Distributed Estimation and Inference with Statistical Guarantees

Heather Battey*† Jianqing Fan* Han Liu* Junwei Lu* Ziwei Zhu*

September 21, 2015

Abstract

This paper studies hypothesis testing and parameter estimation in the context of the divide and conquer algorithm. In a unified likelihood based framework, we propose new test statistics and point estimators obtained by aggregating various statistics from k subsamples of size n/k, where n is the sample size. In both low dimensional and high dimensional settings, we address the important question of how to choose k as n grows large, providing a theoretical upper bound on k such that the information loss due to the divide and conquer algorithm is negligible. In other words, the resulting estimators have the same inferential efficiencies and estimation rates as a practically infeasible oracle with access to the full sample. Thorough numerical results are provided to back up the theory.

1 Introduction

In recent years, the field of statistics has developed apace in response to the opportunities and challenges spawned from the 'data revolution', which marked the dawn of an era characterized by the availability of enormous datasets. An extensive toolkit of methodology is now in place for addressing a wide range of high dimensional problems, whereby the number of unknown parameters, d, is much larger than the number of observations, n. However, many modern datasets are instead characterized by n and d both large. The latter presents intimidating practical challenges resulting from storage and computational limitations, as well as numerous statistical challenges (Fan et al., 2014). It is important that statistical methodology targeting modern application areas does not lose sight of the practical burdens associated with manipulating such large scale datasets. In this vein, incisive new algorithms have been developed for exploiting modern computing architectures and recent advances in distributed computing. These algorithms enjoy computational efficiency

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540; Email: {hbattey,jqfan,hanliu,junweil,ziweiz} @princeton.edu

†Department of Mathematics, Imperial College London, London, SW7 2AZ; Email: h.battey@imperial.ac.uk

Heather Battey was supported in part by the NIH grant 2R01-GM072611-11. Jianqing Fan was supported in part by NSF Grants DMS-1206464 and DMS-1406266, and NIH grants 2R01-GM072611-11.

and facilitate data handling and storage, but come with a statistical overhead if inappropriately tuned.

With increased mindfulness of the algorithmic difficulties associated with large datasets, the statistical community has witnessed a surge in recent activity in the statistical analysis of various divide and conquer (DC) algorithms, which randomly partition the n observations into k subsamples of size $n_k = n/k$, construct statistics based on each subsample, and aggregate them in a suitable way. In splitting the dataset, a single, very large scale estimation or inference problem with computational complexity $O(\gamma(n))$, for a given function $\gamma(\cdot)$ that depends on the underlying problem, is transformed into k high dimensional (large d smaller n_k) problems each with computational complexity $O\big(\gamma(n/k)\big)$ on each machine. What gets lost in this process is the interactions of split subsamples in each machine. They are not recoverable. However, the information got lost is not much statistically, as the spilt subsamples are supposed to be independent. It is thus of significant practical interest to derive a theoretical upper bound on the number of subsamples k that delivers the same statistical performance as the computationally infeasible "oracle" procedure based on the full sample. We develop communication efficient generalizations of the Wald and Rao score tests for the high dimensional scheme, as well as communication efficient estimators for the parameters of the high dimensional and low dimensional linear and generalized linear models. In all cases we give the upper bound on k for preserving the statistical error of the analogous full sample procedure.

While hypothesis testing in a low dimensional context is straightforward, in the high dimensional setting, nuisance parameters introduce a non-negligable bias, causing classical low dimensional theory to break down. In our high dimensional Wald construction, the phenomenon is remedied through a debiasing of the estimator, which gives rise to a test statistic with tractable limiting distribution, as documented in the $k = 1$ (no sample split) setting in Zhang and Zhang (2014) and van de Geer et al. (2014). For the high dimensional analogue of Rao's score statistic, the incorporation of a correction factor increases the convergence rate of higher order terms, thereby vanquishing the effect of the nuisance parameters. The approach is introduced in the $k = 1$ setting in Ning and Liu (2014), where the test statistic is shown to possess a tractable limit distribution. However, the computation complexity for the debiased estimators increases by an order of magnitude, due to solving d high-dimensional regularization problems. This motivates us to appeal to the divide and conquer strategy.

We develop the theory and methodology for DC versions of these tests. In the case $k = 1$, each of the above test statistics can be decomposed into a dominant term with tractable limit distribution and a negligible remainder term. The DC extension requires delicate control of these remainder terms to ensure the error accumulation remains sufficiently small so as not to materially contaminate the leading term. In obtaining the upper bound on the number of permitted subsamples, k, we provide an upper bound on k subject to a statistical guarantee. We find that the theoretical upper bound on the number of subsamples guaranteeing the same inferential or estimation efficiency as the procedure without DC is $k = o((s \log d)^{-1}\sqrt{n})$ in the linear model, where s is the sparsity of the parameter vector. In the generalized linear model the scaling is $k = o(((s \vee s_1) \log d)^{-1}\sqrt{n})$, where s_1 is the sparsity of the inverse information matrix.

For high dimensional estimation problems, we use the same debiasing trick introduced in the high dimensional testing problems to obtain a thresholded divide and conquer estimator that achieves the full sample minimax rate. The appropriate scaling is found to be $k = O(\sqrt{n/(s^2 \log d)})$ for the linear model and $k = O(\sqrt{n/((s \vee s_1)^2 \log d)})$ for the generalized linear model. Moreover, we find that the loss incurred by the divide and conquer strategy, as quantified by the distance between the DC estimator and the full sample estimator, is negligible in comparison to the statistical error of the full sample estimator provided that k is not too large. In the context of estimation, the optimal scaling of k with n and d is also developed for the low dimensional linear and generalized linear model. This theory is of independent interest. It also allows us to study a refitted estimation procedure under a minimal signal strength assumption.

A partial list of references covering DC algorithms from a statistical perspective is Chen and Xie (2012), Zhang et al. (2013), Kleiner et al. (2014), and Zhao et al. (2014a). For the high dimensional estimation setting, the same debiasing approach of van de Geer et al. (2014) is proposed independently by Lee et al. (2015) for divide-and-conquer estimation. Our paper differs from that of Lee et al. (2015) in that we additionally cover high dimensional hypothesis testing and refitted estimation in the DC setting. Our results on hypothesis testing reveals a different phenomenon to that found in estimation, as we observe through the different requirements on the scaling of k. On the estimation side, our results also differ from those of Lee et al. (2015) in that the additional refitting step allows us to achieve the oracle rate under the same scaling of k.

The rest of the paper is organized as follows. Section 2 collects notation and details of a generic likelihood based framework. Section 3 covers testing, providing high dimensional DC analogues of the Wald test (Section 3.1) and Rao score test (Section 3.2), in each case deriving a tractable limit distribution for the corresponding test statistic under standard assumptions. Section 4 covers distributed estimation, proposing an aggregated estimator of $\boldsymbol{\beta}^*$ for low dimensional and high dimensional linear and generalized linear models, as well as a refitting procedure that improves the estimation rate, with the same scaling, under a minimal signal strength assumption. Section 5 provides numerical experiments to back up the developed theory. In Section 6 we discuss our results together with remaining future challenges. Proofs of our main results are collected in Section 7, while the statement and proofs of a number of technical lemmas are deferred to the appendix.

2 Background and Notation

We first collect the general notation, before providing a formal statement of our statistical problems. More specialized notation is introduced in context.

2.1 Generic Notation

We adopt the common convention of using boldface letters for vectors only, while regular font is used for both matrices and scalars, with the context ensuring no ambiguity. $|\cdot|$ denotes both absolute value and cardinality of a set, with the context ensuring no ambiguity. For $\boldsymbol{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, and $1 \leq q \leq \infty$, we define $\|\boldsymbol{v}\|_q = \left(\sum_{j=1}^d |v_j|^q\right)^{1/q}$, $\|\boldsymbol{v}\|_0 = |\operatorname{supp}(\boldsymbol{v})|$, where $\operatorname{supp}(\boldsymbol{v}) = \{j : v_j \neq 0\}$ and $|A|$ is the cardinality of the set A. Write $\|\boldsymbol{v}\|_\infty = \max_{1 \leq j \leq d} |v_j|$, while for a matrix $M = [M_{jk}]$,

let $\|M\|_{\max} = \max_{j,k} |M_{jk}|$, $\|M\|_1 = \sum_{j,k} |M_{jk}|$. For any matrix M we use $\boldsymbol{M}_\ell$ to index the transposed ℓ^{th} row of M and $[\boldsymbol{M}]_\ell$ to index the ℓ^{th} column. The sub-Gaussian norm of a scalar random variable X is defined as $\|X\|_{\psi_2} = \sup_{q \geq 1} q^{-1/2} (\mathbb{E}|X|^q)^{1/q}$. For a random vector $\boldsymbol{X} \in \mathbb{R}^d$, its sub-Gaussian norm is defined as $\|\boldsymbol{X}\|_{\psi_2} = \sup_{x \in \mathbb{S}^{d-1}} \|\langle \boldsymbol{X}, \boldsymbol{x} \rangle\|_{\psi_2}$, where $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$. Let I_d denote the $d \times d$ identity matrix; when the dimension is clear from the context, we omit the subscript. We also denote the Hadamard product of two matrices A, B as $A \circ B$ and $(A \circ B)_{jk} = A_{jk} B_{jk}$ for any j, k. $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$ denotes the canonical basis for $\mathbb{R}^d$. For a vector $\boldsymbol{v} \in \mathbb{R}^d$ and a set of indices $\mathcal{S} \subseteq \{1, \ldots, d\}$, $\boldsymbol{v}_{\mathcal{S}}$ is the vector of length $|\mathcal{S}|$ whose components are $\{v_j : j \in \mathcal{S}\}$. Additionally, for a vector $\boldsymbol{v}$ with j^{th} element v_j, we use the notation $\boldsymbol{v}_{-j}$ to denote the remaining vector when the j^{th} element is removed. With slight abuse of notation, we write $\boldsymbol{v} = (v_j, \boldsymbol{v}_{-j})$ when we wish to emphasize the dependence of $\boldsymbol{v}$ on v_j and $\boldsymbol{v}_{-j}$ individually. The gradient of a function $f(\boldsymbol{x})$ is denoted by $\nabla f(\boldsymbol{x})$, while $\nabla_{\boldsymbol{x}} f\big((\boldsymbol{x}, \boldsymbol{y})\big)$ denotes the gradient of $f\big((\boldsymbol{x}, \boldsymbol{y})\big)$ with respect to $\boldsymbol{x}$, and $\nabla^2_{\boldsymbol{xy}} f\big((\boldsymbol{x}, \boldsymbol{y})\big)$ denotes the matrix of cross partial derivatives with respect to the elements of $\boldsymbol{x}$ and $\boldsymbol{y}$. For a scalar η, we simply write $f'(\eta) := \nabla_\eta f(\eta)$ and $f''(\eta) := \nabla^2_{\eta\eta} f(\eta)$. For a random variable X and a sequence of random variables, X_n, we write $X_n \rightsquigarrow X$ when X_n converges weakly to X. If X is a random variable with standard distribution, say F_X, we simply write $X_n \rightsquigarrow F_X$. Given $a, b \in \mathbb{R}$, let $a \vee b$ and $a \wedge b$ denote the maximum and minimum of a and b. We also make use of the notation $a_n \lesssim b_n$ ($a_n \gtrsim b_n$) if a_n is less than (greater than) b_n up to a constant, and $a_n \asymp b_n$ if a_n is the same order as b_n. Finally, for an arbitrary function f, we use $\text{argzero}_\theta f(\theta)$ to denote the solution to $f(\theta) = 0$.

2.2 General Likelihood based Framework

Let $(\boldsymbol{X}_1^T, Y_1)^T, \ldots, (\boldsymbol{X}_n^T, Y_n)^T$ be n i.i.d. copies of the random vector $(\boldsymbol{X}^T, Y)^T$, whose realizations take values in $\mathbb{R}^d \times \mathcal{Y}$. Write the collection of these n i.i.d. random couples as $\mathcal{D} = \{(\boldsymbol{X}_1^T, Y_1)^T, \ldots, (\boldsymbol{X}_n^T, Y_n)^T\}$ with $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $X = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T \in \mathbb{R}^{n \times d}$. Conditional on $\boldsymbol{X}_i$, we assume Y_i is distributed as $F_{\boldsymbol{\beta}^*}$ for all $i \in \{1, \ldots, n\}$, where $F_{\boldsymbol{\beta}^*}$ has a density or mass function $f_{\boldsymbol{\beta}^*}$. We thus define the negative log-likelihood function, $\ell_n(\boldsymbol{\beta})$, as

$$\ell_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \ell_i(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^n \log f_{\boldsymbol{\beta}}(Y_i | \boldsymbol{X}_i). \tag{2.1}$$

We use $J^* = J(\boldsymbol{\beta}^*)$ to denote the information matrix and Θ^* to denote $(J^*)^{-1}$, where $J(\boldsymbol{\beta}) = \mathbb{E}\big[\nabla^2_{\boldsymbol{\beta\beta}} \ell_n(\boldsymbol{\beta})\big]$.

For testing problems, our goal is to test $H_0 : \beta_v^* = \beta_v^H$ for any $v \in \{1, \ldots, d\}$. We partition $\boldsymbol{\beta}^*$ as $\boldsymbol{\beta}^* = (\beta_v^*, \boldsymbol{\beta}_{-v}^{*T})^T \in \mathbb{R}^d$, where $\boldsymbol{\beta}_{-v}^* \in \mathbb{R}^{d-1}$ is a vector of nuisance parameters and β_v^* is the parameter of interest. To handle the curse of dimensionality, we exploit a penalized M-estimator defined as,

$$\widehat{\boldsymbol{\beta}}^\lambda = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \ell_n(\boldsymbol{\beta}) + \mathcal{P}_\lambda(\boldsymbol{\beta}) \right\}, \tag{2.2}$$

with $\mathcal{P}_\lambda(\boldsymbol{\beta})$ a sparsity inducing penalty function with a regularization parameter λ. Examples of $\mathcal{P}_\lambda(\boldsymbol{\beta})$ include the convex ℓ_1 penalty, $\mathcal{P}_\lambda(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1 = \lambda \sum_{v=1}^d |\beta_v|$ which, in the context of the

linear model, gives rise to the LASSO estimator (Tibshirani, 1996),

$$\widehat{\boldsymbol{\beta}}^{\lambda}_{\text{LASSO}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}. \tag{2.3}$$

Other penalties include folded concave penalties such as the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) and minimax concave MCP penalty (Zhang, 2010), which eliminate the estimation bias and attain the oracle rates of convergence (Loh and Wainwright, 2013; Wang et al., 2014a). The SCAD penalty is defined as

$$\mathcal{P}_\lambda(\boldsymbol{\beta}) = \sum_{v=1}^{d} p_\lambda(\beta_v), \text{ where } p_\lambda(t) = \int_0^{|t|} \left\{ \lambda \mathbf{1}(z \le \lambda) + \frac{a\lambda - z}{a-1} \mathbf{1}(z > \lambda) \right\} dz, \tag{2.4}$$

for a given parameter $a > 0$ and MCP penalty is given by

$$\mathcal{P}_\lambda(\boldsymbol{\beta}) = \sum_{v=1}^{d} p_\lambda(\beta_v), \text{ where } p_\lambda(t) = \lambda \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz \tag{2.5}$$

where $b > 0$ is a fixed parameter. The only requirement we have on $\mathcal{P}_\lambda(\boldsymbol{\beta})$ is that it induces an estimator satisfying the following condition.

Condition 2.1 . For any $\delta \in (0,1)$, if $\lambda \asymp \sqrt{\log(d/\delta)/n}$,

$$\mathbb{P}\Big(\|\widehat{\boldsymbol{\beta}}^{\lambda} - \boldsymbol{\beta}^*\|_1 > Csn^{-1/2}\sqrt{\log(d/\delta)} \Big) \le \delta, \tag{2.6}$$

where s is the sparsity of $\boldsymbol{\beta}^*$, i.e., $s = \|\boldsymbol{\beta}^*\|_0$.

Condition 2.1 holds for the LASSO, SCAD and MCP. See Bühlmann and van de Geer (2011); Fan and Li (2001); Zhang (2010) respectively and Zhang and Zhang (2012).

