',$$

and

$$a_3' = \max\{(2B_2 + 3a_2'/2)/\min\{B_3/3, g_3/2\}, 8 + 2B_2/\{\tau B_1\sqrt{2(a_0+1)/a_1}\}\}.$$

Similarly, one can show that with probability at least $1 - \exp(-g_4 n_1 - g_1 \log p) - (2+e)p^{-a_0'}$,

$$\|\boldsymbol{H}_2^{(1)} - \boldsymbol{H}_\tau\|_\infty \leq M_\tau \sqrt{\frac{s_0 \log p}{n_1}}. \tag{83}$$

Both (82) and (83) indicate that with probability at least $1 - \exp(-g_4 n_1 - g_1 \log p) - (4 + 2e)p^{-a_0'}$,

$$\max\{\|\boldsymbol{H}_1^{(1)} - \boldsymbol{H}_\tau\|_\infty, \|\boldsymbol{H}_2^{(1)} - \boldsymbol{H}_\tau\|_\infty\} \leq M_\tau \sqrt{\frac{s_0 \log p}{n_1}}. \tag{84}$$

Assume that with probability at least $1 - 4(s-2)p^{-a_0'} - \sum_{j=1}^{s-1}\{\exp(-g_4 n_j - g_1 \log p) + 2ep^{-a_0' N_j/n_j}\}$,

$$\begin{aligned}& ||\hat{\boldsymbol{\Delta}}_1^{(s-1)}||_2 \leq d_{s-1}' \sqrt{\frac{s_0 \log p}{N_{s-1}}}, \quad ||\hat{\boldsymbol{\Delta}}_1^{(s-1)}||_1 \leq d_{s-1}'^2 s_0 \sqrt{\frac{\log p}{N_{s-1}}}, \\ & ||\hat{\boldsymbol{\Delta}}_2^{(s-1)}||_2 \leq d_{s-1}' \sqrt{\frac{s_0 \log p}{N_{s-1}}}, \quad ||\hat{\boldsymbol{\Delta}}_2^{(s-1)}||_1 \leq d_{s-1}'^2 s_0 \sqrt{\frac{\log p}{N_{s-1}}}, \\ & ||\hat{\boldsymbol{\beta}}_{ave}^{(s-1)} - \boldsymbol{\beta}_\tau^*||_2 \leq d_{s-1}' \sqrt{\frac{s_0 \log p}{N_2}}, \quad \text{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(s-1)} - \boldsymbol{\beta}_\tau^*||_1 \leq d_{s-1}'^2 s_0 \sqrt{\frac{\log p}{N_{s-1}}},\end{aligned} \tag{85}$$

and with probability at least $1 - 4(s-1)p^{-a_0'} - \sum_{j=1}^{s-1}\{\exp(-g_4 n_j - g_1 \log p) + 2ep^{-a_0' N_j/n_j}\}$,

$$\begin{aligned}& \max\{\|\frac{1}{N_{s-1}}\sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)} - \boldsymbol{H}_\tau\|_\infty, \|\frac{1}{N_{s-1}}\sum_{j=1}^{s-1} n_j \boldsymbol{H}_2^{(j)} - \boldsymbol{H}_\tau\|_\infty\} \\ \leq & \frac{1}{N_{s-1}}\sum_{j=1}^{s-1} n_j M_\tau^j \max\{\sqrt{\frac{s_0 \log p}{n_j}}, \sqrt{s_0}\frac{\log p}{n_j}\},\end{aligned} \tag{86}$$

where $d'_{s-1} = a_3'^{s-2} d'_1$. Let $\lambda_s = c'_{1s}\sqrt{\log p/N_s}$ and $\gamma_s = c'_{2s}\sqrt{\log p/N_s}$, where c'_{1s} and c'_{2s} could be any constants which belong to $[2\tau B_1\sqrt{2(a'_0+1)/a_1}, a'_2]$. According to (75), (76), (85), (86), and conditions (C4) and (E3), similar to (45), one can prove that with probability at least $1-4(s-1)p^{-a'_0} - \sum_{j=1}^{s}\{\exp(-g_4 n_j - g_1\log p) + 2ep^{-a'_0 N_j/n_j}\}$,

$$
\begin{aligned}
&||\hat{\boldsymbol{\Delta}}_1^{(s)}||_2 \le d'_s\sqrt{\frac{s_0\log p}{N_s}}, \quad ||\hat{\boldsymbol{\Delta}}_1^{(s)}||_1 \le d_s'^2 s_0\sqrt{\frac{\log p}{N_s}},\\
&||\hat{\boldsymbol{\Delta}}_2^{(s)}||_2 \le d'_s\sqrt{\frac{s_0\log p}{N_s}}, \quad ||\hat{\boldsymbol{\Delta}}_2^{(s)}||_1 \le d_s'^2 s_0\sqrt{\frac{\log p}{N_s}},\\
&||\hat{\boldsymbol{\beta}}_{ave}^{(s)} - \boldsymbol{\beta}^*||_2 \le d'_s\sqrt{\frac{s_0\log p}{N_s}}, \quad \text{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(s)} - \boldsymbol{\beta}^*||_1 \le d_s'^2 s_0\sqrt{\frac{\log p}{N_s}},
\end{aligned}
\tag{87}
$$

where $d'_s = a_3'^{s-1}d'_1$. Similar to (78), we can show that with probability at least $1-4(s-1)p^{-a'_0} - \sum_{j=1}^{s}\{\exp(-g_4 n_j - g_1\log p) + 2ep^{-a'_0 N_j/n_j}\}$,

$$
\|E(\boldsymbol{H}_1^{(s)}|\hat{\boldsymbol{\beta}}_2^{(s)}) - \boldsymbol{H}_\tau\|_\infty \le 4\sqrt{2}L_\tau B_1^3 d'_s\sqrt{\frac{s_0\log p}{n_1}}. \tag{88}
$$