The DC algorithm randomly and evenly partitions $\mathcal{D}$ into k disjoint subsets $\mathcal{D}_1, \ldots, \mathcal{D}_k$, so that $\mathcal{D} = \cup_{j=1}^k \mathcal{D}_j$, $\mathcal{D}_j \cap \mathcal{D}_\ell = \emptyset$ for all $j, \ell \in \{1, \ldots, k\}$, and $|\mathcal{D}_1| = |\mathcal{D}_2| = \cdots = |\mathcal{D}_k| = n_k = n/k$, where it is implicitly assumed that n can be divided evenly. Let $\mathcal{I}_j \subset \{1, \ldots, n\}$ be the index set corresponding to the elements of $\mathcal{D}_j$. Then for an arbitrary $n \times d$ matrix A, $A^{(j)} = [A_{i\ell}]_{i \in \mathcal{I}_j, 1 \le \ell \le d}$. For an arbitrary estimator $\widehat{\tau}$, we write $\widehat{\tau}(\mathcal{D}_j)$ when the estimator is constructed based only on $\mathcal{D}_j$. What information gets lost in this process is the interactions of data across subsamples $\{\mathcal{D}_j\}_{j=1}^{n/k}$. Taking the oridinary least-squares regression, for example, the cross-covariances of the subsamples will not be able to get recovered. However, they do not contain much information about the unknown parameters, as the subsamples are nearly independent. Finally, we write $\ell_{n_k}^{(j)}(\boldsymbol{\beta}) = \sum_{i \in \mathcal{I}_j} \ell_i(\boldsymbol{\beta})$ to denote the negative log-likelihood function of equation (2.1) based on $\mathcal{D}_j$.

While the results of this paper hold in a general likelihood based framework, for simplicity we state conditions at the population level for the generalized linear model (GLM) with canonical link. A much more general set of statements appear in the auxiliary lemmas upon which our main results are based. Under GLM with the canonical link, the response follows the distribution,

$$f_n(\boldsymbol{Y}; X, \boldsymbol{\beta}^*) = \prod_{i=1}^{n} f(Y_i; \eta_i^*) = \prod_{i=1}^{n} \left\{ c(Y_i) \exp\left[\frac{Y_i \eta_i^* - b(\eta_i^*)}{\phi} \right] \right\}, \tag{2.7}$$

where $\eta_i^* = \boldsymbol{X}_i^T\boldsymbol{\beta}^*$. The negative log-likelihood corresponding to (2.7) is given, up to an affine transformation, by

$$\ell_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} -Y_i\boldsymbol{X}_i^T\boldsymbol{\beta} + b(\boldsymbol{X}_i^T\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} -Y_i\eta_i + b(\eta_i) = \frac{1}{n}\sum_{i=1}^{n}\ell_i(\boldsymbol{\beta}), \tag{2.8}$$

and the gradient and Hessian of $\ell_n(\boldsymbol{\beta})$ are respectively

$$\nabla\ell_n(\boldsymbol{\beta}) = -\frac{1}{n}X^T(\boldsymbol{Y} - \boldsymbol{\mu}(X\boldsymbol{\beta})) \quad \text{and} \quad \nabla^2\ell_n(\boldsymbol{\beta}) = \frac{1}{n}X^T D(X\boldsymbol{\beta})X,$$

where $\boldsymbol{\mu}(\boldsymbol{\beta}) = (b'(\eta_1), \ldots, b'(\eta_n))^T$ and $D(\boldsymbol{\beta}) = \text{diag}\{b''(\eta_1), \ldots, b''(\eta_n)\}$. In this setting, $J(\boldsymbol{\beta}) = \mathbb{E}[b''(\boldsymbol{X}_1^T\boldsymbol{\beta})\boldsymbol{X}_1\boldsymbol{X}_1^T]$ and $J^* = \mathbb{E}[b''(\boldsymbol{X}_1^T\boldsymbol{\beta}^*)\boldsymbol{X}_1\boldsymbol{X}_1^T]$.

3 Divide and Conquer Hypothesis Tests

In the context of the two classical testing frameworks, the Wald and Rao score tests, our objective is to construct a test statistic $\overline{S}_n$ with low communication cost and a tractable limiting distribution F. Let β_v^* be the v^{th} component of $\boldsymbol{\beta}^*$. From this statistic we define a test of size α of the null hypothesis, $H_0 : \beta_v^* = \beta_v^H$, against the alternative, $H_1 : \beta_v^* \neq \beta_v^H$, as a partition of the sample space described by

$$T_n^\alpha = \begin{cases} 0 & \text{if } |\overline{S}_n| \leq F^{-1}(1-\alpha/2) \\ 1 & \text{if } |\overline{S}_n| > F^{-1}(1-\alpha/2) \end{cases} \tag{3.1}$$

for a two sided test.

3.1 Two Divide and Conquer Wald Type Constructions

For the high dimensional linear model, Zhang and Zhang (2014), van de Geer et al. (2014) and Javanmard and Montanari (2014) propose methods for debiasing the LASSO estimator with a view to constructing high dimensional analogues of Wald statistics and confidence intervals for low-dimensional coordinates. As pointed out by Zhang and Zhang (2014), the debiased estimator does not impose the minimum signal condition used in establishing oracle properties of regularized estimators (Fan and Li, 2001; Fan and Lv, 2011; Loh and Wainwright, 2015; Wang et al., 2014b; Zhang and Zhang, 2012) and hence has wider applicability than those inferences based on the oracle properties. The method of van de Geer et al. (2014) is appealing in that it accommodates a general penalized likelihood based framework, while the Javanmard and Montanari (2014) approach is appealing in that it optimizes asymptotic variance and requires a weaker condition than van de Geer et al. (2014) in the specific case of the linear model. We consider the DC analogues of Javanmard and Montanari (2014) and van de Geer et al. (2014) in Sections 3.1.1 and 3.1.2 respectively.

3.1.1 LASSO based Wald Test for the Linear Model

The linear model assumes

$$Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta}^* + \varepsilon_i, \tag{3.2}$$

where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. with mean zero and variance σ^2. For concreteness, we focus on a LASSO based method, but our procedure is also valid when other pilot estimators are used. We describe a modification of the bias correction method introduced in Javanmard and Montanari (2014) as a means to testing hypotheses on low dimensional coordinates of $\boldsymbol{\beta}^*$ via pivotal test statistics.

On each subset $\mathcal{D}_j$, we compute the debiased estimator of $\boldsymbol{\beta}^*$ as in Javanmard and Montanari (2014) as

$$\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j) = \widehat{\boldsymbol{\beta}}^{\lambda}_{\text{LASSO}}(\mathcal{D}_j) + \frac{1}{n_k} M^{(j)} \big(X^{(j)}\big)^T \big(Y^{(j)} - X^{(j)} \widehat{\boldsymbol{\beta}}^{\lambda}_{\text{LASSO}}(\mathcal{D}_j)\big), \tag{3.3}$$

where the superscript d is used to indicate the debiased version of the estimator, $M^{(j)} = (\mathbf{m}_1^{(j)}, \ldots, \mathbf{m}_d^{(j)})^T$ and $\mathbf{m}_v$ is the solution of

$$\begin{aligned} \mathbf{m}_v^{(j)} = \underset{\mathbf{m}}{\operatorname{argmin}}\, \mathbf{m}^T \widehat{\Sigma}^{(j)} \mathbf{m} \quad \text{s.t.} \quad & \|\widehat{\Sigma}^{(j)} \mathbf{m} - \boldsymbol{e}_v\|_\infty \le \vartheta_1, \\ & \|X^{(j)} \mathbf{m}\|_\infty \le \vartheta_2, \end{aligned} \tag{3.4}$$

where the choice of tuning parameters ϑ_1 and ϑ_2 is discussed in Javanmard and Montanari (2014) and Zhao et al. (2014a). Above, $\widehat{\Sigma}^{(j)} = {n_k}^{-1} \sum_{i \in \mathcal{I}_j} \boldsymbol{X}_i^{(j)} \boldsymbol{X}_i^{(j)T}$ is the sample covariance based on $\mathcal{D}_j$, whose population counterpart is $\Sigma = \mathbb{E}(\boldsymbol{X}_1 \boldsymbol{X}_1^T)$ and $M^{(j)}$ is its regurlized inverse. The second term in (3.3) is a bias correction term, while $\sigma^2 \mathbf{m}_v^{(j)T} \widehat{\Sigma}^{(j)} \mathbf{m}_v^{(j)} / n_k$ is shown in Javanmard and Montanari (2014) to be the variance of the v^{th} component of $\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$. The parameter ϑ_1, which tends to zero, controls the bias of the debiased estimator (3.3) and the optimization in (3.4) minimizes the variance of the resulting estimator.

Solving d optimization problems in (3.4) increase an order of magnitude of computation complexity even for $k = 1$. Thus, it is necessary to appeal to the divide and conquer strategy to reduce the computation burden. This gives rise to the question how large k can be in order to maintain the same statistical properties as the whole sample one ($k = 1$).

Because our DC procedure gives rise to smaller samples, we overcome the singularity in $\widehat{\Sigma}$ through a change of variables. More specifically, noting that $M^{(j)}$ is not required explicitly, but rather the product $M^{(j)}(X^{(j)})^T$, we propose

$$\begin{aligned} \boldsymbol{b}_v^{(j)} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \frac{\boldsymbol{b}^{(j)T} \boldsymbol{b}^{(j)}}{n_k} \quad \text{s.t.} \quad & \Big\| \frac{X^{(j)T} \boldsymbol{b}^{(j)}}{n_k} - \boldsymbol{e}_v \Big\|_\infty \le \vartheta_1, \\ & \|\boldsymbol{b}^{(j)}\|_\infty \le \vartheta_2, \end{aligned}$$

from which we construct $M^{(j)}(X^{(j)})^T = B^T$, where $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d)$.

The following conditions on the data generating process and the tail behavior of the design vectors are imposed in Javanmard and Montanari (2014). Both conditions are used to derive the theoretical properties of the DC Wald test statistic based on the aggregated debiased estimator, $\overline{\boldsymbol{\beta}}^d = k^{-1} \sum_{j=1}^k \widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$.

Condition 3.1 . $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ are i.i.d. and Σ satisfies $0 < C_{\min} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le C_{\max}$.

Condition 3.2 . The rows of X are sub-Gaussian with $\|\boldsymbol{X}_i\|_{\psi_2} \le \kappa$, $i = 1, \ldots, n$.

Note that under the two conditions above, there exists a constant $\kappa_1 > 0$ such that $\|\boldsymbol{X}_1\Sigma^{-\frac{1}{2}}\|_{\psi_2} \leq \kappa_1$. Without loss of generality, we can set $\kappa_1 = \kappa$. Our first main theorem provides the relative scaling of the various tuning parameters involved in the construction of $\overline{\boldsymbol{\beta}}^d$.

Theorem 3.3. Suppose Conditions 2.1, 3.1 and 3.2 are fulfilled. Suppose $\mathbb{E}[\varepsilon_1^4] < \infty$ and choose ϑ_1, ϑ_2 and k such that $\vartheta_1 \asymp \sqrt{k \log d/n}$, $\vartheta_2 n^{-1/2} = o(1)$ and $k = o((s \log d)^{-1}\sqrt{n})$. For any $v \in \{1, \ldots, d\}$,

$$\sqrt{n}\frac{1}{k}\sum_{j=1}^{k} \frac{\widehat{\beta}_v^d(\mathcal{D}_j) - \beta_v^*}{\widehat{Q}_v^{(j)}} \rightsquigarrow N(0, \sigma^2), \tag{3.5}$$

where $\widehat{Q}_v^{(j)} = \big(\mathbf{m}_v^{(j)T}\widehat{\Sigma}^{(j)}\mathbf{m}_v^{(j)}\big)^{1/2}$.

Theorem 3.3 entertains the prospect of a divide and conquer Wald statistic of the form

$$\overline{S}_n = \sqrt{n}\frac{1}{k}\sum_{j=1}^{k} \frac{\widehat{\beta}_v^d(\mathcal{D}_j) - \beta_v^H}{\overline{\sigma}\big(\mathbf{m}_v^{(j)T}\widehat{\Sigma}^{(j)}\mathbf{m}_v^{(j)}\big)^{1/2}} \tag{3.6}$$

for β_v^*, where $\overline{\sigma}$ is an estimator for σ based on the k subsamples. On the left hand side of equation (3.6) we suppress the dependence on v to simplify notation. As an estimator for σ, a simple suggestion with the same computational complexity is $\overline{\sigma}$ where

$$\overline{\sigma}^2 = \frac{1}{k}\sum_{j=1}^{k}\widehat{\sigma}^2(\mathcal{D}_j) \quad \text{and} \quad \widehat{\sigma}^2(\mathcal{D}_j) = \frac{1}{n_k}\sum_{i\in\mathcal{I}_j}(Y_i^{(j)} - \boldsymbol{X}_i^{(j)T}\widehat{\boldsymbol{\beta}}^{\lambda}_{\text{LASSO}}(\mathcal{D}_j))^2. \tag{3.7}$$

One can use the refitted cross-validation procedure of Fan et al. (2012) to reduce the bias of the estimate. In Lemma 3.4 we show that with the scaling of k and λ required for the weak convergence results of Theorem 3.3, consistency of $\overline{\sigma}^2$ is also achieved.

Lemma 3.4. Suppose $\mathbb{E}[\varepsilon_i|\boldsymbol{X}_i] = 0$ for all $i \in \{1, \ldots, n\}$. Then with $\lambda \asymp \sqrt{k \log d/n}$ and $k = o\big(\sqrt{n}(s \log d)^{-1}\big)$, $|\overline{\sigma}^2 - \sigma^2| = o_{\mathbb{P}}(1)$.

With Lemma 3.4 and Theorem 3.3 at hand, we establish in Corollary 3.5 the asymptotic distribution of $\overline{S}_n$ under the null hypothesis $H_0 : \beta_v^* = \beta_v^H$. This holds for each component $v \in \{1, \ldots, d\}$.

Corollary 3.5. Suppose Conditions 3.1 and 3.2 are fulfilled, $\mathbb{E}[\varepsilon_1^4] < \infty$, and λ, ϑ_1 and ϑ_2 are chosen as $\lambda \asymp \sqrt{k \log d/n}$, $\vartheta_1 \asymp \sqrt{k \log d/n}$ and $\vartheta_2 n^{-1/2} = o(1)$. Then provided $k = o((s \log d)^{-1}\sqrt{n})$, under $H_0 : \beta_v^* = \beta_v^H$, we have

$$\lim_{n\to\infty}\sup_{t\in\mathbb{R}}\big|\mathbb{P}\big(\overline{S}_n \leq t\big) - \Phi(t)\big| = 0, \tag{3.8}$$

where $\Phi(\cdot)$ is the cdf of a standard normal distribution.

3.1.2 Wald Test in the Likelihood Based Framework

An alternative route to debiasing the LASSO estimator of $\boldsymbol{\beta}^*$ is the one proposed in van de Geer et al. (2014). Their so called desparsified estimator of $\boldsymbol{\beta}^*$ is more general than the debiased estimator

of Javanmard and Montanari (2014) in that it accommodates generic estimators of the form (2.2) as pilot estimators, but the latter optimizes the variance of the resulting estimator. The desparsified estimator for subsample $\mathcal{D}_j$ is

$$\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j) = \widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \widehat{\Theta}^{(j)} \nabla \ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)), \tag{3.9}$$

where $\widehat{\Theta}^{(j)}$ is a regularized inverse of the Hessian matrix of second order derivatives of $\ell_{n_k}^{(j)}(\boldsymbol{\beta})$ at $\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)$, denoted by $\widehat{J}^{(j)} = \nabla^2 \ell_{n_k}^{(j)}\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)\big)$. We will make this explicit in due course. The estimator resembles the classical one-step estimator (Bickel, 1975), but now in the high-dimensional setting via regularized inverse of the Hessian matrix $\widehat{J}^{(j)}$, which reduces to the empirical covariance of the design matrix in the case of the linear model. From equation (3.9), the aggregated debiased estimator over the k subsamples is defined as $\overline{\boldsymbol{\beta}}^d = k^{-1}\sum_{j=1}^k \widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$.

We now use the nodewise LASSO (Meinshausen and Bühlmann, 2006) to approximately invert $\widehat{J}^{(j)}$ via L_1-regularization. The basic idea is to find the regularized invert row by row via a penalized L_1-regression, which is the same as regressing the variable X_v on $\boldsymbol{X}_{-v}$ but expressed in the sample covariance form. For each row $v \in 1, \ldots, d$, consider the optimization

$$\widehat{\boldsymbol{\kappa}}_v(\mathcal{D}_j) = \underset{\boldsymbol{\kappa}\in\mathbb{R}^{d-1}}{\operatorname{argmin}} \ \big(\widehat{J}_{vv}^{(j)} - 2\widehat{J}_{v,-v}^{(j)}\boldsymbol{\kappa} + \boldsymbol{\kappa}^T \widehat{J}_{-v,-v}^{(j)}\boldsymbol{\kappa} + 2\lambda_v\|\boldsymbol{\kappa}\|_1\big), \tag{3.10}$$

where $\widehat{J}_{v,-v}^{(j)}$ denotes the v^{th} row of $\widehat{J}^{(j)}$ without the $(v,v)^{\text{th}}$ diagonal element, and $\widehat{J}_{-v,-v}^{(j)}$ is the principal submatrix without the v^{th} row and v^{th} column. Introduce

$$\widehat{C}^{(j)} := \begin{pmatrix} 1 & -\widehat{\kappa}_{1,2}(\mathcal{D}_j) & \ldots & -\widehat{\kappa}_{1,d}(\mathcal{D}_j) \\ -\widehat{\kappa}_{2,1}(\mathcal{D}_j) & 1 & \ldots & -\widehat{\kappa}_{2,d}(\mathcal{D}_j) \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\kappa}_{d,1}(\mathcal{D}_j) & -\widehat{\kappa}_{d,2}(\mathcal{D}_j) & \ldots & 1 \end{pmatrix} \tag{3.11}$$

and $\widehat{\Xi}^{(j)} = \operatorname{diag}\big(\widehat{\tau}_1(\mathcal{D}_j), \ldots, \widehat{\tau}_d(\mathcal{D}_j)\big)$, where $\widehat{\tau}_v(\mathcal{D}_j)^2 = \widehat{J}_{vv}^{(j)} - \widehat{J}_{v,-v}^{(j)}\widehat{\boldsymbol{\kappa}}_v(\mathcal{D}_j)$. $\widehat{\Theta}^{(j)}$ in equation (3.9) is given by

$$\widehat{\Theta}^{(j)} = (\widehat{\Xi}^{(j)})^{-2}\widehat{C}^{(j)}, \tag{3.12}$$

and we define $\widehat{\boldsymbol{\Theta}}_v^{(j)}$ as the transposed v^{th} row of $\widehat{\Theta}^{(j)}$.

Theorem 3.8 establishes the limit distribution of the term,

$$\overline{S}_n = \sqrt{n}\frac{1}{k}\sum_{j=1}^k \frac{\widehat{\beta}_v^d(\mathcal{D}_j) - \beta_v^H}{\sqrt{\Theta_{vv}^*}} \tag{3.13}$$

for any $v \in \{1, \ldots, d\}$ under the null hypothesis $H_0 : \beta_v^* = \beta_v^H$. This provides the basis for the statistical inference based on divide-and-conquer. We need the following condition. Recall that $J^* = \mathbb{E}\big[\nabla_{\boldsymbol{\beta}\boldsymbol{\beta}}\ell_n(\boldsymbol{\beta}^*)\big]$ and consider the generalized linear model (2.7).

Condition 3.6 . (i) $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ are i.i.d., $0 < C_{\min} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le C_{\max}$, $\lambda_{\min}(J^*) \ge L_{\min} > 0$, $\|J^*\|_{\max} < U_1 < \infty$. (ii) For some constant $M < \infty$, $\max_{1\le i\le n}\big|\boldsymbol{X}_i^T\boldsymbol{\beta}^*\big| \le M$ and $\max_{1\le i\le n}\|\boldsymbol{X}_i\|_\infty \le M$. (iii) There exist finite constants $U_2, U_3 > 0$ such that $b''(\eta) < U_2$ and $b'''(\eta) < U_3$ for all $\eta \in \mathbb{R}$.

The same assumptions appear in van de Geer et al. (2014). In the case of the Gaussian GLM, the condition on $\lambda_{\min}(J^*)$ reduces to the requirement that the covariance of the design has minimal eigenvalue bounded away from zero, which is a standard assumption. We require $\|J^*\|_{\max} < \infty$ to control the estimation error of different functionals of J^*. The restriction in (ii) on the covariates and the projection of the covariates are imposed for technical simplicity; it can be extended to the case of exponential tails (see Fan and Song, 2010). Note that $\mathrm{Var}(Y_i) = \phi b''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)$ where ϕ is the dispersion parameter in (2.7), so $b''(\eta) < U_2$ essentially implies an upper bound on the variance of the response. In fact, Lemma A.2 shows that $b''(\eta) < U_2$ can guarantee that the response is sub-gaussian. $b'''(\eta) < U_3$ is used to derive the Lipschitz property of $b''(\boldsymbol{X}_i^T\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ as shown in Lemma A.5. We emphasize that no requirement in Condition 3.6 is specific to the divide and conquer framework.

The assumption of bounded design in (ii) can be relaxed to the sub-gaussian design. However, the price to pay is that the allowable number of subsets k is smaller than the bounded case, which means we need a larger sub-sample size. To be more precise, the order of maximum k for the sub-gaussian design has an extra factor, which is a polynomial of $\sqrt{\log d}$, compared to the order for the bounded design. This logrithmic factor comes from different Lipschitz properties of $b''(\boldsymbol{X}_i^T\boldsymbol{\beta})$ in the two designs, which is fully explained in Lemma A.5 of the appendix. In the following theorems, we only present results for the case of bounded design for technical simplicity.

In addition, recalling that $\Theta^* = (J^*)^{-1}$, where $J^* := J(\boldsymbol{\beta}^*) = \mathbb{E}\big[\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}}\,\ell_n(\boldsymbol{\beta}^*)\big]$, we impose Condition 3.7 on Θ^* and its estimator $\widehat{\Theta}$.

Condition 3.7 . (i) $\min_{1\le v\le d}\Theta^*_{vv} > \theta_{\min} > 0$. (ii) $\max_{1\le i\le n}\|\boldsymbol{X}_i^T\boldsymbol{\Theta}^*\|_\infty \le M$. (iii) For $v = 1,\ldots,d$, whenever $\lambda_v \asymp \sqrt{k\log d/n}$ in (3.10), we have

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\Theta}}_v - \boldsymbol{\Theta}^*_v\|_1 \ge C s_1\sqrt{\log d/n}\right) \le d^{-1},$$

where C is a constant and s_1 is such that $\|\boldsymbol{\Theta}^*_v\|_0 \lesssim s_1$ for all $v \in \{1,\ldots,d\}$.

Part (i) of Corollary 3.7 ensures that the variances of each component of the debiased estimator exist, guaranteeing the existence of the Wald statistic. Parts (ii) and (iii) are imposed directly for technical simplicity. Results of this nature have been established under a similar set of assumptions in van de Geer et al. (2014) and Negahban et al. (2009) for convex penalties and in Wang et al. (2014a) and Loh and Wainwright (2015) for folded concave penalties.

As a step towards deriving the limit distribution of the proposed divide and conquer Wald statistic in the GLM framework, we establish the asymptotic behavior of the aggregated debiased estimator $\overline{\beta}^d_v$ for every given $v \in [d]$.

Theorem 3.8. Under Conditions 2.1, 3.6 and 3.7, with $\lambda \asymp \sqrt{k\log d/n}$, we have

$$\overline{\beta}^d_v - \beta^*_v = -\frac{1}{k}\sum_{j=1}^{k}\widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla\,\ell^{(j)}_{n_k}(\boldsymbol{\beta}^*) + o_{\mathbb{P}}(n^{-1/2}) \tag{3.14}$$

for any $k \ll d$ satisfying $k = o\big(((s\vee s_1)\log d)^{-1}\sqrt{n}\big)$, where $\widehat{\boldsymbol{\Theta}}_v^{(j)}$ is the transposed v^{th} row of $\widehat{\Theta}^{(j)}$.

A corollary of Theorem 3.8 provides the asymptotic distribution of the Wald statistic in equation (3.13) under the null hypothesis.

Corollary 3.9. Let $\overline{S}_n$ be as in equation (3.13), with Θ^*_{vv} replaced with an estimator $\widetilde{\Theta}_{vv}$. Then under the conditions of Theorem 3.8 and $H_0 : \beta^*_v = \beta^H_v$, provided $|\widetilde{\Theta}_{vv} - \Theta_{vv}| = o_{\mathbb{P}}(1)$ under the scaling $k = o\big(((s \vee s_1)\log d)^{-1}\sqrt{n}\big)$, we have

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \big|\mathbb{P}\big(\overline{S}_n \le t\big) - \Phi(t)\big| = 0.$$

Remark 3.10. Although Theorem 3.8 and Corollary 3.9 are stated only for the GLM, their proofs are in fact an application of two more general results. Further details are available in Lemmas A.7 and A.8 of the appendix.

We return to the issue of estimating Θ^*_{vv} in Section 4, where we introduce an consistent estimator of Θ^*_{vv} that preserves the scaling of Theorem 3.8 and Corollary 3.9.

3.2 Divide and Conquer Score Test

In this section, we use $\nabla_v f(\boldsymbol{\beta})$ and $\nabla_{-v} f(\boldsymbol{\beta})$ to denote, respectively, the partial derivative of f with respect to β_v and the partial derivative vector of f with respect to $\boldsymbol{\beta}_{-v}$. $\nabla^2_{vv} f(\boldsymbol{\beta})$, $\nabla^2_{v,-v} f(\boldsymbol{\beta})$, $\nabla^2_{-v,v} f(\boldsymbol{\beta})$ and $\nabla^2_{-v,-v} f(\boldsymbol{\beta})$ are analogously defined.

In the low dimensional setting (where d is fixed), Rao's score test of $H_0 : \beta^*_v = \beta^H_v$ against $H_1 : \beta^*_v \neq \beta^H_v$ is based on $\nabla_v \ell_n(\beta^H_v, \widetilde{\boldsymbol{\beta}}_{-v})$, where $\widetilde{\boldsymbol{\beta}}_{-v}$ is a constrained maximum likelihood estimator of $\boldsymbol{\beta}^*_{-v}$, constructed as $\widetilde{\boldsymbol{\beta}}_{-v} = \operatorname{argmin}_{\boldsymbol{\beta}_{-v}} \ell_n(\beta^H_v, \boldsymbol{\beta}_{-v}) = \operatorname{argmax}_{\boldsymbol{\beta}_{-v}}\{-\ell_n(\beta^H_v, \boldsymbol{\beta}_{-v})\}$. If H_0 is false, imposing the constraint postulated by H_0 significantly violates the first order conditions from M-estimation with high probability; this is the principle underpinning the classical score test. Under regularity conditions, it can be shown (e.g. Cox and Hinkley, 1974) that

$$\sqrt{n}\big(\nabla_v \ell_n(\beta^H_v, \widetilde{\boldsymbol{\beta}}_{-v})\big) J^{*-1/2}_{v|-v} \rightsquigarrow N(0,1),$$

where $J^*_{v|-v}$ is given by $J^*_{v|-v} = J^*_{v,v} - \boldsymbol{J}^*_{v,-v} J^{*-1}_{-v,-v} \boldsymbol{J}^*_{-v,v}$, with $J^*_{v,v}$, $\boldsymbol{J}^*_{v,-v}$, $J^*_{-v,-v}$ and $\boldsymbol{J}^*_{-v,v}$ the partitions of the information matrix $J^* = J(\boldsymbol{\beta}^*)$,

$$J(\boldsymbol{\beta}) = \begin{pmatrix} J_{v,v} & \boldsymbol{J}_{v,-v} \\ \boldsymbol{J}_{-v,v} & J_{-v,-v} \end{pmatrix} = \begin{pmatrix} \mathbb{E}\nabla^2_{v,v}\ell_n(\boldsymbol{\beta}) & \mathbb{E}\nabla^2_{v,-v}\ell_n(\boldsymbol{\beta}) \\ \mathbb{E}\nabla^2_{-v,v}\ell_n(\boldsymbol{\beta}) & \mathbb{E}\nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}) \end{pmatrix}. \tag{3.15}$$

The problems associated with the use of the classical score statistic in the presence of a high dimensional nuisance parameter are brought to light by Ning and Liu (2014), who propose a remedy via the decorrelated score. The problem stems from the inversion of the matrix $J^*_{-v,-v}$ in high dimensions. The decorrelated score is defined as

$$S(\beta^*_v, \boldsymbol{\beta}^*_{-v}) = \nabla_v \ell_n(\beta^*_v, \boldsymbol{\beta}^*_{-v}) - \boldsymbol{w}^{*T} \nabla_{-v} \ell_n(\beta^*_v, \boldsymbol{\beta}^*_{-v}), \text{ where } \boldsymbol{w}^{*T} = \boldsymbol{J}^*_{v,-v} J^{*-1}_{-v,-v}. \tag{3.16}$$

For a regularized estimator $\widehat{\boldsymbol{w}}$ of $\boldsymbol{w}^*$, to be defined below, a mean value expansion of

$$\widehat{S}(\beta^*_v, \widehat{\boldsymbol{\beta}}^\lambda_{-v}) := \nabla_v \ell_n(\beta^*_v, \widehat{\boldsymbol{\beta}}^\lambda_{-v}) - \widehat{\boldsymbol{w}}^T \nabla_{-v} \ell_n(\beta^*_v, \widehat{\boldsymbol{\beta}}^\lambda_{-v}) \tag{3.17}$$

around $\boldsymbol{\beta}^*_{-v}$ gives

$$\begin{aligned} \widehat{S}(\beta^*_v, \widehat{\boldsymbol{\beta}}^\lambda_{-v}) \;=\; & \nabla_v \ell_n(\beta^*_v, \boldsymbol{\beta}^*_{-v}) - \widehat{\boldsymbol{w}}^T \nabla_{-v} \ell_n(\beta^*_v, \boldsymbol{\beta}^*_{-v}) \\ & + \big[\nabla^2_{v,-v}\ell_n(\beta^*_v, \boldsymbol{\beta}_{-v,\alpha}) - \widehat{\boldsymbol{w}}^T \nabla^2_{-v,-v}\ell_n(\beta^*_v, \boldsymbol{\beta}_{-v,\alpha})\big](\widehat{\boldsymbol{\beta}}^\lambda_{-v} - \boldsymbol{\beta}^*_{-v}), \end{aligned} \tag{3.18}$$

where $\boldsymbol{\beta}_{-v,\alpha} = \alpha\widehat{\boldsymbol{\beta}}^{\lambda}_{-v} + (1-\alpha)\boldsymbol{\beta}^*_{-v}$ for $\alpha \in [0,1]$. The key to understanding how the decorrelated score remedies the problems faced by the classical score test is the observation that

$$\begin{aligned} & \big[\nabla^2_{v,-v}\ell_n(\beta^*_v, \boldsymbol{\beta}_{-v,\alpha}) - \widehat{\boldsymbol{w}}^T\nabla^2_{-v,-v}\ell_n(\beta^*_v, \boldsymbol{\beta}_{-v,\alpha})\big] \\ \approx \;& \mathbb{E}\big[\nabla^2_{v,-v}\ell_n(\beta^*_v, \boldsymbol{\beta}^*_{-v}) - \boldsymbol{w}^{*T}\nabla^2_{-v,-v}\ell_n(\beta^*_v, \boldsymbol{\beta}^*_{-v})\big] = \boldsymbol{J}^*_{v,-v} - \boldsymbol{J}^*_{v,-v}J^{*-1}_{-v,-v}J^*_{-v,-v} = 0, \end{aligned} \tag{3.19}$$

where $\boldsymbol{w}^{*T} = \boldsymbol{J}^*_{v,-v}J^{*-1}_{-v,-v}$. Hence, provided $\boldsymbol{w}^*$ is sufficiently sparse to avoid excessive noise accumulation, we are able to achieve rate acceleration in equation (3.18), ultimately giving rise to a tractable limit distribution of a suitable rescaling of $\widehat{S}(\beta^*_v, \widehat{\boldsymbol{\beta}}^{\lambda}_{-v})$. Since β^*_v is restricted under the null hypothesis, $H_0 : \beta^*_v = \beta^H_v$, the statistic in equation (3.17) is accessible once H_0 is imposed. As Ning and Liu (2014) point out, $\boldsymbol{w}^*$ is the solution to

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, \mathbb{E}\big[\nabla_v\ell_n(\beta^H_v, \boldsymbol{\beta}^*_{-v}) - \boldsymbol{w}^T\nabla_{-v}\ell_n(\beta^H_v, \boldsymbol{\beta}^*_{-v})\big]^2$$

under $H_0 : \beta^*_v = \beta^H_v$. We thus see that the population analogue of the decorrelation device is the linear combination $\boldsymbol{w}^{*T}\nabla_{-v}\ell_n(\beta^H_v, \boldsymbol{\beta}^*_{-v})$ that best approximates $\nabla_v\ell_n(\beta^H_v, \boldsymbol{\beta}^*_{-v})$ in a least squares sense.