Similar to (80), for any $x>0$, we can show

$$
P(\|\boldsymbol{H}_1^{(s)} - E(\boldsymbol{H}_1^{(s)}|\hat{\boldsymbol{\beta}}_2^{(s)})\|_\infty \ge x|\hat{\boldsymbol{\beta}}_2^{(s)}) \le 2p^2\exp\{-a'_4\min(\frac{x^2 n_s}{32B_1^4}, \frac{x n_s}{8B_1^2})\}.
$$

Let $x = \max\{\sqrt{32B_1^4(a'_0+2)/a'_4}, 8B_1^2(a'_0+2)/a'_4\}\max\{\sqrt{\log p/n_s}, \log p/n_s\}$. Then similar to (81), we can obtain

$$
P\big(\|\boldsymbol{H}_1^{(s)} - E(\boldsymbol{H}_1^{(s)}|\hat{\boldsymbol{\beta}}_2^{(s)})\|_\infty \ge x\big) \le 2p^{-a'_0}. \tag{89}
$$

By (88), (89) and the triangle inequality, we can obtain that with probability at least $1-2p^{-a'_0}-4(s-1)p^{-a'_0} - \sum_{j=1}^{s}\{\exp(-g_4 n_j - g_1\log p) + 2ep^{-a'_0 N_j/n_j}\}$,

$$
\|\boldsymbol{H}_1^{(s)} - \boldsymbol{H}_\tau\|_\infty \le M_\tau^s\max\{\sqrt{\frac{s_0\log p}{n_s}}, \sqrt{s_0}\frac{\log p}{n_s}\}. \tag{90}
$$

Similarly, one can show that with probability at least $1-2p^{-a'_0}-4(s-1)p^{-a'_0} - \sum_{j=1}^{s}\{\exp(-g_4 n_j - g_1\log p) + 2ep^{-a'_0 N_j/n_j}\}$,

$$
\|\boldsymbol{H}_2^{(s)} - \boldsymbol{H}_\tau\|_\infty \le M_\tau^s\max\{\sqrt{\frac{s_0\log p}{n_s}}, \sqrt{s_0}\frac{\log p}{n_s}\}. \tag{91}
$$

Both (90) and (91) imply that with probability at least $1-4sp^{-a'_0} - \sum_{j=1}^{s}\{\exp(-g_4 n_j - g_1\log p) + 2ep^{-a'_0 N_j/n_j}\}$,

$$
\max\{\|\boldsymbol{H}_1^{(s)} - \boldsymbol{H}_\tau\|_\infty, \|\boldsymbol{H}_2^{(s)} - \boldsymbol{H}_\tau\|_\infty\} \le M_\tau^s\max\{\sqrt{\frac{s_0\log p}{n_s}}, \sqrt{s_0}\frac{\log p}{n_s}\}. \tag{92}
$$

It follows from (86), (92) and the triangle inequality that with probability at least $1 - 4sp^{-a_0'} - \sum_{j=1}^{s}\{\exp(-g_4 n_j - g_1 \log p) + 2ep^{-a_0' N_j/n_j}\}$,

$$\begin{aligned}
&\max\{\|\frac{1}{N_s}\sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)} - \boldsymbol{H}_\tau\|_\infty, \|\frac{1}{N_s}\sum_{j=1}^{s} n_j \boldsymbol{H}_2^{(j)} - \boldsymbol{H}_\tau\|_\infty\} \\
\leq&\max\{\frac{N_{s-1}}{N_s}\|\frac{1}{N_{s-1}}\sum_{j=1}^{s-1} n_j \boldsymbol{H}_1^{(j)} - \boldsymbol{H}_\tau\|_\infty + \frac{n_s}{N_s}\|\boldsymbol{H}_1^{(s)} - \boldsymbol{H}_\tau\|_\infty, \frac{N_{s-1}}{N_s}\|\frac{1}{N_{s-1}}\sum_{j=1}^{s-1} n_j \boldsymbol{H}_2^{(j)} \\
&- \boldsymbol{H}_\tau\|_\infty + \frac{n_s}{N_s}\|\boldsymbol{H}_2^{(s)} - \boldsymbol{H}_\tau\|_\infty\}, \\
\leq&\frac{1}{N_s}\sum_{j=1}^{s} n_j M_\tau^j \max\{\sqrt{\frac{s_0 \log p}{n_j}}, \sqrt{s_0}\frac{\log p}{n_j}\}.
\end{aligned}$$

The proof of Corollary 6 is completed. ■

B.2 Proof of Corollary 8

Proof It is sufficient to show that conditions (D2) and (D3) are satisfied. By Corollary 6, we can show that for any $1 \leq s \leq m$, with probability at least $1 - 4sp^{-a_0'} - \sum_{j=1}^{s}\{\exp(-g_4 n_j - g_1 \log p) + 2ep^{-a_0' N_j/n_j}\}$,

$$\begin{aligned}
&\max\{\|\frac{1}{N_s}\sum_{j=1}^{s} n_j \boldsymbol{H}_1^{(j)} - \boldsymbol{H}_\tau\|_\infty, \|\frac{1}{N_s}\sum_{j=1}^{s} n_j \boldsymbol{H}_2^{(j)} - \boldsymbol{H}_\tau\|_\infty\} \\
\leq&\frac{1}{N_s}\sum_{j=1}^{s} n_j M_\tau^j \max\{\sqrt{\frac{s_0 \log p}{n_j}}, \sqrt{s_0}\frac{\log p}{n_j}\} \\
\leq& s M_\tau^s \sqrt{s_0}\sqrt{\frac{\log p}{N_s}} \\
\leq&\min\{\frac{h_s}{\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty}}, \frac{\kappa_s}{\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty}}\}. \qquad (93)
\end{aligned}$$