Our divide and conquer score statistic under $H_0 : \beta^*_v = \beta^H_v$ is

$$\overline{S}(\beta^H_v) = \frac{1}{k}\sum_{j=1}^{k}\widehat{S}^{(j)}\big(\beta^H_v, \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)\big), \tag{3.20}$$

where $\widehat{S}^{(j)}\big(\beta_v, \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)\big) = \nabla_v\ell^{(j)}_{n_k}\big(\beta_v, \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)\big) - \widehat{\boldsymbol{w}}(\mathcal{D}_j)^T\nabla_{-v}\ell^{(j)}_{n_k}\big(\beta_v, \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)\big)$ and

$$\widehat{\boldsymbol{w}}(\mathcal{D}_j) = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, \|\boldsymbol{w}\|_1, \text{ s.t. } \Big\|\nabla^2_{-v,v}\ell^{(j)}_{n_k}\big(\widehat{\beta}^{\lambda}_v(\mathcal{D}_j), \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)\big) - \boldsymbol{w}^T\nabla^2_{-v,-v}\ell^{(j)}_{n_k}\big(\widehat{\beta}^{\lambda}_v(\mathcal{D}_j), \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)\big)\Big\|_{\infty} \le \mu. \tag{3.21}$$

Equation (3.21) is the Dantzig selector of Candes and Tao (2007).

Theorem 3.11. Let $\widehat{J}_{v|-v}$ be a consistent estimator of $J^*_{v|-v}$ and

$$S^{(j)}(\beta^H_v, \boldsymbol{\beta}^*_{-v}) = \nabla_v\ell^{(j)}_{n_k}(\beta^H_v, \boldsymbol{\beta}^*_{-v}) - \boldsymbol{w}^{*T}\nabla_{-v}\ell^{(j)}_{n_k}(\beta^H_v, \boldsymbol{\beta}^*_{-v}).$$

Suppose $\|\boldsymbol{w}^*\|_1 \lesssim s_1$ and Conditions 2.1 and 3.6 are fulfilled. Then under $H_0 : \beta^*_v = \beta^H_v$ with $\lambda \asymp \mu \asymp \sqrt{k\log d/n}$,

$$\sqrt{n}\,\overline{S}(\beta^H_v) = \sqrt{n}\frac{1}{k}\sum_{j=1}^{k}S^{(j)}(\beta^H_v, \boldsymbol{\beta}^*_{-v}) + o_{\mathbb{P}}(1) \quad \text{and} \quad \lim_{n\to\infty}\sup_{t\in\mathbb{R}}\big|\mathbb{P}\big(\overline{S}(\beta^H_v)\widehat{J}^{-1/2}_{v|-v} \le t\big) - \Phi(t)\big| = 0,$$

for any $k \ll d$ satisfying $k = o\big(((s\vee s_1)\log d)^{-1}\sqrt{n}\big)$, where $\overline{S}(\beta^H_v)$ is defined in equation (3.20).

Remark 3.12. By the definition of $\boldsymbol{w}^*$ and the block matrix inversion formula for $\Theta^* = (J^*)^{-1}$, sparsity of $\boldsymbol{w}^*$ is implied by sparsity of Θ^* as assumed in van de Geer et al. (2014) and Condition 3.7 of Section 3.1.2. In turn, $\|\boldsymbol{w}^*\|_0 \lesssim s_1$ implies $\|\boldsymbol{w}^*\|_1 \lesssim s_1$ provided that the elements of $\boldsymbol{w}^*$ are bounded.

Remark 3.13. Although Theorem 3.11 is stated in the penalized GLM setting, the result holds more generally; further details are available in Lemma A.13 of Appendix A in the Supplementary Material.

To maintain the same computational complexity, an estimator of the conditional information needs to be constructed using a DC procedure. For this, we propose to use

$$\overline{J}_{v|-v} = k^{-1}\sum_{j=1}^{k}\big(\nabla^2_{v,v}\,\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v}) - \overline{\boldsymbol{w}}^T\,\nabla^2_{-v,v}\,\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v})\big),$$

where $\overline{\beta}^d_v = k^{-1}\sum_{j=1}^k \widehat{\beta}^d_v(\mathcal{D}_j)$, $\overline{\boldsymbol{\beta}}_{-v} = k^{-1}\sum_{j=1}^k \widehat{\boldsymbol{\beta}}^\lambda_{-v}(\mathcal{D}_j)$ and $\overline{\boldsymbol{w}} = k^{-1}\sum_{j=1}^k \widehat{\boldsymbol{w}}(\mathcal{D}_j)$. By Lemma 3.14, this estimator is asymptotically consistent.

Lemma 3.14. Suppose $\|\boldsymbol{w}^*\|_1 = O(s_1)$ and Conditions 2.1 and 3.6 are fulfilled. Then for any $k \ll d$ satisfying $k = o\big(((s \vee s_1)\log d)^{-1}\sqrt{n}\big)$, $|\overline{J}_{v|-v} - J^*_{v|-v}| = o_{\mathbb{P}}(1)$.

4 Accuracy of Distributed Estimation

As explained in Section 2.2, the information got lost in the divide-and-conquer process is not very much. This motivates us to consider $\|\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d\|_2$, the loss incurred by the divide and conquer strategy in comparison with the computationally infeasible full sample debiased estimator $\widehat{\boldsymbol{\beta}}^d$. Indeed, it turns out that, for k not too large, $\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d$ appears only as a higher order term in the decomposition of $\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*$ and thus $\|\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d\|_2$ is negligible compared to the statistical error, $\|\widehat{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_2$. In other words, the divide-and-conquer errors are statistically negligible.

When the minimum signal strength is sufficiently strong, thresholding $\overline{\boldsymbol{\beta}}^d$ achieves exact support recovery, motivating a refitting procedure based on the low dimensional selected variables. As a means to understanding the theoretical properties of this refitting procedure, as well as for independent interest, this section develops new theory and methodology for the low dimensional ($d < n$) linear and generalized linear models in addition to their high dimensional ($d \gg n$) counterparts. It turns out that simple averaging of low dimensional OLS or GLM estimators (denoted uniformly as $\widehat{\boldsymbol{\beta}}^{(j)}$, without superscript d as debias is not necessary) suffices to preserve the statistical error, i.e., achieving the same statistical accuracy as the estimator based on the whole data set. This is because, in contrast to the high dimensional setting, parameters are not penalized in the low dimensional case. With $\overline{\boldsymbol{\beta}}$ the average of $\widehat{\boldsymbol{\beta}}^{(j)}$ over the k machines and $\widehat{\boldsymbol{\beta}}$ the full sample counterpart ($k = 1$), we derive the rate of convergence of $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2$. Refitted estimation using only the selected covariates allows us to eliminate a $\log d$ term in the statistical rate of convergence of the estimator. We present the high dimensional and low dimensional results separately, with the analysis of the refitting procedures appearing as corollaries to the low dimensional analysis.

4.1 The High-Dimensional Linear Model

Recall that the high dimensional DC estimator is $\overline{\boldsymbol{\beta}}^d = k^{-1}\sum_{j=1}^k \widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$, where $\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$ for $1 \le j \le k$ is the debiased estimator defined in (3.3). We also denote the debiased LASSO estimator

using the entire dataset as $\widehat{\boldsymbol{\beta}}^d = \widehat{\boldsymbol{\beta}}^d(\cup_{j=1}^k \mathcal{D}_j)$. The following lemma shows that not only is $\overline{\boldsymbol{\beta}}^d$ asymptotically normal, it approximates the full sample estimator $\widehat{\boldsymbol{\beta}}^d$ so well that it has the same statistical error as $\widehat{\boldsymbol{\beta}}^d$ provided the number of subsamples k is not too large.

Lemma 4.1. Under the Conditions 3.1 and 3.2, if λ, ϑ_1 and ϑ_2 are chosen as $\lambda \asymp \sqrt{k \log d/n}$, $\vartheta_1 \asymp \sqrt{k \log d/n}$ and $\vartheta_2 n^{-1/2} = o(1)$, we have with probability $1 - c/d$,

$$\|\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d\|_\infty \leq C\frac{sk \log d}{n} \text{ and } \|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty \leq C\Big(\sqrt{\frac{\log d}{n}} + \frac{sk \log d}{n}\Big). \tag{4.1}$$

Remark 4.2. The term $\sqrt{\frac{\log d}{n}}$ in (4.1) is the estimation error of $\|\widehat{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty$. Lemma 4.1 does not rely on any specific choice of k, however, in order for the aggregated estimator $\overline{\boldsymbol{\beta}}^d$ to attain the same $\|\cdot\|_\infty$ norm estimation error as the full sample LASSO estimator, $\widehat{\boldsymbol{\beta}}_{\text{LASSO}}$, the required scaling is $k = O(\sqrt{n/(s^2 \log d)})$. This is a weaker scaling requirement than that of Theorem 3.3 because the latter entails a guarantee of asymptotic normality, which is a stronger result. It is for the same reason that our estimation results only require $O(\cdot)$ scaling whilst those for testing require $o(\cdot)$ scaling.

Although $\overline{\boldsymbol{\beta}}^d$ achieves the same rate as the LASSO estimator under the infinity norm, it cannot achieve the minimax rate in ℓ_2 norm since it is not a sparse estimator. To obtain an estimator with the ℓ_2 minimax rate, we sparsify $\overline{\boldsymbol{\beta}}^d$ by hard thresholding. For any $\boldsymbol{\beta} \in \mathbb{R}^d$, define the hard thresholding operator $\mathcal{T}_\nu$ such that the j-th entry of $\mathcal{T}_\nu(\boldsymbol{\beta})$ is

$$[\mathcal{T}_\nu(\boldsymbol{\beta})]_j = \boldsymbol{\beta}_j \mathbb{1}\{|\boldsymbol{\beta}_j| \geq \nu\}, \text{ for } 1 \leq j \leq d. \tag{4.2}$$

According to (4.1), if $\boldsymbol{\beta}_j^* = 0$, we have $|\overline{\boldsymbol{\beta}}_j^d| \leq C(\sqrt{\log d/n} + sk \log d/n)$ with high probability. The following theorem characterizes the estimation rate of the thresholded estimator $\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)$.

Theorem 4.3. Suppose Conditions 3.1 and 3.2 are fulfilled and choose $\lambda \asymp \sqrt{k \log d/n}$, $\vartheta_1 \asymp \sqrt{k \log d/n}$ and $\vartheta_2 n^{-1/2} = o(1)$. Take the parameter of the hard threshold operator in (4.2) as $\nu = C_0\sqrt{\log d/n}$ for some sufficiently large constant C_0. If the number of subsamples satisfies $k = O(\sqrt{n/(s^2 \log d)})$, for large enough d and n, we have with probability $1 - c/d$,

$$\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_\nu(\widehat{\boldsymbol{\beta}}^d)\|_2 \leq C\frac{s^{3/2}k \log d}{n}, \ \|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_\infty \leq C\sqrt{\frac{\log d}{n}} \text{ and } \|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2 \leq C\sqrt{\frac{s \log d}{n}}. \tag{4.3}$$

Remark 4.4. In fact, in the proof of Theorem 4.3, we show that if the thresholding parameter ν satisfies $\nu \geq \|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty$, we have $\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2 \leq 2\sqrt{2s} \cdot \nu$; it is for this reason that we choose $\nu \asymp \sqrt{\log d/n}$. Unfortunately, the constant is difficult to choose in practice. In the following paragraphs we propose a practical method to select the tuning parameter ν.

Let $(M^{(j)}X^{(j)T})_\ell$ denote the transposed ℓ^{th} row of $M^{(j)}X^{(j)T}$. Inspection of the proof of Theorem 3.3 reveals that the leading term of term of $\sqrt{n}\|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty$ satisfies

$$T_0 = \max_{1 \leq \ell \leq d} \frac{1}{\sqrt{k}} \sum_{j=1}^k \frac{1}{\sqrt{n_k}} (M^{(j)}X^{(j)T})_\ell^T \boldsymbol{\varepsilon}^{(j)}.$$

Chernozhukov et al. (2013) propose the Gaussian multiplier bootstrap to estimate the quantile of T_0. Let $\{\xi_i\}_{i=1}^n$ be i.i.d. standard normal random variable independent of $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$. Consider the statistic

$$W_0 = \max_{1\le \ell\le d} \frac{1}{\sqrt{k}} \sum_{j=1}^{k} \frac{1}{\sqrt{n_k}} (M^{(j)} X^{(j)T})_\ell^T (\widehat{\boldsymbol{\varepsilon}}^{(j)} \circ \boldsymbol{\xi}^{(j)}),$$

where $\widehat{\boldsymbol{\varepsilon}}^{(j)} \in \mathbb{R}^{n_k}$ is an estimator of $\boldsymbol{\varepsilon}^{(j)}$ such that for any $i \in \mathcal{I}_j$, $\widehat{\varepsilon}_i^{(j)} = Y_i^{(j)} - \boldsymbol{X}_i^{(j)} \widehat{\boldsymbol{\beta}}(\mathcal{D}_j)$, and $\boldsymbol{\xi}^{(j)}$ is a subvector of $\{\xi_i\}_{i=1}^n$ with indices in $\mathcal{I}_j$. Recall that "$\circ$" denotes the Hadamard product. The α-quantile of W_0 conditioning on $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$ is defined as $c_{W_0}(\alpha) = \inf\{t \,|\, \mathbb{P}\big(W_0 \le t \,|\, \boldsymbol{Y}, X\big) \ge \alpha\}$. We can estimate $c_{W_0}(\alpha)$ by Monte-Carlo and thus choose $\nu_0 = c_{W_0}(\alpha)/\sqrt{n}$. This choice ensures

$$\big\|\mathcal{T}_{\nu_0}(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\big\|_2 = O_{\mathbb{P}}(\sqrt{s \log d/n}),$$

which coincides with the ℓ_2 convergence rate of the LASSO.

Remark 4.5. Lemma 4.1 and Theorem 4.3 show that if the number of subsamples satisfies $k = o(\sqrt{n/(s^2 \log d)})$, $\|\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d\|_\infty = o_{\mathbb{P}}\big(\sqrt{\log d/n}\big)$ and $\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_\nu(\widehat{\boldsymbol{\beta}}^d)\|_2 = o_{\mathbb{P}}(\sqrt{s \log d/n})$, and thus the error incurred by the divide and conquer procedure is negligible compared to the statistical minimax rate. The reason for this contraction phenomenon is that $\overline{\boldsymbol{\beta}}^d$ and $\widehat{\boldsymbol{\beta}}^d$ share the same leading term in their Taylor expansions around $\boldsymbol{\beta}^*$. The difference between them is only the difference of two remainder terms which is smaller order than the leading term. We uncover a similar phenomenon in the low dimensional case covered in Section 4.3. However, in the low dimensional case ℓ_2 norm consistency is automatic while the high dimensional case requires an additional thresholding step to guarantee sparsity and, consequently, ℓ_2 norm consistency.

4.2 The High-Dimensional Generalized Linear Model

We can generalize the DC estimation of the linear model to GLM. Recall that $\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$ is the debiased estimator defined in (3.9) and the aggregated estimator is $\overline{\boldsymbol{\beta}}^d = k^{-1}\sum_{j=1}^k \widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$. We still denote $\widehat{\boldsymbol{\beta}}^d = \widehat{\boldsymbol{\beta}}^d(\cup_{j=1}^k \mathcal{D}_j)$. The next lemma bounds the error incurred by splitting the sample and the statistical rate of convergence of $\overline{\boldsymbol{\beta}}^d$ in terms of the infinity norm.