By appealing to (93) and condition (E11), and following the proof of Theorem 6 of Cai et al. (2011), we can prove that for any $1 \leq s \leq m$, with probability at least $1 - 4sp^{-a_0'} - \sum_{j=1}^{s}\{\exp(-g_4 n_j - g_1 \log p) + 2ep^{-a_0' N_j/n_j}\}$,

$$\begin{aligned}
&\max\{\|\hat{\boldsymbol{\Omega}}_1^{(s)} - \boldsymbol{\Omega}_\tau\|_{\infty,\infty}, \|\hat{\boldsymbol{\Omega}}_2^{(s)} - \boldsymbol{\Omega}_\tau\|_{\infty,\infty}\} \\
\leq& 12v(p)\max\{(4\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty} h_s)^{1-\omega}, (4\|\boldsymbol{\Omega}_\tau\|_{\infty,\infty}\kappa_s)^{1-\omega}\}.
\end{aligned}$$

Based on condition (E6), for any $1 \leq s \leq m$, we can show

$$\lim_{p\to\infty} 1 - 4sp^{-a_0'} - \sum_{j=1}^{s}\{\exp(-g_4 n_j - g_1 \log p) + 2ep^{-a_0' N_j/n_j}\} = 1.$$

Then for any $1 \leq s \leq m$, we have

$$\max\{\|\hat{\mathbf{\Omega}}_1^{(s)} - \mathbf{\Omega}_\tau\|_{\infty,\infty}, \|\hat{\mathbf{\Omega}}_2^{(s)} - \mathbf{\Omega}_\tau\|_{\infty,\infty}\} = O_p((\|\mathbf{\Omega}_\tau\|_{\infty,\infty}^4 s^2 M_\tau^{2s} s_0 \log p/N_s)^{(1-\omega)/2} v(p)).$$

For any $1 \leq s \leq m$, in light of Corollary 6, conditions (C4) and (E12), similar to (78), we can show

$$\begin{aligned}
&\|E\big(\{\sum_{j=1}^s n_j \boldsymbol{H}_1^{(j)}(\boldsymbol{\beta}_\tau^* - \hat{\boldsymbol{\beta}}_2^{(j)}) + \sum_{j=1}^s n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}) - \sum_{j=1}^s n_j \nabla l_1^{(j)}(\boldsymbol{\beta}_\tau^*)\}/N_s | \hat{\boldsymbol{\beta}}_2^{(1)}, \ldots, \hat{\boldsymbol{\beta}}_2^{(s)}\big)\|_\infty \\
=&O_p(\frac{1}{N_s}\sum_{j=1}^s n_j d_j'^2 \frac{\log p}{N_j}) \\
=&O_p(n_1^{\alpha_1}\frac{\log p}{n_1^{\alpha_1}}\frac{1}{N_s}\sum_{j=1}^s \frac{n_j}{N_j} d_j'^2) \\
=&o_p(n_1^{\alpha_1} N_s^{-1} s d_s'^2) \\
=&o_p(s d_s'^2 N_s^{\alpha_1 - 1}) \\
=&o_p(N_s^{-1/2}\|\mathbf{\Omega}_\tau\|_{\infty,\infty}^{-1}).
\end{aligned} \tag{94}$$

For any $1 \leq s \leq m$, using Corollary 6, conditions (E2) and (E12), similar to (81), one can prove

$$\begin{aligned}
&\|\{\sum_{j=1}^s n_j \boldsymbol{H}_1^{(j)}(\boldsymbol{\beta}_\tau^* - \hat{\boldsymbol{\beta}}_2^{(j)}) + \sum_{j=1}^s n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}) - \sum_{j=1}^s n_j \nabla l_1^{(j)}(\boldsymbol{\beta}_\tau^*)\}/N_s \\
&-E\big(\{\sum_{j=1}^s n_j \boldsymbol{H}_1^{(j)}(\boldsymbol{\beta}_\tau^* - \hat{\boldsymbol{\beta}}_2^{(j)}) + \sum_{j=1}^s n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}) - \sum_{j=1}^s n_j \nabla l_1^{(j)}(\boldsymbol{\beta}_\tau^*)\}/N_s | \hat{\boldsymbol{\beta}}_2^{(1)}, \ldots, \hat{\boldsymbol{\beta}}_2^{(s)}\big)\|_\infty \\
=&O_p(\sqrt{\frac{\log p}{N_s}} \max_{1\leq j\leq s} d_j' \sqrt{\frac{s_0 \log p}{N_j}}) \\
=&O_p(N_s^{-1/2} \max_{1\leq j\leq s} d_j' \sqrt{\frac{s_0 \log^2 p}{N_j}}) \\
=&o_p(N_s^{-1/2}\|\mathbf{\Omega}_\tau\|_{\infty,\infty}^{-1}).
\end{aligned} \tag{95}$$