Lemma 4.6. Under Conditions 2.1, 3.6 and 3.7, for $\widehat{\boldsymbol{\beta}}^\lambda$ with $\lambda \asymp \sqrt{k \log d/n}$, we have with probability $1 - c/d$,

$$\big\|\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d\big\|_\infty \le C\frac{(s \vee s_1) k \log d}{n} \text{ and } \big\|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\big\|_\infty \le C\Big(\sqrt{\frac{\log d}{n}} + \frac{(s \vee s_1) k \log d}{n}\Big). \tag{4.4}$$

Applying a similar thresholding step as in the linear model, we obtain the following estimation rate in ℓ_2 norm.

Theorem 4.7. Under Conditions 2.1 - 3.7, choose $\lambda \asymp \sqrt{k \log d/n}$ and $\lambda_v \asymp \sqrt{k \log d/n}$. Take the parameter of the hard threshold operator in (4.2) as $\nu = C_0\sqrt{\log d/n}$ for some sufficiently large constant C_0. If the number of subsamples satisfies $k = O(\sqrt{n/((s \vee s_1)^2 \log d)})$, for large enough d and n, we have with probability $1 - c/d$,

$$\big\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_\nu(\widehat{\boldsymbol{\beta}}^d)\big\|_2 \le C\frac{(s \vee s_1) s^{1/2} k \log d}{n}, \quad \big\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\big\|_\infty \le C\sqrt{\frac{\log d}{n}} \tag{4.5}$$

and $\left\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\right\|_2 \leq C\sqrt{s\log d/n}$.

Remark 4.8. As in the case of the linear model, Theorem 4.7 reveals that the loss incurred by the divide and conquer procedure is negligible compared to the statistical minimax estimation error provided $k = o\big(\sqrt{n/(s_1 \vee s)^2 s\log d}\big)$.

A similar proof strategy to that of Theorem 4.7 allows us to construct an estimator of Θ^*_{vv} that achieves the required consistency with the scaling of Corollary 3.9. Our estimator is $\widetilde{\Theta}_{vv} := \big[\mathcal{T}_\zeta(\overline{\Theta})\big]_{vv}$, where $\overline{\Theta} = k^{-1}\sum_{j=1}^k \widehat{\Theta}^{(j)}$ and $\mathcal{T}_\zeta(\cdot)$ is the thresholding operator defined in equation (4.2) with $\zeta = C_1\sqrt{\log d/n}$ for some sufficiently large constant C_1.

Corollary 4.9. Under the conditions and scaling of Theorem 3.8, $\big|\widetilde{\Theta}_{vv} - \Theta^*_{vv}\big| = o_{\mathbb{P}}(1)$.

Substituting this estimator in Corollary 3.9 delivers a practically implementable test statistic based on $k = o\big(((s \vee s_1)\log d)^{-1}\sqrt{n}\big)$ subsamples.

4.3 The Low-Dimensional Linear Model

As mentioned earlier, the infinity norm bound derived in Lemma 4.1 can be used to do model selection, after which the selected support can be shared across all the local agents, significantly reducing the dimension of the problem as we only need to refit the data on the selected model. The remaining challenge is to implement the divide and conquer strategy in the low dimensional setting, which is also of independent interest. Here we focus on the linear model, while the generalized linear model is covered in Section 4.4.

In this section d still stands for dimension, but in contrast with the rest of this paper in which $d \gg n$, here we consider $d < n$. More specifically, we consider the linear model (3.2) with $d < n$ and i.i.d sub-gaussian noise $\{\varepsilon_i\}_{i=1}^n$. It is well known that the ordinary least square (OLS) estimator of $\boldsymbol{\beta}^*$ is defined as $\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{Y}$. In the massive data setting, the communication cost of estimating and inverting covariance matrices is very high (order $O(kd^2)$). However, as pointed out by Chen and Xie (2012), this estimator exactly coincides with the DC estimator,

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{j=1}^k X^{(j)T}X^{(j)}\right)^{-1}\sum_{j=1}^k X^{(j)T}\boldsymbol{Y}^{(j)}.$$

In this section, we study the DC strategy to approximate $\widehat{\boldsymbol{\beta}}$ with the communication cost only $O(kd)$, which implies that we can only communicate d dimensional vectors.

The OLS estimator based on the subsample $\mathcal{D}_j$ is defined as $\widehat{\boldsymbol{\beta}}(\mathcal{D}_j) = (X^{(j)T}X^{(j)})^{-1}X^{(j)T}\boldsymbol{Y}^{(j)}$. In order to estimate $\boldsymbol{\beta}^*$, a simple and natural idea is to take the average of $\{\widehat{\boldsymbol{\beta}}(\mathcal{D}_j)\}_{j=1}^k$, which we denote by $\overline{\boldsymbol{\beta}}$. The question is whether this estimator preserves the statistical error as $\widehat{\boldsymbol{\beta}}$. The following theorem gives an upper bound of the gap between $\overline{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$, and shows that this gap is negligible compared with the statistical error of $\widehat{\boldsymbol{\beta}}$ as long as k is not large.

Theorem 4.10. Consider the linear model (3.2). Suppose Conditions 3.1 and 3.2 hold and $\{\varepsilon_i\}_{i=1}^n$ are i.i.d sub-Gaussian random variables with $\|\varepsilon_i\|_{\psi_2} \leq \sigma_1$. If the number of subsamples satisfies

$k = O(nd/(d \vee \log n)^2)$, then for sufficiently large n and d it follows that

$$\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 = O_{\mathbb{P}}\Big(\frac{\sqrt{k}(d \vee \log n)}{n}\Big), \quad \|\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}\big(\sqrt{d/n}\big). \tag{4.6}$$

Remark 4.11. By taking $k = o\big(nd/(d \vee \log n)^2\big)$, the loss incurred by the divide and conquer procedure, i.e., $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2$, converges at a faster rate than the statistical error of the full sample estimator $\widehat{\boldsymbol{\beta}}$.

We now take a different viewpoint by returning to the high dimensional setting of Section 4.1 ($d \gg n$) and applying Theorem 4.10 in the context of a refitting estimator. In this refitting setting, the sparsity s of Lemma 4.1 becomes the dimension of a low dimensional parameter estimation problem on the selected support. Our refitting estimator is defined as

$$\overline{\boldsymbol{\beta}}^r := \frac{1}{k}\sum_{j=1}^{k}(X_{\widehat{S}}^{(j)T}X_{\widehat{S}}^{(j)})^{-1}X_{\widehat{S}}^{(j)T}\boldsymbol{Y}^{(j)}, \tag{4.7}$$

where $\widehat{S} := \{j : |\overline{\beta}_j^d| > 2C\sqrt{\log d/n}\}$ and C is the same constant as in (4.1).

Corollary 4.12. Suppose $\beta^*_{\min} > 2C\sqrt{\log d/n}$, where $\beta^*_{\min} := \min_{1\le j\le d}|\beta_j^*|$ and C is the same constant as in (4.1). Define the full sample oracle estimator as $\widehat{\boldsymbol{\beta}}^o = (X_S^TX_S)^{-1}X_S^T\boldsymbol{Y}$, where S is the true support of $\boldsymbol{\beta}^*$. If $k = O(\sqrt{n/(s^2\log d)})$, then for sufficiently large n and d we have

$$\|\overline{\boldsymbol{\beta}}^r - \widehat{\boldsymbol{\beta}}^o\|_2 = O_{\mathbb{P}}\Big(\frac{\sqrt{k}(s \vee \log n)}{n}\Big), \quad \|\overline{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}\big(\sqrt{s/n}\big). \tag{4.8}$$

We see from Corollary 4.12 that $\overline{\boldsymbol{\beta}}^r$ achieves the oracle rate when the minimum signal strength is not too weak and the number of subsamples k is not too large.

4.4 The Low-Dimensional Generalized Linear Model

The next theorem quantifies the gap between $\overline{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$, where $\overline{\boldsymbol{\beta}}$ is the average of subsambled GLM estimators and $\widehat{\beta}$ is the full sample GLM estimator.

Theorem 4.13. Under Condition 3.6, if $k = O(\sqrt{n}/(d \vee \log n))$, then we have for sufficiently large d and n,

$$\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 = O_{\mathbb{P}}\Big(\frac{k\sqrt{d}(d \vee \log n)}{n}\Big), \quad \|\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}\big(\sqrt{d/n}\big). \tag{4.9}$$

Remark 4.14. In analogy to Theorem 4.10, by constraining the growth rate of the number of subsamples according to $k = o\big(\sqrt{n}/(d \vee \log n)\big)$, the error incurred by the divide and conquer procedure, i.e., $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2$ decays at a faster rate than that of the statistical error of the full sample estimator $\widehat{\boldsymbol{\beta}}$.

As in the linear model, Lemma 4.6 together with Theorem 4.13 allow us to study the theoretical properties of a refitting estimator for the high dimensional GLM. Estimation on the estimated

support set is again a low dimensional problem, thus the d of Theorem 4.13 corresponds to the s of Lemma 4.6 in this refitting setting. The refitted GLM estimator is defined as

$$\overline{\boldsymbol{\beta}}^r = \frac{1}{k}\sum_{j=1}^{k} \widehat{\boldsymbol{\beta}}^r(\mathcal{D}_j), \tag{4.10}$$

where $\widehat{\boldsymbol{\beta}}^r(\mathcal{D}_j) = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d, \boldsymbol{\beta}_{\widehat{S}^c}=0} \ell_{n_k}^{(j)}(\boldsymbol{\beta})$ and $\widehat{S} := \{j : |\overline{\boldsymbol{\beta}}_j^d| > 2C\sqrt{\log d/n}\}$. The following corollary quantifies the statistical rate of $\overline{\boldsymbol{\beta}}^r$.

Corollary 4.15. Suppose $\beta^*_{\min} > 2C\sqrt{\log d/n}$, where $\beta^*_{\min} := \min_{1\le j\le d}|\beta_j^*|$ and C is the same constant as in (4.4). Define the full sample oracle estimator as $\widehat{\boldsymbol{\beta}}^o = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d, \boldsymbol{\beta}_{S^c}=0} \ell_n(\boldsymbol{\beta})$, where S is the true support of $\boldsymbol{\beta}^*$. If $k = O\big(\sqrt{n/((s\vee s_1)^2\log d)}\big)$, then for sufficiently large n and d we have

$$\|\overline{\boldsymbol{\beta}}^r - \widehat{\boldsymbol{\beta}}^o\|_2 = O_{\mathbb{P}}\Big(\frac{k\sqrt{s}(s\vee\log n)}{n}\Big), \quad \|\overline{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}\big(\sqrt{s/n}\big). \tag{4.11}$$

We thus see that $\overline{\boldsymbol{\beta}}^r$ achieves the oracle rate when the minimum signal strength is not too weak and the number of subsamples k is not too large.

5 Numerical Analysis

In this section, we illustrate and validate our theoretical findings through simulations. For inference, we use QQ plots to compare the distribution of p-values for divide and conquer test statistics to their theoretical uniform distribution. We also investigate the estimated type I error and power of the divide and conquer tests. For estimation, we validate our claim of Sections 4.3 and 4.4 that the loss incurred by the divide and conquer strategy is negligible compared to the statistical error of the corresponding full sample estimator in the low dimensional case. An analogous empirical verification of the theory is performed for the high dimensional case, where we also compare the performance of the divide and conquer thresholding estimator of Section 4.1 to the full sample LASSO and the average LASSO over subsamples.

5.1 Results on Inference

We explore the probability of rejection of a null hypothesis of the form $H_0 : \beta_1^* = 0$ when data $(Y_i, \boldsymbol{X}_i)_{i=1}^n$ are generated according to the linear model,

$$\boldsymbol{Y}_i = \boldsymbol{X}_i^T\boldsymbol{\beta}^* + \varepsilon_i, \quad \varepsilon_i \sim N\big(0, \sigma_\varepsilon^2\big),$$

for $\sigma_\varepsilon^2 = 1$ and $\boldsymbol{\beta}^*$ an s sparse d dimensional vector with $d = 850$ and $s = 3$. In each Monte Carlo replication, we split the initial sample of size n into k subsamples of size n/k. In particular we choose $n = 840$ because it has a large number of factors $k \in \{1, 2, 5, 10, 15, 20, 24, 28, 30, 35, 40\}$. The number of simulations is 250. Using $\widehat{\boldsymbol{\beta}}_{\text{LASSO}}$ as a preliminary estimator of $\boldsymbol{\beta}^*$, we construct Wald and Rao score test statistics as described in Sections 3.1.2 and 3.2 respectively.

Panels (A) and (B) of Figure 1 are QQ plots of the p-values of the divide and conquer Wald and score test statistics under the null hypothesis against the theoretical quantiles of of the uniform

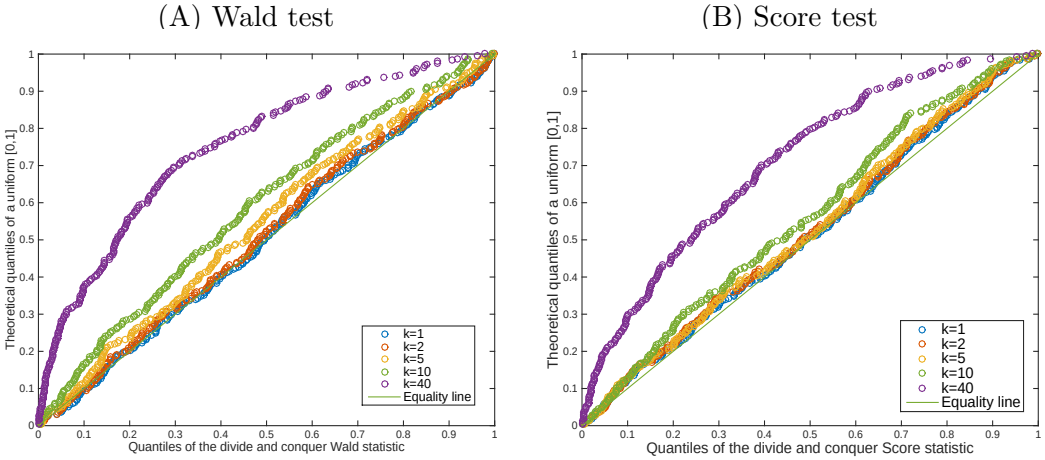

Figure 1: QQ plots of the p-values of the Wald (A) and score (B) divide and conquer test statistics against the theoretical quantiles of the uniform [0,1] distribution under the null hypothesis.

[0,1] distribution for four different values of k. For both test constructions, the distributions of the p-values are close to uniform and remain so as we split the data set. When $k = 40$, the distribution of the corresponding p-values is visibly non-uniform, as expected from the theory developed in Sections 3.1.2 and 3.2. Panel (A) of Figure 2 also supports our theoretical findings, showing that, for both test constructions, the empirical level of the test is close to both the nominal $\alpha = 0.05$ level and the level of the full sample oracle OLS estimator which knows the true support of $\boldsymbol{\beta}^*$. Moreover, it remains at this level as long as we do not split the data set too many times. Panel (B) of Figure 2 displays the power of the test for two different signal strengths, $\beta_1^* = 0.125$ and $\beta_1^* = 0.15$. We see that the power is also comparable with that of the full sample oracle OLS estimator which knows the true support of $\boldsymbol{\beta}^*$.

5.2 Results on Estimation

In this section, we turn our attention to experimental validation of our divide and conquer estimation theory, focusing first on the low dimensional case and then on the high dimensional case.

5.2.1 The Low-Dimensional Linear Model

All $n \times d$ entries of the design matrix X are generated as i.i.d. standard normal random variables and the errors $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. standard normal as well. The true regression vector $\boldsymbol{\beta}^*$ satisfies $\beta_j^* = 10/\sqrt{d}$ for $j = 1, \ldots, d/2$ and $\beta_j^* = -10/\sqrt{d}$ for $j > d/2$, which guarantees that $\|\boldsymbol{\beta}^*\|_2 = 10$. Then we generate the response variable $\{Y_i\}_{i=1}^n$ according to the model (3.2). Denote the full sample ordinary least-squares estimator and the divide and conquer estimator by $\widehat{\boldsymbol{\beta}}$ and $\overline{\boldsymbol{\beta}}$ respectively. Figure 3(A) illustrates the change in the ratio $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 / \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ as the sample size increases, where k assumes three different growth rates and $d = \sqrt{n}/2$. Figure 3(B) focuses on the relationship between the statistical error of $\overline{\boldsymbol{\beta}}$ and $\log k$ under three different scalings of n and d. All the data points are obtained based on average over 100 Monte Carlo replications.

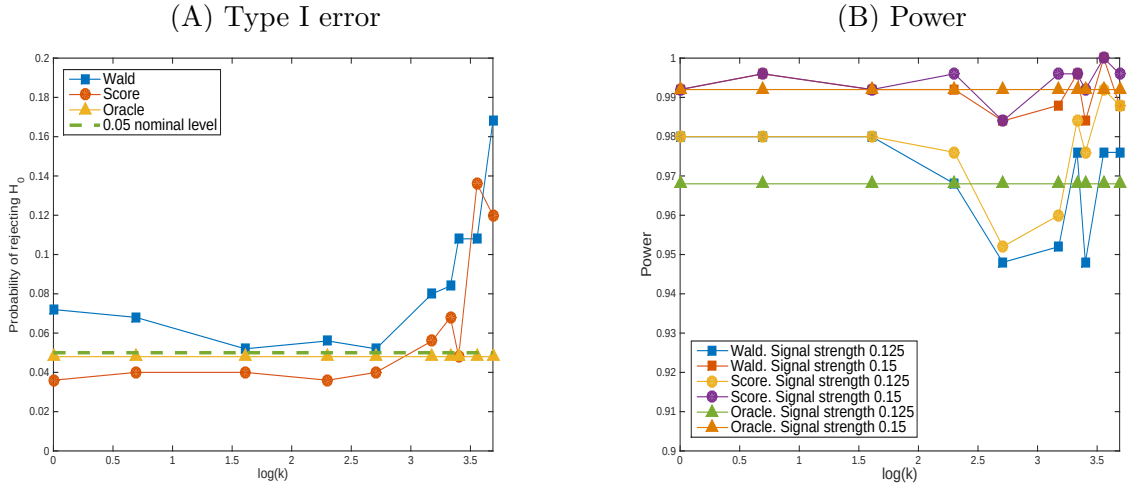

Figure 2: (A) Estimated probabilities of type I error for the Wald and score tests as a function of k. (B) Estimated power with signal strength 0.125 and 0.15 for the Wald, and score tests as a function of k.

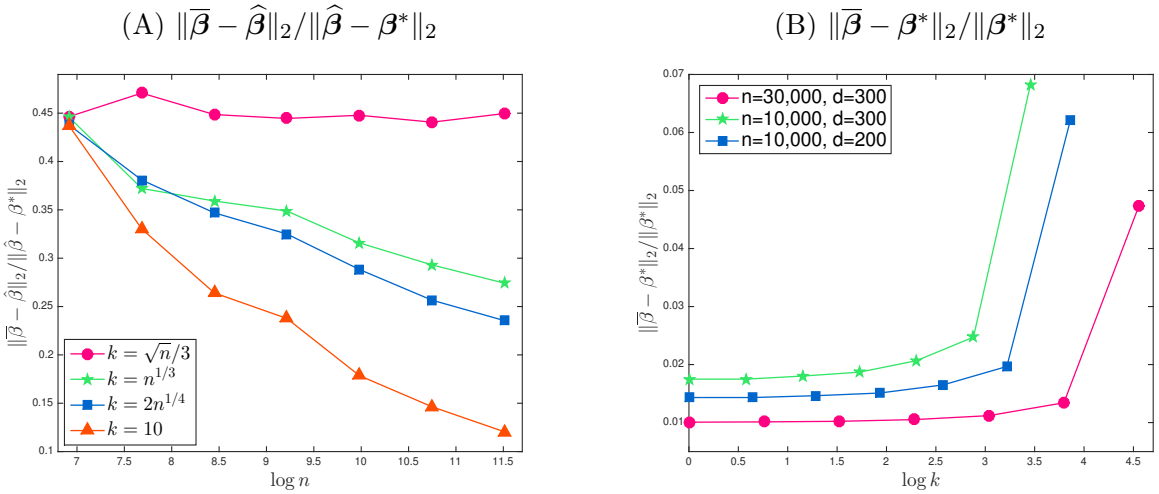

Figure 3: (A) The ratio between the loss of the divide and conquer procedure and the statistical error of the estimator based on the whole sample with $d = \sqrt{n}/2$ and different growth rates of k. (B) Statistical error of the DC estimator against $\log k$.

As Figure 3(A) demonstrates, when $k = O(n^{1/3})$, $O(n^{1/4})$ or $O(1)$, the ratio decreases with ever faster rates, which is consistent with the argument of Remark 4.11 that the ratio goes to zero when $k = o(n/d) = o(\sqrt{n})$. When $k = O(\sqrt{n})$, however, we observe that the ratio is essentially constant, which suggests the rate we derived in Theorem 4.10 is sharp.

From Figure 3(B), we see that when k is not large, the statistical error of $\overline{\boldsymbol{\beta}}$ is very small because the loss incurred by the divide and conquer procedure is negligible compared to the statistical error of $\widehat{\boldsymbol{\beta}}$. However, when k is larger than a threshold, there is a surge in the statistical error, since the loss of the divide and conquer begins to dominate the statistical error of $\widehat{\boldsymbol{\beta}}$. We also notice that the

(A) $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 / \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ (B) $\|\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 / \|\boldsymbol{\beta}^*\|_2$

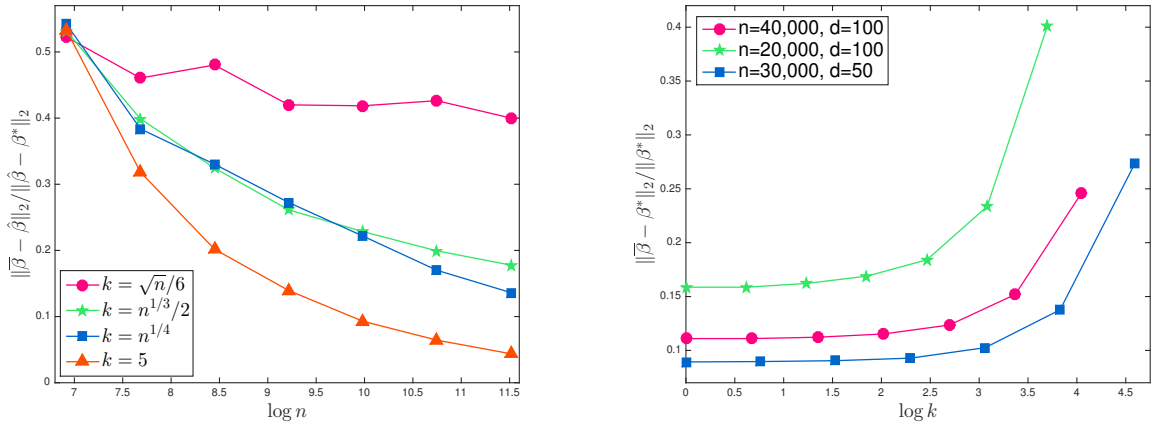

Figure 4: (A) The ratio between the loss of the divide and conquer procedure and the statistical error of the estimator based on the whole sample when $d = 20$. (B) Statistical error of the DC estimator.

larger the ratio n/d, the larger the threshold of $\log k$, which is again consistent with Remark 4.11.

5.2.2 The Low-Dimensional Logistic Regression

In logistic regression, given covariates $\boldsymbol{X}$, the response $Y|\boldsymbol{X} \sim \text{Ber}(\eta(\boldsymbol{X}))$, where $\text{Ber}(\eta)$ denotes the Bernoulli distribution with expectation η and

$$\eta(\boldsymbol{X}) = \frac{1}{1 + \exp(-\boldsymbol{X}^T \boldsymbol{\beta}^*)}.$$

We see that $\text{Ber}(\eta(\boldsymbol{X}))$ is in exponential dispersion family canonical form (2.7) with $b(\theta) = \log(1 + e^\theta)$, $\phi = 1$ and $c(y) = 1$. The use of the canonical link,

$$\eta(\boldsymbol{X}) = \frac{1}{1 + e^{-\theta(\boldsymbol{X})}},$$

leads to the simplification $\theta(\boldsymbol{X}) = \boldsymbol{X}^T \boldsymbol{\beta}^*$.

In our Monte Carlo experiments, all $n \times d$ entries of the design matrix X are generated as i.i.d. standard normal random variables. The true regression vector $\boldsymbol{\beta}^*$ satisfies $\beta_j^* = 1/\sqrt{d}$ for $j \le d/2$ and $\beta_j^* = -1/\sqrt{d}$ for $j > d/2$, which guarantees that $\|\boldsymbol{\beta}^*\|_2 = 1$. Finally, we generate the response variables $\{Y_i\}_{i=1}^n$ according to $\text{Ber}(\eta(\boldsymbol{X}))$. Figure 4(A) illustrates the change of the ratio $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 / \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ as the sample size increases, where k assumes three different growths rates and $d = 20$. Figure 4(B) focuses on the relationship between the statistical error of $\overline{\boldsymbol{\beta}}$ and $\log k$ under three different scalings of n and d. All the data points are obtained based on an average over 100 Monte Carlo replications.

Figure 4 reveals similar phenomena to those revealed in Figure 3 of the previous subsection. More specifically, Figure 4(A) shows that when $k = O(n^{1/3})$, $O(n^{1/4})$ or $O(1)$, the ratio decreases with even faster rates, which is consistent with the argument of Remark 4.14 that the ratio converges

to zero when $k = o(\sqrt{n}/d) = o(\sqrt{n})$. When $k = O(\sqrt{n})$, however, we observe that the ratio remains essentially constant when $\log n$ is large, which suggests the rate we derived in Theorem 4.10 is sharp.

As for Figure 4(B), we again observe that the statistical error of $\overline{\boldsymbol{\beta}}$ is very small when k is sufficiently small, but grows fast when k becomes large. The reasoning is the same as in the linear model, i.e. when k is large, the loss incurred by the divide and conquer procedure is non-negligible as compared with the statistical error of $\|\widehat{\boldsymbol{\beta}}\|_2$. In addition, as Figure 4(B) reveals, the larger is $\sqrt{n}/d$, the larger the threshold of k, which is again consistent with the threshold rate pointed out in Remark 4.14.

5.2.3 The High Dimensional Linear Model

We now consider the same setting of Section 5.1 with $n = 1400$, $d = 1500$ and $\beta_j^* = 10$ for all j in the support of $\boldsymbol{\beta}^*$. In this context, we analyze the performance of the thresholded averaged debiased estimator of Section 4.1. Figure 5(A) depicts the average over 100 Monte Carlo replications of $\|\boldsymbol{b} - \boldsymbol{\beta}^*\|_2$ for three different estimators: debiased divide-and-conquer $\boldsymbol{b} = \mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)$, the LASSO estimator based on the whole sample $\boldsymbol{b} = \widehat{\boldsymbol{\beta}}_{\text{LASSO}}$ and the estimator obtained by naïvely averaging the LASSO estimators from the k subsamples $\boldsymbol{b} = \overline{\boldsymbol{\beta}}_{\text{LASSO}}$. The parameter ν is taken as $\nu = \sqrt{\log d/n}$ in the specification of $\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)$. As expected, the performance of $\overline{\boldsymbol{\beta}}_{\text{LASSO}}$ deteriorates sharply as k increases. $\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)$ outperforms $\widehat{\boldsymbol{\beta}}_{\text{LASSO}}$ as long as k is not too large. This is expected because, for sufficiently large signal strength, both $\widehat{\boldsymbol{\beta}}_{\text{LASSO}}$ and $\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)$ recover the correct support, however $\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)$ is unbiased for those β_j^* in the support of $\boldsymbol{\beta}^*$, whilst $\widehat{\boldsymbol{\beta}}_{\text{LASSO}}$ is biased. Figure 5(B) shows the error incurred by the divide and conquer procedure $\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_\nu(\widehat{\boldsymbol{\beta}}^d)\|_2$ relative to the statistical error of the full sample estimator, $\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2$, for four different scalings of k. We observe that, with $k = O(\sqrt{n/s^2 \log d})$ and n not too small, the relative error incurred by the divide and conquer procedure is essentially constant across n, demonstrating the theory developed in Theorem 4.3.

6 Discussion

With the advent of the data revolution comes the need to modernize the classical statistical toolkit. For very large scale datasets, distribution of data across multiple machines is the only practical way to overcome storage and computational limitations. It is thus essential to build aggregation procedures for conducting inference based on the combined output of multiple machines. We successfully achieve this objective, deriving divide and conquer analogues of the Wald and score statistics and providing statistical guarantees on their performance as the number of sample splits grows to infinity with the full sample size. Tractable limit distributions of each DC test statistic are derived. These distributions are valid as long as the number of subsamples, k, does not grow too quickly. In particular, $k = o\big(((s \vee s_1) \log d)^{-1}\sqrt{n}\big)$ is required in a general likelihood based framework. If k grows faster than $((s \vee s_1) \log d)^{-1}\sqrt{n}$, remainder terms become non-negligible and contaminate the tractable limit distribution of the leading term. When attention is restricted to the linear model, a faster growth rate of $k = o\big((s \log d)^{-1}\sqrt{n}\big)$ is allowed.

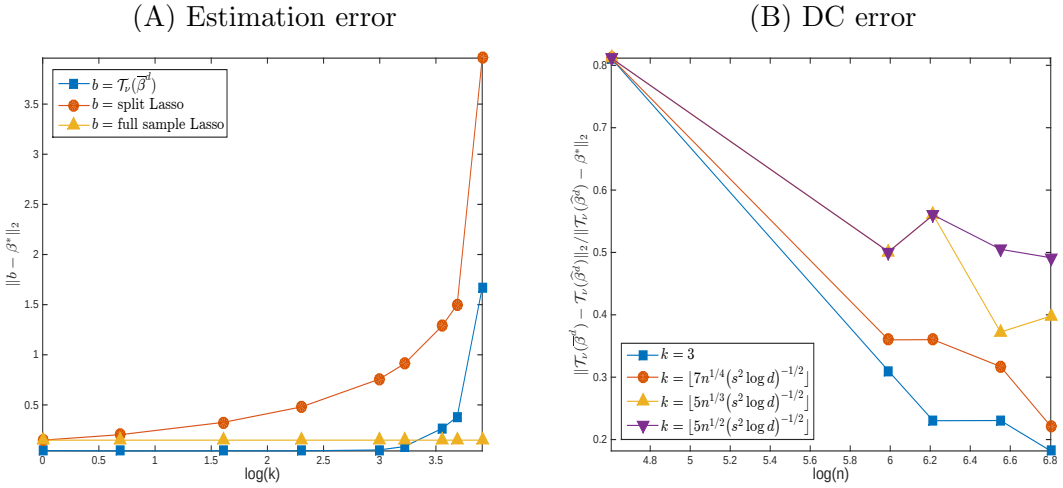

Figure 5: (A): Statistical error of the DC estimator, split LASSO and the full sample LASSO for $k \in \{1, 2, 5, 10, 20, 25, 35, 40, 50\}$ when $n = 1400$, $d = 1500$. (B): Euclidean norm difference between the DC thresholded debiased estimator and its full sample analogue.

The divide and conquer strategy is also successfully applied to estimation of regression parameters. We obtain the rate of the loss incurred by the divide and conquer strategy. Based on this result, we derive an upper bound on the number of subsamples for preserving the statistical error. For low-dimensional models, simple averaging is shown to be effective in preserving the statistical error, so long as $k = O(n/d)$ for the linear model and $k = O(\sqrt{n}/d)$ for the generalized linear model. For high-dimensional models, the debiased estimator used in the Wald construction is also successfully employed, achieving the same statistical error as the LASSO based on the full sample, so long as $k = O(\sqrt{n/s^2 \log d})$.

Our contribution advances the understanding of distributed inference and estimation in the presence of large scale and distributed data, but there is still a great deal of work to be done in the area. We focus here on the fundamentals of statistical inference and estimation in the divide and conquer setting. Beyond this, there is a whole toolkit of statistical methodology designed for the single sample setting, whose split sample asymptotic properties are yet to be understood.

7 Proofs

In this section, we present the proofs of the main theorems appearing in Sections 3.1-4. The statements and proofs of several auxiliary lemmas appear in the Supplementary Material. To simplify notation, we take $\beta_v^H = 0$ without loss of generality.

7.1 Proofs for Section 3.1

The proof of Theorem 3.3, relies on the following lemma, which bounds the probability that optimization problems in (3.4) are feasible.