Both (94) and (95) imply

$$\|\mathbf{\Omega}_\tau\|_{\infty,\infty}\|\{\sum_{j=1}^s n_j \boldsymbol{H}_1^{(j)}(\boldsymbol{\beta}_\tau^* - \hat{\boldsymbol{\beta}}_2^{(j)}) + \sum_{j=1}^s n_j \nabla l_1^{(j)}(\hat{\boldsymbol{\beta}}_2^{(j)}) - \sum_{j=1}^s n_j \nabla l_1^{(j)}(\boldsymbol{\beta}_\tau^*)\}/N_s^{1/2}\|_\infty = o_p(1).$$

Similarly, we can show

$$\|\mathbf{\Omega}_\tau\|_{\infty,\infty}\|\{\sum_{j=1}^s n_j \boldsymbol{H}_2^{(j)}(\boldsymbol{\beta}_\tau^* - \hat{\boldsymbol{\beta}}_1^{(j)}) + \sum_{j=1}^s n_j \nabla l_2^{(j)}(\hat{\boldsymbol{\beta}}_1^{(j)}) - \sum_{j=1}^s n_j \nabla l_2^{(j)}(\boldsymbol{\beta}_\tau^*)\}/N_s^{1/2}\|_\infty = o_p(1).$$

We complete the proof of Corollary 8. ∎

B.3 Proof of Corollary 10

Proof It is straightforward to verify

$$\boldsymbol{Z} = \boldsymbol{X}\zeta_2,$$

where $\zeta_2 = \exp(\boldsymbol{X}^\top \boldsymbol{\beta}^*)/\{1+\exp(\boldsymbol{X}^\top \boldsymbol{\beta}^*)\} - Y$. In light of $|\zeta_2| \leq 1$ and condition (E2), we have

$$\|\boldsymbol{Z}\|_{\psi_2} \leq B_1. \tag{96}$$

According to Lemma 15, we have that for any $1 \leq j \leq m$, with probability at least $1-\exp(-g_3' n_j - g_1' \log p)$,

$$l_1^{(j)}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - l_1^{(j)}(\boldsymbol{\beta}^*) - \boldsymbol{\Delta}^\top \nabla l_1^{(j)}(\boldsymbol{\beta}^*) \geq g_2' ||\boldsymbol{\Delta}||_2^2 - g_7' \sqrt{\frac{\log p}{n_j}} ||\boldsymbol{\Delta}||_1 ||\boldsymbol{\Delta}||_2,$$

and

$$l_2^{(j)}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - l_2^{(j)}(\boldsymbol{\beta}^*) - \boldsymbol{\Delta}^\top \nabla l_2^{(j)}(\boldsymbol{\beta}^*) \geq g_2' ||\boldsymbol{\Delta}||_2^2 - g_7' \sqrt{\frac{\log p}{n_j}} ||\boldsymbol{\Delta}||_1 ||\boldsymbol{\Delta}||_2, \tag{97}$$

for all $||\boldsymbol{\Delta}||_2 \leq 1$. Similar to the proof of Corollary 6, we only need to show that condition (C7) is satisfied by mathematical induction. Let $\lambda_1 = c_{11}'' \sqrt{\log p/n_1}$, $\gamma_1 = c_{21}'' \sqrt{\log p/n_1}$, and $d_1'' = \max\{3a_2''/g_2', 4\}$, where c_{11}'' and c_{21}'' could be any constants which belongs to $[2B_1\sqrt{2(a_0''+1)/a_1}, a_2'']$. By (96) and (97), similar to (22), we can show that with probability at least $1-\exp(-g_3' n_1 - g_1' \log p) - 2ep^{-a_0''}$,

$$\begin{aligned}
&||\hat{\boldsymbol{\Delta}}_1^{(1)}||_2 \leq d_1'' \sqrt{\frac{s_0 \log p}{n_1}}, \quad ||\hat{\boldsymbol{\Delta}}_1^{(1)}||_1 \leq d_1''^2 s_0 \sqrt{\frac{\log p}{n_1}},\\
&||\hat{\boldsymbol{\Delta}}_2^{(1)}||_2 \leq d_1'' \sqrt{\frac{s_0 \log p}{n_1}}, \quad ||\hat{\boldsymbol{\Delta}}_2^{(1)}||_1 \leq d_1''^2 s_0 \sqrt{\frac{\log p}{n_1}},\\
&||\hat{\boldsymbol{\beta}}_{ave}^{(1)} - \boldsymbol{\beta}^*||_2 \leq d_1'' \sqrt{\frac{s_0 \log p}{n_1}}, \quad \text{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(1)} - \boldsymbol{\beta}^*||_1 \leq d_1''^2 s_0 \sqrt{\frac{\log p}{n_1}}.
\end{aligned} \tag{98}$$

According to condition (E2), the mean value theorem, the Cauchy-Schwarz inequality and (98), we can obtain that with probability at least $1-\exp(-g_3'n_1-g_1'\log p)-ep^{-a_0''}$,

$$\begin{aligned}
&\|E(\boldsymbol{H}_1^{(1)}|\hat{\boldsymbol{\beta}}_2^{(1)})-\boldsymbol{H}\|_\infty\\
=&\|E\{\boldsymbol{X}_1^{(1)}\boldsymbol{X}_1^{(1)\top}\frac{\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}{1+\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}|\hat{\boldsymbol{\beta}}_2^{(1)}\}-E\{\boldsymbol{X}_1^{(1)}\boldsymbol{X}_1^{(1)\top}\frac{\exp(\boldsymbol{X}_1^{(1)\top}\boldsymbol{\beta}^*)}{1+\exp(\boldsymbol{X}_1^{(1)\top}\boldsymbol{\beta}^*)}\|_\infty\\
\le&\max_{\substack{1\le j\le p\\1\le k\le p}}E\{|X_{1,j}^{(1)}X_{1,k}^{(1)}||\boldsymbol{X}_1^{(1)\top}(\hat{\boldsymbol{\beta}}_2^{(1)}-\boldsymbol{\beta}^*)||\hat{\boldsymbol{\beta}}_2^{(1)}\}\\
\le&\max_{\substack{1\le j\le p\\1\le k\le p}}\{EX_{1,j}^{(1)2}X_{1,k}^{(1)2}\}^{1/2}(E[\{\boldsymbol{X}_1^{(1)\top}(\hat{\boldsymbol{\beta}}_2^{(1)}-\boldsymbol{\beta}^*)\}^2|\hat{\boldsymbol{\beta}}_2^{(1)}])^{1/2}\\
\le&\max_{\substack{1\le j\le p\\1\le k\le p}}\{EX_{1,j}^{(1)4}\}^{1/4}\{EX_{1,k}^{(1)4}\}^{1/4}(E[\{\boldsymbol{X}_1^{(1)\top}(\hat{\boldsymbol{\beta}}_2^{(1)}-\boldsymbol{\beta}_\tau^*)\}^2|\hat{\boldsymbol{\beta}}_2^{(1)}])^{1/2}\\
\le&4\sqrt{2}B_1^3\|\hat{\boldsymbol{\beta}}_2^{(1)}-\boldsymbol{\beta}_\tau^*\|_2\\
\le&4\sqrt{2}B_1^3d_1''\sqrt{\frac{s_0\log p}{n_1}}.
\end{aligned}\tag{99}$$

For any $1\le j,k\le p$ and $l\ge 1$, in light of the Cauchy-Schwarz inequality and condition (E2), we can obtain

$$[E\{|X_{1,j}^{(1)l}X_{1,k}^{(1)l}|\frac{\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}{1+\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}|\hat{\boldsymbol{\beta}}_2^{(1)}\}]^{1/l}/l\le(EX_{1,j}^{(1)2l})^{1/2l}(EX_{1,k}^{(1)2l})^{1/2l}/l\le 2B_1^2,$$

which implying

$$\|X_{1,j}^{(1)}X_{1,k}^{(1)}\frac{\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}{1+\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}|\hat{\boldsymbol{\beta}}_2^{(1)}\|_{\psi_1}\le 2B_1^2.$$

It follows from the triangle inequality, the Cauchy-Schwarz inequality and condition (E2) that

$$\begin{aligned}
&\|X_{1,j}^{(1)}X_{1,k}^{(1)}\frac{\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}{1+\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}-E\{X_{1,j}^{(1)}X_{1,k}^{(1)}\frac{\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}{1+\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}|\hat{\boldsymbol{\beta}}_2^{(1)}\}|\hat{\boldsymbol{\beta}}_2^{(1)}\|_{\psi_1}\\
\le&\|X_{1,j}^{(1)}X_{1,k}^{(1)}\frac{\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}{1+\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}|\hat{\boldsymbol{\beta}}_2^{(1)}\|_{\psi_1}+|E\{X_{1,j}^{(1)}X_{1,k}^{(1)}\frac{\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}{1+\exp(\boldsymbol{X}_1^{(1)\top}\hat{\boldsymbol{\beta}}_2^{(1)})}\}|\hat{\boldsymbol{\beta}}_2^{(1)}|\\
\le&2B_1^2+(EX_{1,j}^{(1)2})^{1/2}(EX_{1,k}^{(1)2})^{1/2}\\
\le&4B_1^2.
\end{aligned}\tag{100}$$

According to (100), a Bernstein-type inequality (Vershynin, 2012, Proposition 5.16), the union inequality and the Law of Total Probability, similar to (81), we have

$$\begin{aligned}
&P\big(\|\boldsymbol{H}_1^{(1)}-E(\boldsymbol{H}_1^{(1)}|\hat{\boldsymbol{\beta}}_2^{(1)})\|_\infty\ge\max\{\sqrt{32B_1^4(a_0''+2)/a_4'},8B_1^2(a_0''+2)/a_4'\}\sqrt{\log p/n_1}\big)\\
\le&2p^{-a_0''}.
\end{aligned}\tag{101}$$

By (100), (101) and the triangle inequality, one can prove that with probability at least $1-\exp(-g_3'n_1-g_1'\log p)-(2+e)p^{-a_0''}$,

$$\begin{aligned}\|\boldsymbol{H}_1^{(1)}-\boldsymbol{H}\|_\infty \leq&\max\{\sqrt{32B_1^4(a_0''+2)a_4'},8B_1^2(a_0''+2)a_4'\}\sqrt{\log p/n_1}+4\sqrt{2}B_1^3d_1''\sqrt{\frac{s_0\log p}{n_1}}\\ \leq&\tilde{M}\sqrt{\frac{s_0\log p}{n_1}},\end{aligned}\tag{102}$$

where $\tilde{M}=[\max\{\sqrt{32B_1^4(a_0''+2)/a_4'},8B_1^2(a_0''+2)/a_4'\}+4\sqrt{2}B_1^3+1]a_3''d_1''$, and

$$a_3''=\max\{(2M_3+3a_2''/2)/\min\{M_2/3,g_2'/2\},8+2M_3/\{B_1\sqrt{2(a_0''+1)/a_1}\}\}.$$