Lemma 7.1. Assume $\Sigma = \mathbb{E}(\boldsymbol{X}_i \boldsymbol{X}_i^T)$ satisfies $C_{\min} < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_{\max}$ as well as $\|\Sigma^{-1/2}\boldsymbol{X}_1\|_{\psi_2} = \kappa$, then we have

$$\mathbb{P}\left(\max_{j=1,\ldots,k} \|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max} \leq a\sqrt{\frac{\log d}{n}}\right) \geq 1 - 2kd^{-c_2}, \text{ where } c_2 = \frac{a^2 C_{\min}}{24e^2\kappa^4 C_{\max}} - 2.$$

Proof. The proof is an application of the union bound in Lemma 6.2 of Javanmard and Montanari (2014). □

Using Lemma 7.1 we now prove Theorem 7.2, from which Theorem 3.3 easily follows. The term $\boldsymbol{Z}$ in the decomposition of $\sqrt{n}(\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*)$ in Theorem 7.2 is responsible for the asymptotic normality of the proposed DC Wald statistic in Theorem 3.3, while the upper bound on k ensures $\boldsymbol{\Delta}$ is asymptotically negligible.

Theorem 7.2. Suppose Conditions 3.1 and 3.2 are fulfilled. Let $\lambda \asymp \sqrt{k\log d/n}$ and $\vartheta_1 \asymp \sqrt{k\log d/n}$. With $k = o((s\log d)^{-1}\sqrt{n})$, $\sqrt{n}(\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*) = \boldsymbol{Z} + \boldsymbol{\Delta}$, where $\boldsymbol{Z} = \frac{1}{\sqrt{k}}\sum_{j=1}^k \frac{1}{\sqrt{n_k}} M^{(j)} X^{(j)T}\boldsymbol{\varepsilon}^{(j)}$ and $\|\boldsymbol{\Delta}\|_\infty = o_{\mathbb{P}}(1)$.

Proof. For notational convenience, we write $\widehat{\boldsymbol{\beta}}^\lambda_{\text{LASSO}}(\mathcal{D}_j)$ simply as $\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)$. Decompose $\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*$ as

$$\begin{aligned}
\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^* &= \frac{1}{k}\sum_{j=1}^k \Big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^* + \frac{1}{n_k} M^{(j)} X^{(j)T} X^{(j)}\big(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)\big)\Big) + \frac{1}{k}\sum_{j=1}^k \frac{1}{n_k} M^{(j)} X^{(j)T}\varepsilon^{(j)} \\
&= \frac{1}{k}\sum_{j=1}^k \big(I - M^{(j)}\widehat{\Sigma}^{(j)}\big)\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big) + \frac{1}{k}\sum_{j=1}^k \frac{1}{n_k} M^{(j)} X^{(j)T}\varepsilon^{(j)},
\end{aligned}$$

hence $\sqrt{n}(\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*) = \boldsymbol{Z} + \boldsymbol{\Delta}$, where

$$\boldsymbol{Z} = \frac{1}{\sqrt{k}}\sum_{j=1}^k \frac{1}{\sqrt{n_k}} M^{(j)} X^{(j)T}\varepsilon^{(j)} \quad \text{and} \quad \boldsymbol{\Delta} = \sqrt{n}\frac{1}{k}\sum_{j=1}^k \big(I - M^{(j)}\widehat{\Sigma}^{(j)}\big)\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big).$$

Defining $\boldsymbol{\Delta}^{(j)} = \big(I - M^{(j)}\widehat{\Sigma}^{(j)}\big)\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)$, we have

$$\|\boldsymbol{\Delta}^{(j)}\|_\infty \leq \|\boldsymbol{\Delta}^{(j)}\|_1 \leq \|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1$$

by Hölder's inequality, where $\|I - M^{(j)}\widehat{\Sigma}^{(j)}\|_{\max} \leq \vartheta_1$ by the definition of $M^{(j)}$ and, for $\lambda = C\sigma^2\sqrt{\log d/n_k}$,

$$\mathbb{P}\left(\big\|\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big\|_1^2 > C\frac{s^2\log(2d)}{n_k} + t\right) \leq \exp\left(-\frac{cn_k t}{s^2\sigma^2}\right) \tag{7.1}$$

by Bühlmann and van de Geer (2011). We thus bound the expectation of the ℓ_1 loss by

$$\mathbb{E}\left[\big\|\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big\|_1^2\right] \leq \frac{2Cs^2\log(2d)}{n_k} + \int_0^\infty \exp\left(-\frac{cn_k t}{s^2\sigma^2}\right)dt \leq \frac{2Cs^2\log(2d)}{n_k} + \frac{s^2\sigma^2}{cn_k}. \tag{7.2}$$

Define the event $\mathcal{E}^{(j)} := \{\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1 \leq s\sqrt{C\log(2d)/n_k}\}$ for $j = 1, \ldots, k$. $\|\boldsymbol{\Delta}^{(j)}\|_\infty \leq \Delta_1^{(j)} + \Delta_2^{(j)} + \Delta_3^{(j)}$ where

$$\begin{aligned}
\Delta_1^{(j)} &= \|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1 \mathbb{1}\{\mathcal{E}^{(j)}\} \\
&\quad - \mathbb{E}\big[\|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1 \mathbb{1}\{\mathcal{E}^{(j)}\}\big] \\
\Delta_2^{(j)} &= \|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1 \mathbb{1}\{\mathcal{E}^{(j)c}\} \\
&\quad - \mathbb{E}[\|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1 \mathbb{1}\{\mathcal{E}^{(j)c}\}] \quad \text{and} \\
\Delta_3^{(j)} &= \mathbb{E}[\|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1].
\end{aligned}$$

Consider $\Delta_1^{(j)}$, $\Delta_2^{(j)}$ and $\Delta_3^{(j)}$ in turn. By Hoeffding's inequality, we have for any $t > 0$,

$$\mathbb{P}\left(\frac{1}{k}\sum_{j=1}^{k}\Delta_1^{(j)} > t\right) \leq \exp\left(-\frac{n_k k t^2}{Cs^2\vartheta_1^2\log(2d)}\right) \leq \exp\left(-\frac{n_k n t^2}{Cs^2\log^2(2d)}\right). \tag{7.3}$$

By Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{k}\sum_{j=1}^{k}\Delta_2^{(j)} > t\right) &\leq \frac{\sum_{j=1}^{k}\mathbb{E}[\Delta_2^{(j)}]}{kt} \leq 2t^{-1}\mathbb{E}\big[\|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1 \mathbb{1}\{\mathcal{E}^{(j)c}\}\big] \\
&\leq 2t^{-1}\vartheta_1\sqrt{\mathbb{E}\big[\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1^2\big]\mathbb{P}(\mathcal{E}^{(j)c})} \\
&\leq Ct^{-1}\sqrt{\frac{\log d}{n_k}\cdot\frac{s^2\log(2d)}{n_k}d^{-c}} \leq Ct^{-1}sn_k^{-1}d^{-c/2}\log d,
\end{aligned} \tag{7.4}$$

where the penultimate inequality follows from Jensen's inequality. Finally, by Jensen's inequality again,

$$\begin{aligned}
\frac{1}{k}\sum_{j=1}^{k}\Delta_3^{(j)} &= \mathbb{E}[\|M^{(j)}\widehat{\Sigma}^{(j)} - I\|_{\max}\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1] \\
&\leq \vartheta_1\sqrt{\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1^2\right]} \leq C\frac{s\log d}{n_k}.
\end{aligned} \tag{7.5}$$

Combining (7.3), (7.4) and (7.5),

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{\Delta}\|_\infty > 3C\sqrt{n}\cdot\frac{s\log d}{n_k}\right) &\leq \sum_{u=1}^{3}\mathbb{P}\left(\frac{1}{k}\sum_{j=1}^{k}\Delta_u^{(j)} > C\sqrt{n}\cdot\frac{s\log d}{n_k}\right) \\
&\leq \exp(-ckn) + d^{-c/2} \to 0,
\end{aligned} \tag{7.6}$$

and taking $k = o\big((s\log d)^{-1}\sqrt{n}\big)$ delivers $\|\boldsymbol{\Delta}\|_\infty = o_{\mathbb{P}}(1)$. □

Proof of Theorem 3.3. We verify the requirements of the Lindeberg-Feller central limit theorem (e.g. Kallenberg, 1997, Theorem 4.12). Write

$$\overline{V}_n := \sqrt{n}\frac{1}{k}\sum_{j=1}^{k}\frac{Z_v^{(j)}}{\widehat{Q}^{(j)}} = \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\xi_{iv}^{(j)}, \quad \text{where} \quad \xi_{iv}^{(j)} := \frac{\mathbf{m}_v^{(j)T}\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}}{\big(n\mathbf{m}_v^{(j)T}\widehat{\Sigma}^{(j)}\mathbf{m}_v^{(j)}\big)^{1/2}}.$$

By the fact that ε_i is independent of X for all i and $\mathbb{E}[\varepsilon_i] = 0$,

$$\begin{aligned}\mathbb{E}(\xi_{iv}^{(j)}|X) &= \mathbb{E}\left[\mathbf{m}_v^{(j)T}\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}/\big(n\mathbf{m}_v^{(j)T}\widehat{\Sigma}^{(j)}\mathbf{m}_v^{(j)}\big)^{1/2}|X\right] \\ &= \big(n\mathbf{m}_v^{(j)T}\widehat{\Sigma}^{(j)}\mathbf{m}_v^{(j)}\big)^{-1/2}\mathbf{m}_v^{(j)T}\boldsymbol{X}_i^{(j)}\mathbb{E}\big(\varepsilon_i^{(j)}\big) = 0.\end{aligned}$$

By independence of $\{\varepsilon_i\}_{i=1}^n$ and the definition of $\widehat{\Sigma}^{(j)}$, we also have

$$\begin{aligned}&\mathrm{Var}\left(\overline{V}_n\Big|X\right) = \sum_{j=1}^k\sum_{i\in\mathcal{I}_j}\mathrm{Var}\big(\xi_{iv}^{(j)}|X\big) \\ = \;& k^{-1}\sum_{j=1}^k n_k^{-1}\big(\mathbf{m}_v^{(j)T}\widehat{\Sigma}^{(j)}\mathbf{m}_v^{(j)}\big)^{-1}\sum_{i\in\mathcal{I}_j}\mathbf{m}_v^{(j)T}\boldsymbol{X}_i^{(j)}\boldsymbol{X}_i^{(j)T}\mathbf{m}_v^{(j)}\,\mathrm{Var}\big(\varepsilon_i^{(j)}|X\big) = \sigma^2.\end{aligned}$$

It only remains to verify the Lindeberg condition, i.e.,

$$\lim_{k\to\infty}\lim_{n_k\to\infty}\frac{1}{\sigma^2}\sum_{j=1}^k\sum_{i\in\mathcal{I}_j}\mathbb{E}\left[(\xi_{iv}^{(j)})^2\,\mathbb{1}\big\{\big|\xi_{iv}^{(j)}\big| > \varepsilon\sigma\big\}\big|X\right] = 0, \quad \forall\,\varepsilon > 0. \tag{7.7}$$

By Lemma A.1, $\big|\xi_{iv}^{(j)}\big| \le n^{-1/2}c_{n_k}^{-1}|\mathbf{m}_v^{(j)T}\boldsymbol{X}_i^{(j)}||\varepsilon_i^{(j)}| \le n^{-1/2}c_{n_k}^{-1}\vartheta_2|\varepsilon_i^{(j)}|$, where $\liminf_{n_k} c_{n_k} = c_\infty > 0$, hence the event $\big\{|\xi_{iv}^{(j)}| > \varepsilon\sigma\big\}$ is contained in the event $\big\{|\varepsilon_i^{(j)}| > \varepsilon\sigma c_{n_k}\vartheta_2^{-1}\sqrt{n}\big\}$ and we have

$$\begin{aligned}&\frac{1}{\sigma^2}\sum_{j=1}^k\sum_{i\in\mathcal{I}_j}\mathbb{E}\left[(\xi_{iv}^{(j)})^2\,\mathbb{1}\big\{|\xi_{iv}^{(j)}| > \varepsilon\sigma\big\}|X\right] \le \frac{1}{\sigma^2}\sum_{j=1}^k\sum_{i\in\mathcal{I}_j}\mathbb{E}\left[(\xi_{iv}^{(j)})^2\,\mathbb{1}\big\{|\varepsilon_i^{(j)}| > \varepsilon\sigma c_{n_k}\vartheta_2^{-1}\sqrt{n}\big\}|X\right] \\ = \;& \frac{1}{\sigma^2}\frac{1}{k}\sum_{j=1}^k\big(\mathbf{m}_v^{(j)T}\widehat{\Sigma}\mathbf{m}_v^{(j)}\big)^{-1}\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\mathbf{m}_v^{(j)T}\boldsymbol{X}_i^{(j)}\boldsymbol{X}_i^{(j)T}\mathbf{m}_v^{(j)}\mathbb{E}\Big[(\varepsilon_i^{(j)})^2\,\mathbb{1}\big\{|\varepsilon_i^{(j)}| > \varepsilon\sigma c_{n_k}\vartheta_2^{-1}\sqrt{n}\big\}\Big] \\ = \;& \frac{1}{\sigma^2}\mathbb{E}\Big[(\varepsilon_i^{(j)})^2\,\mathbb{1}\big\{|\varepsilon_i^{(j)}| > \varepsilon\sigma c_{n_k}\vartheta_2^{-1}\sqrt{n_k}\sqrt{k}\big\}\Big].\end{aligned}$$

Let $\delta = \varepsilon\sigma c_{n_k}\vartheta_2^{-1}\sqrt{n}$. Then, for any $\eta > 0$,

$$\mathbb{E}\big[\big(\varepsilon_i^{(j)}\big)^2\,\mathbb{1}\big\{\big|\varepsilon_i^{(j)}\big| > \delta\big\}\big] \le \mathbb{E}\Big[\big(\varepsilon_i^{(j)}\big)^2\frac{\big|\varepsilon_i^{(j)}\big|^\eta}{\delta^\eta}\,\mathbb{1}\big\{\big|\varepsilon_i^{(j)}\big| > \delta\big\}\Big] \le \delta^{-\eta}\mathbb{E}\big[\big|\varepsilon_i^{(j)}\big|^{2+\eta}\big]. \tag{7.8}$$

Since $\vartheta_2 n^{-1/2} = o(1)$ by the statement of the theorem, the choice $\eta = 2$ delivers

$$\begin{aligned}&\frac{1}{\sigma^2}\lim_{k\to\infty}\lim_{n_k\to\infty}\sum_{j=1}^k\sum_{i\in\mathcal{I}_j}\mathbb{E}\left[(\xi_{iv}^{(j)})^2\,\mathbb{1}\big\{\big|\xi_{iv}^{(j)}\big| > \varepsilon\sigma\big\}\big|X\right] \\ \le\;& \lim_{k\to\infty}\lim_{n_k\to\infty}k^{-1}n_k^{-1}\vartheta_2 c_{n_k}^{-2}\varepsilon^{-2}\sigma^{-2}\mathbb{E}\left(\big(\varepsilon_i^{(j)}\big)^4\right) = 0\end{aligned} \tag{7.9}$$

by the bounded forth moment assumption. By the law of iterated expectations, all conditional results hold in unconditional form as well. Hence, $\overline{V}_n \rightsquigarrow N(0,\sigma^2)$ by the Lindeberg-Feller central limit theorem. □

Proof of Corollary 3.5. Similar to (7.9), we also have

$$\frac{1}{\sigma^3}\lim_{k\to\infty}\lim_{n_k\to\infty}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\mathbb{E}\left[(\xi_{iv}^{(j)})^4\,\mathbb{1}\{|\xi_{iv}^{(j)}|>\varepsilon\sigma\}\big|X\right]=0.$$

The proof is complete through an application of the self-normalized Berry-Essen inequality (de la Peña et al., 2009), noting that $\overline{S}_n=\overline{V}_n+o_P(1)$, as demonstrated in the previous proof. □