Similarly, we can show that with probability at least $1-\exp(-g_3'n_1-g_1'\log p)-(2+e)p^{-a_0''}$,

$$\|\boldsymbol{H}_2^{(1)}-\boldsymbol{H}\|_\infty \leq\tilde{M}\sqrt{\frac{s_0\log p}{n_1}}.\tag{103}$$

Both (102) and (103) imply that with probability at least $1-\exp(-g_3'n_1-g_1'\log p)-(4+2e)p^{-a_0''}$,

$$\max\{\|\boldsymbol{H}_1^{(1)}-\boldsymbol{H}\|_\infty,\|\boldsymbol{H}_2^{(1)}-\boldsymbol{H}\|_\infty\}\leq\tilde{M}\sqrt{\frac{s_0\log p}{n_1}}.\tag{104}$$

Assume that with probability at least $1-4(s-2)p^{-a_0''}-\sum_{j=1}^{s-1}\{\exp(-g_3'n_j-g_1'\log p)+2ep^{-a_0''N_j/n_j}\}$,

$$\begin{aligned}&||\hat{\boldsymbol{\Delta}}_1^{(s-1)}||_2\leq d_{s-1}''\sqrt{\frac{s_0\log p}{N_{s-1}}},\quad ||\hat{\boldsymbol{\Delta}}_1^{(s-1)}||_1\leq d_{s-1}''^2s_0\sqrt{\frac{\log p}{N_{s-1}}},\\ &||\hat{\boldsymbol{\Delta}}_2^{(s-1)}||_2\leq d_{s-1}''\sqrt{\frac{s_0\log p}{N_{s-1}}},\quad ||\hat{\boldsymbol{\Delta}}_2^{(s-1)}||_1\leq d_{s-1}''^2s_0\sqrt{\frac{\log p}{N_{s-1}}},\\ &||\hat{\boldsymbol{\beta}}_{ave}^{(s-1)}-\boldsymbol{\beta}^*||_2\leq d_{s-1}''\sqrt{\frac{s_0\log p}{N_2}},\quad\text{and}\quad ||\hat{\boldsymbol{\beta}}_{ave}^{(s-1)}-\boldsymbol{\beta}^*||_1\leq d_{s-1}''^2s_0\sqrt{\frac{\log p}{N_{s-1}}},\end{aligned}\tag{105}$$

and with probability at least $1-4(s-1)p^{-a_0''}-\sum_{j=1}^{s-1}\{\exp(-g_3'n_j-g_1'\log p)+2ep^{-a_0''N_j/n_j}\}$,

$$\begin{aligned}&\max\{\|\frac{1}{N_{s-1}}\sum_{j=1}^{s-1}n_j\boldsymbol{H}_1^{(j)}-\boldsymbol{H}\|_\infty,\|\frac{1}{N_{s-1}}\sum_{j=1}^{s-1}n_j\boldsymbol{H}_2^{(j)}-\boldsymbol{H}\|_\infty\}\\ \leq&\frac{1}{N_{s-1}}\sum_{j=1}^{s-1}n_j\tilde{M}^j\max\{\sqrt{\frac{s_0\log p}{n_j}},\sqrt{s_0}\frac{\log p}{n_j}\},\end{aligned}\tag{106}$$

where $d_{s-1}''=a_3''^{s-2}d_1''$. Let $\lambda_s=c_{1s}''\sqrt{\log p/N_s}$ and $\gamma_s=c_{2s}''\sqrt{\log p/N_s}$, where c_{1s}'' and c_{2s}'' could be any constants which belong to $[2B_1\sqrt{2(a_0''+1)/a_1},a_2'']$. By (96), (97), (105), and

(106), and conditions (C4) and (C5), similar to (45), we can show that with probability at least $1-4(s-1)p^{-a_0''}-\sum_{j=1}^{s}\{\exp(-g_3'n_j-g_1'\log p)+2ep^{-a_0''N_j/n_j}\}$,

$$
\begin{aligned}
&||\hat{\boldsymbol{\Delta}}_1^{(s)}||_2 \le d_s''\sqrt{\frac{s_0\log p}{N_s}}, \quad ||\hat{\boldsymbol{\Delta}}_1^{(s)}||_1 \le d_s''s_0\sqrt{\frac{\log p}{N_s}},\\
&||\hat{\boldsymbol{\Delta}}_2^{(s)}||_2 \le d_s''\sqrt{\frac{s_0\log p}{N_s}}, \quad ||\hat{\boldsymbol{\Delta}}_2^{(s)}||_1 \le d_s''s_0\sqrt{\frac{\log p}{N_s}},\\
&||\hat{\boldsymbol{\beta}}_{ave}^{(s)}-\boldsymbol{\beta}^*||_2 \le d_s''\sqrt{\frac{s_0\log p}{N_s}}, \quad \text{and} \quad ||\hat{\boldsymbol{\beta}}_{ave}^{(s)}-\boldsymbol{\beta}^*||_1 \le d_s''s_0\sqrt{\frac{\log p}{N_s}},
\end{aligned} \tag{107}
$$

where $d_s''=a_3''^{s-1}d_1''$. In light of (107), similar to (104), one can prove that that with probability at least $1-4sp^{-a_0''}-\sum_{j=1}^{s}\{\exp(-g_3'n_j-g_1'\log p)+2ep^{-a_0''N_j/n_j}\}$,

$$
\max\{\|\boldsymbol{H}_1^{(s)}-\boldsymbol{H}\|_\infty, \|\boldsymbol{H}_2^{(s)}-\boldsymbol{H}\|_\infty\} \le \tilde{M}^s\max\{\sqrt{\frac{s_0\log p}{n_s}}, \sqrt{s_0}\frac{\log p}{n_s}\}. \tag{108}
$$

Based on (106), (108) and the triangle inequality, we can show that with probability at least $1-4sp^{-a_0''}-\sum_{j=1}^{s}\{\exp(-g_3'n_j-g_1'\log p)+2ep^{-a_0''N_j/n_j}\}$,

$$
\begin{aligned}
&\max\{\|\frac{1}{N_s}\sum_{j=1}^{s}n_j\boldsymbol{H}_1^{(j)}-\boldsymbol{H}\|_\infty, \|\frac{1}{N_s}\sum_{j=1}^{s}n_j\boldsymbol{H}_2^{(j)}-\boldsymbol{H}\|_\infty\}\\
\le&\max\{\frac{N_{s-1}}{N_s}\|\frac{1}{N_{s-1}}\sum_{j=1}^{s-1}n_j\boldsymbol{H}_1^{(j)}-\boldsymbol{H}\|_\infty+\frac{n_s}{N_s}\|\boldsymbol{H}_1^{(s)}-\boldsymbol{H}\|_\infty, \frac{N_{s-1}}{N_s}\|\frac{1}{N_{s-1}}\sum_{j=1}^{s-1}n_j\boldsymbol{H}_2^{(j)}-\boldsymbol{H}\|_\infty\\
&+\frac{n_s}{N_s}\|\boldsymbol{H}_2^{(s)}-\boldsymbol{H}\|_\infty\},\\
\le&\frac{1}{N_s}\sum_{j=1}^{s}n_j\tilde{M}^j\max\{\sqrt{\frac{s_0\log p}{n_j}}, \sqrt{s_0}\frac{\log p}{n_j}\}.
\end{aligned}
$$

We complete the proof of Corollary 10. ■

The proof of Corollary 12 is similar to that of Corollary 8, and thus is not reported here.

B.4 Proof of Corollaries 7, 9, 11 and 13

Proof Under condition (E7), it is easy to show

$$
\sup_{\|\boldsymbol{\Delta}\|_2=1}\|\boldsymbol{H}^{1/2}\boldsymbol{\Delta}\|_2^2 \le B_5, \quad \text{and} \quad \sup_{\|\boldsymbol{\Delta}\|_2=1}\|\boldsymbol{H}_\tau^{1/2}\boldsymbol{\Delta}\|_2^2 \le B_5. \tag{109}
$$

For any $t\in\mathbb{R}$ and any $\boldsymbol{a}$ which satisfies $\|\boldsymbol{a}\|_2=1$, by using condition (E7), we can obtain

$$
E\{\exp(t\boldsymbol{a}^\top\boldsymbol{X})\}=\exp(t^2\boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a}/2)\le\exp(t^2B_5/2).
$$

It follows from Lemma 5.5 of Vershynin (2012) that there exists a positive number B_6 which depends on B_5 such that

$$
\|\boldsymbol{X}\|_{\psi_2}\le B_6. \tag{110}
$$

In light of (109) and (110), similar to the proofs of Corollaries 6, 8, 10 and 12, respectively, we can obtain the results in Corollaries 7, 9, 11 and 13. ■

References

Pierre Alquier and Gérard Biau. Sparse single-index model. *The Journal of Machine Learning Research*, 14(1):243–280, 2013.

Sladana Babić, Laetitia Gelbgras, Marc Hallin, and Christophe Ley. Optimal tests for elliptical symmetry: specified and unspecified location. *Bernoulli*, 27(4):2189–2216, 2021.

Vladimir Braverman, Gereon Frahling, Harry Lang, Christian Sohler, and Lin F Yang. Clustering high dimensional dynamic data streams. In *International Conference on Machine Learning*, pages 576–585. PMLR, 2017.

Peter Bühlmann and Sara Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, Berlin Heidelberg, 2011.

Leheng Cai, Xu Guo, Gaorong Li, and Falong Tan. Tests for high-dimensional single-index models. *Electronic Journal of Statistics*, 17(1):429–463, 2023.

Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

Stamatis Cambanis, Steel Huang, and Gordon Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981.

Raymond J Carroll, Jianqing Fan, Irene Gijbels, and Matt P Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

Delphine Cassart, Marc Hallin, and Davy Paindaveine. Optimal detection of fechner-asymmetry. *Journal of Statistical Planning and Inference*, 138(8):2499–2525, 2008.

Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

Xia Cui, Wolfgang Karl Härdle, and Lixing Zhu. The efm approach for single-index models. *The Annals of Statistics*, 39(3):1658–1688, 2011.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(6):165–202, 2012.

Yash Deshpande, Adel Javanmard, and Mohammad Mehrabi. Online debiasing for adaptively collected high-dimensional data with applications to time series analysis. *Journal of the American Statistical Association*, 118(542):1126–1139, 2023.

Salah Ud Din, Jay Kumar, Junming Shao, Cobbinah Bernard Mawuli, and Waldiodio David Ndiaye. Learning high-dimensional evolving data streams with limited labels. *IEEE Transactions on Cybernetics*, 52(11):11373–11384, 2021.

John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *The Journal of Machine Learning Research*, 10:2899–2934, 2009.

Hamid Eftekhari, Moulinath Banerjee, and Yaacov Ritov. Inference in high-dimensional single-index models under symmetric designs. *Journal of Machine Learning Research*, 22 (27):1–63, 2021.

Jianqing Fan, Quefeng Li, and Yuyan Wang. 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):247–265, 2017.

Jianqing Fan, Wenyan Gong, Chris Junchi Li, and Qiang Sun. Statistical sparse online regression: A diffusion approximation perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 1017–1026. PMLR, 2018.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of statistics*, 49(3):1239–1266, 2021.

Ravi Ganti, Nikhil Rao, Laura Balzano, Rebecca Willett, and Robert Nowak. On learning high dimensional structured single index models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 2017.

Alexander Gepperth and Benedikt Pfülb. Gradient-based training of gaussian mixture models for high-dimensional streaming data. *Neural Processing Letters*, 53(6):4331–4348, 2021.

Dongxiao Han, Jian Huang, Yuanyuan Lin, and Guohao Shen. Robust post-selection inference of high-dimensional mean regression with heavy-tailed asymmetric or heteroskedastic errors. *Journal of Econometrics*, 230(2):416–431, 2022.

Dongxiao Han, Miao Han, Jian Huang, and Yuanyuan Lin. Robust inference for high-dimensional single index models. *Scandinavian Journal of Statistics*, 50(4):1590–1615, 2023.

Ruijian Han, Lan Luo, Yuanyuan Lin, and Jian Huang. Online debiased lasso for streaming data. *arXiv preprint arXiv:2106.05925*, 2021.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied Logistic Regression*, volume 398. John Wiley & Sons, Hoboken, New Jersey, 2013.

Jian Huang, Tingni Sun, Zhiliang Ying, Yi Yu, and Cun-Hui Zhang. Oracle inequalities for the lasso in the cox model. *The Annals of Statistics*, 41(3):1142–1165, 2013.

Peter J Huber. Robust estimation of a location parameter. *Annals Mathematics Statistics*, 35:73–101, 1964.

Jana Janková and Sara Van De Geer. Confidence regions for high-dimensional generalized linear models under sparsity. *arXiv preprint arXiv:1610.01353*, 2016.

Wei Lan, Ping-Shou Zhong, Runze Li, Hansheng Wang, and Chih-Ling Tsai. Testing a single regression coefficient in high dimensional linear models. *Journal of Econometrics*, 195(1):154–168, 2016.

John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10(3):719–743, 2009.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, Berlin Heidelberg, 1991.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *The Annals of Statistics*, 17(3):1009–1052, 1989.

Nan Lin and Ruibin Xi. Aggregated estimating equation estimation. *Statistics and its Interface*, 4(1):73–83, 2011.

Po-Ling Loh. Scale calibration for high-dimensional robust regression. *Electronic Journal of Statistics*, 15(2):5933–5994, 2021.

Lan Luo and Peter X-K Song. Renewable estimation and incremental inference in generalized linear models with streaming data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):69–97, 2020.

Lan Luo, Ruijian Han, Yuanyuan Lin, and Jian Huang. Online inference in high-dimensional generalized linear models with streaming data. *Electronic Journal of Statistics*, 17(2):3443 – 3471, 2023.

Rong Ma, T Tony Cai, and Hongzhe Li. Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association*, 116(534):984–998, 2021.

Qing Mai, Di He, and Hui Zou. Coordinatewise gaussianization: Theories and applications. *Journal of the American Statistical Association*, 118(544):2329–2343, 2023.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *arXiv preprint arXiv:1010.2731*, 2010.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Matey Neykov, Jun S Liu, and Tianxi Cai. L1-regularized least squares for support recovery of high dimensional single index models with gaussian designs. *The Journal of Machine Learning Research*, 17(1):2976–3012, 2016.

Peter Radchenko. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282, 2015.

Elizabeth D Schifano, Jing Wu, Chun Wang, Jun Yan, and Ming-Hui Chen. Online updating of statistical inference in the big data setting. *Technometrics*, 58(3):393–403, 2016.

Chengchun Shi, Rui Song, Wenbin Lu, and Runze Li. Statistical inference for high-dimensional models via recursive online-score estimation. *Journal of the American Statistical Association*, 116(535):1307–1318, 2021.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116 (29):14516–14525, 2019.

Falong Tan and Lixing Zhu. Adaptive-to-model checking for regressions with diverging number of predictors. *Annals of Statistics*, 47(4):1960–1994, 2019.

Falong Tan and Lixing Zhu. Integrated conditional moment test and beyond: when the number of covariates is divergent. *Biometrika*, 109(1):103–122, 2022.

Leonard J Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), 2014.

Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*. In Y. Eldar and G. Kutyniok (Eds.), Compressed Sensing: Theory and Applications (pp. 210-268). Cambridge University Press, Cambridge, 2012.

Dantong Wang, Simon Fong, Raymond K Wong, Sabah Mohammed, Jinan Fiaidhi, and Kelvin KL Wong. Robust high-dimensional bioinformatics data streams mining by odr-iovfdt. *Scientific Reports*, 7(1):1–12, 2017.

Hansheng Wang, Runze Li, and Chih-Ling Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.

Lili Wang, Chao Zheng, Wen Zhou, and Wen-Xin Zhou. A new principle for tuning-free huber regression. *Statistica Sinica*, 31(4):2153–2177, 2021.

Yingcun Xia, Howell Tong, Wai Keung Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. In *Exploration of A Nonlinear World: An Appreciation of Howell Tong's Contributions to Statistics*, pages 299–346. World Scientific, 2009.

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(88):2543–2596, 2010.

Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. High-dimensional non-gaussian single index models via thresholded score function estimation. In *International Conference on Machine Learning*, pages 3851–3860. PMLR, 2017.

Yuankun Zhang, Heng Lian, and Yan Yu. Ultra-high dimensional single-index quantile regression. *Journal of Machine Learning Research*, 21(224):1–25, 2020.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.