Proof of Lemma 3.4. We first show that, for any $j\in\{1,\dots,k\}$, $|\widehat{\sigma}^2(\mathcal{D}_j)-\sigma^2|=o_{\mathbb{P}}(k^{-1})$. To this end, letting

$$\widehat{\varepsilon}_i=Y_i^{(j)}-\boldsymbol{X}_i^{(j)T}\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)=Y_i^{(j)}-\boldsymbol{X}_i^{(j)T}\boldsymbol{\beta}^*-\boldsymbol{X}_i^{(j)T}\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^*\big),$$

we write

$$|\widehat{\sigma}^2(\mathcal{D}_j)-\sigma^2|=\Big|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\widehat{\varepsilon}_i^2-\sigma^2\Big|\le\Delta_1^{(j)}+2\Delta_2^{(j)}+\Delta_3^{(j)},$$

$$\begin{aligned}
\Delta_1^{(j)} &:= \Big|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\varepsilon_i^2-\sigma^2\Big|,\quad \Delta_2^{(j)}:=\Big|\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^*\big)\Big(\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}\Big)\Big|,\ \text{and}\\
\Delta_3^{(j)} &:= \Big|\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^*\big)^T\Big(\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\boldsymbol{X}_i^{(j)}\boldsymbol{X}_i^{(j)T}\Big)\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^*\big)\Big|\\
&= \big\|X^{(j)}\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^*\big)\big\|_2^2/n_k=O_{\mathbb{P}}(\lambda^2 s)
\end{aligned}$$

by Theorem 6.1 of Bühlmann and van de Geer (2011). Hence, with $\lambda=C\sigma^2\sqrt{k\log d/n}$, $\Delta_3^{(j)}=o_{\mathbb{P}}(1)$ for $k=o\big((s\log d)^{-1}n\big)$, a fortiori for $k=o\big((s\log d)^{-1}\sqrt{n}\big)$. Letting

$$\begin{aligned}
\Delta_{21}^{(j)} &= \big\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^*\big\|_1\Big\|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}-\mathbb{E}[\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}]\Big\|_\infty,\\
\Delta_{22}^{(j)} &= \big\|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^*\big\|_1\big\|\mathbb{E}[\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}]\big\|_\infty.
\end{aligned}$$

We obtain the bound

$$\Delta_2^{(j)}=\Big|\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^*\big)\Big(\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}-\mathbb{E}[\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}]\Big)+\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^*\big)\mathbb{E}[\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}]\Big|\le\Delta_{21}^{(j)}+\Delta_{22}^{(j)}.$$

By the statement of the Lemma, $\mathbb{E}\big[\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}\big]=\mathbb{E}\big[\boldsymbol{X}_i^{(j)}\mathbb{E}[\varepsilon_i^{(j)}|\boldsymbol{X}_i^{(j)}]\big]=0$, hence $\Delta_{22}^{(j)}=0$, while by the central limit theorem and Theorem 6.1 of Bühlmann and van de Geer (2011),

$$\Delta_{21}^{(j)}\le O_{\mathbb{P}}(\lambda s)O_{\mathbb{P}}(n_k^{-1/2}).$$

We conclude $\Delta_2^{(j)}=O_{\mathbb{P}}\big(\lambda s n_k^{-1/2}\big)$, and with $\lambda\asymp\sigma^2\sqrt{k\log d/n}$, $\Delta_2^{(j)}=o(1)$ with $k=o\big(n(s\log d)^{-2/3}\big)$, a fortiori for $k=o\big(\sqrt{n}(s\log d)^{-1}\big)$. Finally, noting that $\sigma^2=\mathbb{E}[\varepsilon_i^{(j)}]$, $\Delta_1^{(j)}=O_{\mathbb{P}}(n_k^{-1/2})=o_{\mathbb{P}}\big(1\big)$ by the central limit theorem. Combining the bounds, we obtain $|\widehat{\sigma}^2(\mathcal{D}_j)-\sigma^2|=o_{\mathbb{P}}(1)$ for any $j\in\{1,\dots,k\}$ and therefore $|\overline{\sigma}^2-\sigma^2|\le k^{-1}\sum_{j=1}^{k}|\widehat{\sigma}^2(\mathcal{D}_j)-\sigma^2|=o_{\mathbb{P}}(1)$. □

The proofs of Theorem 3.8 and Corollary 3.9 are stated as an application of Lemmas A.7 and A.8, which apply under a more general set of requirements.

Proof of Theorem 3.8. We verify (A1)-(A4) of Lemma A.7. For (A1), decompose the object of interest as

$$\frac{1}{n_k}\|X^{(j)}\widehat{\Theta}^{(j)}\|_{\max} = \frac{1}{n_k}\|X^{(j)}\big(\widehat{\Theta}^{(j)} - \Theta^*\big)\|_{\max} + \frac{1}{n_k}\|X^{(j)}\Theta^*\|_{\max} = \Delta_1 + \Delta_2,$$

where Δ_1 can be further decomposed and bounded by

$$\begin{aligned}\frac{1}{n_k}\big\|X^{(j)}\big(\widehat{\Theta}^{(j)} - \Theta^*\big)\big\|_{\max} &= \frac{1}{n_k}\max_{1\le i\le n}\max_{1\le v\le d}\Big[\big|\boldsymbol{X}_i^{(j)T}\big(\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*\big)\big|\Big] \\ &\le \frac{1}{n_k}\max_{1\le i\le n}\|\boldsymbol{X}_i\|_\infty \max_{1\le v\le d}\|\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*\|_1.\end{aligned}$$

We have

$$\mathbb{P}(\Delta_1 > q/2) \le \mathbb{P}\Big(\max_{1\le v\le d}\|\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*\|_1 > \frac{n}{kM}\frac{q}{2}\Big) < \psi$$

and by Condition 3.7, $\psi = o(d^{-1}) = o(k^{-1})$ for any $q \ge 2CMs_1(k/n)^{3/2}\sqrt{\log d}$, a fortiori for q a constant. Since $\boldsymbol{X}_i$ is sub-Gaussian, a matching probability bound can easily be obtained for Δ_2, thus we obtain $\mathbb{P}\big(n_k^{-1}\big\|X^{(j)}\widehat{\Theta}^{(j)}\big\|_{\max}\big) \le 2\psi$ for $\psi = o(k^{-1})$. (A2) and (A3) of Lemma A.7 are applications of Lemmas A.3 and A.4 respectively. To establish (A4), observe that $\big(\widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla^2\ell_{n_k}^{(j)}\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)\big) - \boldsymbol{e}_v\big) = \Delta_1 + \Delta_2 + \Delta_3$, where $\Delta_1 = \big(\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*\big)^T\nabla^2\ell_{n_k}^{(j)}\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)\big)$, $\Delta_2 = \boldsymbol{\Theta}_v^{*T}\big(\nabla^2\ell_{n_k}^{(j)}\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)\big) - \nabla^2\ell_{n_k}^{(j)}(\boldsymbol{\beta}^*)\big)$ and $\Delta_3 = \boldsymbol{\Theta}_v^{*T}\nabla^2\ell_{n_k}^{(j)}(\boldsymbol{\beta}^*) - \boldsymbol{e}_v$. We thus consider $\big|\Delta_\ell(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)\big|$ for $\ell = 1, 2, 3$.

$$\begin{aligned}\big|\Delta_2(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)\big| &= \Big|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\boldsymbol{\Theta}_v^{*T}\boldsymbol{X}_i\boldsymbol{X}_i^T\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\big[b''(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)) - b''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\big]\Big| \\ &\le U_3\max_{1\le i\le n}\big|\boldsymbol{\Theta}_v^{*T}\boldsymbol{X}_i\big|\frac{1}{n_k}\big\|X\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\big\|_2^2\end{aligned}$$

$\mathbb{P}\Big(\big\|X\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\big\|_2^2 \gtrsim n^{-1}sk\log(d/\delta)\Big) < \delta$ by Lemma A.4, thus $\mathbb{P}\Big(\big|\Delta_2(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)\big| > t\Big) < \delta$ for $t \asymp MU_3n^{-1}sk\log(d/\delta)$. Invoking Hölder's inequality, Hoeffding's inequality and Condition 2.1, we also obtain, for $t \asymp n^{-1}sk\log(d/\delta)$,

$$\mathbb{P}\Big(\big|\Delta_3(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)\big| > t\Big) \le \mathbb{P}\Big(\Big\|\boldsymbol{\Theta}_v^{*T}\Big(\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}b''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\boldsymbol{X}_i\boldsymbol{X}_i^T\Big) - \boldsymbol{e}_v\Big\|_{\max}\big\|\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big\|_1 > t\Big).$$

Therefore $\mathbb{P}\Big(\big|\Delta_2(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)\big| > t\Big) < 2\delta$. Finally, with $t \asymp n^{-1}(s \vee s_1)k\log(d/\delta)$,

$$\mathbb{P}\Big(\big|\Delta_1(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)\big| > t\Big) \le \mathbb{P}\Big(\Big\|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}X_i^T\big(\widehat{\boldsymbol{\Theta}}_v - \boldsymbol{\Theta}_v\big)b''(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j))\Big\|_2\Big\|\frac{1}{n_k}X^{(j)}\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\Big\|_2 > t\Big),$$

hence $\mathbb{P}\Big(\big|\Delta_1(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)\big| > t\Big) < 2\delta$. This follows because $\mathbb{P}\Big(\Big\|\frac{1}{n_k}X^{(j)}\big(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\Big\|_2 \gtrsim n^{-1/2}\sqrt{sk\log(d/\delta)}\Big) < \delta$ by Lemma A.4 and

$$\mathbb{P}\Big(\Big\|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}X_i^T\big(\widehat{\boldsymbol{\Theta}}_v - \boldsymbol{\Theta}_v\big)b''(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j))\Big\|_2 \gtrsim n^{-1/2}\sqrt{s_1k\log(d/\delta)}\Big) < \delta$$

by Lemma C.4 of Ning and Liu (2014). □

Proof of Corollary 3.9. We verify (A5)-(A9) of Lemma A.8. (A5) is satisfied because $\widetilde{\Theta}_{vv}$ is consistent under the required scaling by the statement of the corollary. (A6) is satisfied by Condition 3.7. To verify (A7), first note that $\nabla \ell_i(\boldsymbol{\beta}^*) = (b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*) - Y_i)\boldsymbol{X}_i$. According to Lemma A.2, we know that conditional on X, $b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*) - Y_i$ is a sub-gaussian random variable. Therefore Lemma B.5 delivers

$$\mathbb{P}\left(\|\frac{1}{n}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j} \nabla \ell_i(\boldsymbol{\beta}^*)\|_\infty > t \,|\, X \right) \leq d\exp\left(1 - \frac{ct^2}{nM^2}\right),$$

which implies that with probability $1 - c/d$,

$$\|\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j} \nabla \ell_i(\boldsymbol{\beta}^*)\|_\infty = C\sqrt{n\log d} \tag{7.10}$$

It only remains to verify (A8). Let $\xi_{iv}^{(j)} = \boldsymbol{\Theta}_v^{*T}\nabla \ell_i^{(j)}(\boldsymbol{\beta}^*)/\sqrt{n\Theta_{vv}^*}$. By the definition of the log likelihood,

$$\mathbb{E}[\xi_{iv}^{(j)}] = \frac{\boldsymbol{\Theta}_v^{*T}\mathbb{E}[\nabla \ell_i^{(j)}(\boldsymbol{\beta}^*)]}{(n\Theta_{vv}^*)^{1/2}} = 0$$

and by independence of $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$,

$$\begin{aligned}
&\mathrm{Var}\Big(\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\xi_{iv}^{(j)}\Big) = \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\mathrm{Var}\big(\xi_{iv}^{(j)}\big) = \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\mathbb{E}[(\xi_{iv}^{(j)})^2] \\
= \;& \frac{1}{n}\sum_{i=1}^{n}(\Theta_{vv}^*)^{-1}\boldsymbol{\Theta}_v^{*T}\mathbb{E}\big[\big(\nabla \ell_i(\boldsymbol{\beta}^*)\big)\big(\nabla \ell_i(\boldsymbol{\beta}^*)\big)^T\big]\boldsymbol{\Theta}_v^* = \frac{1}{n}\sum_{i=1}^{n}(\Theta_{vv}^*)^{-1}[\Theta^* J^* \Theta^*]_{vv} = 1.
\end{aligned}$$

By Condition 3.6, $\theta_{\min} > 0$, the event $\{|\xi_{iv}^{(j)}| > \varepsilon\}$ coincides with the event $\big\{\big|\boldsymbol{\Theta}_v^{*T}\nabla \ell_i(\boldsymbol{\beta}^*)\big| > \varepsilon\sqrt{\theta_{\min} n}\big\} = \big\{\big|\boldsymbol{\Theta}_v^{*T}\boldsymbol{X}_i(Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*))\big| > \varepsilon\sqrt{\theta_{\min} n}\big\}$. Furthermore, since $\big|\boldsymbol{\Theta}_v^{*T}\boldsymbol{X}_i\big| \leq M$ by Condition 3.7, this event is contained in the event $\big\{\big|Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\big| > \delta\big\}$, where $\delta = \varepsilon\sqrt{\theta_{\min} n}/M$. By an analogous calculation to that of equation (7.8), we have

$$\mathbb{E}\Big[\big(Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\big)^2 \mathbb{1}\{|Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)| > \delta\}|X\Big] \leq \delta^{-\eta}\mathbb{E}\big[\big(Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\big)^{2+\eta}|X\big].$$

Hence, setting $\eta = 2$ and noting that $\mathbb{E}\big[(Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*))^{2+\eta}|X\big] \leq C\sqrt{2+\eta}\phi U_2$ by Lemma A.2, it follows that

$$\begin{aligned}
&\lim_{k\to\infty}\lim_{n_k\to\infty}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\mathbb{E}\big[(\xi_{i,v}^{(j)})^2\,\mathbb{1}\{|\xi_{i,v}^{(j)}| > \varepsilon\}\big] \\
\leq\;& (\theta_{\min})^{-1}\lim_{k\to\infty}\lim_{n_k\to\infty} n^{-1}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\boldsymbol{\Theta}_v^{*T}\mathbb{E}[\boldsymbol{X}_i\boldsymbol{X}_i^T]\boldsymbol{\Theta}_v^*\delta^{-2} \\
\leq\;& (\theta_{\min})^{-1}\lim_{k\to\infty}\lim_{n_k\to\infty} M^3 s_1^2/(n\varepsilon^2\theta_{\min}) = 0,
\end{aligned} \tag{7.11}$$

where the last inequality follows because $\|\Sigma\|_{\max} = \|\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T]\|_{\max} < M^2$ by Condition 3.6. Similarly, we have for any $\varepsilon > 0$,

$$\varepsilon^{-3} \lim_{k\to\infty} \lim_{n_k\to\infty} \sum_{j=1}^{k} \sum_{i\in\mathcal{I}_j} \mathbb{E}\big[(\xi_{i,v}^{(j)})^3 \mathbb{1}\{|\xi_{i,v}^{(j)}| > \varepsilon\}\big] = 0.$$

Applying the self-normalized Berry-Essen inequality, we complete the proof of this corollary. □

7.2 Proofs for Section 3.2

The proof of Theorem 3.11 relies on several preliminary lemmas, collected in the Supplementary Material. Without loss of generality we set $H_0 : \beta_v^* = 0$ to ease notation.

Proof of Theorem 3.11. Since $\overline{S}(0) = k^{-1}\sum_{j=1}^{k} \widehat{S}^{(j)}(0, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j))$, and (B1)-(B4) of Condition A.9 are fulfilled under Conditions 3.6 and 2.1 by Lemma A.10 (see Appendix A). The proof is now simply an application of Lemma A.13 with $\beta_v^* = 0$ under the restriction of the null hypothesis. □

Proof of Lemma 3.14. The proof is an application of Lemma A.16, noting that (B1)-(B5) of Condition A.9 are fulfilled under Conditions 3.6 and 2.1 by Lemmas A.10 and A.11. □

7.3 Proofs for Section 4

Recall from Section 2 that for an arbitrary matrix M, $\boldsymbol{M}_\ell$ denotes the transposed ℓ^{th} row of M and $[\boldsymbol{M}]_\ell$ denotes the ℓ^{th} column of M.

Proof of Lemma 4.1. According to Theorem 7.2, we have $\sqrt{n}(\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*) = \boldsymbol{Z} + \boldsymbol{\Delta}$, where $\boldsymbol{Z} = \frac{1}{\sqrt{k}}\sum_{j=1}^{k} \frac{1}{\sqrt{n_k}} M^{(j)} X^{(j)T} \boldsymbol{\varepsilon}^{(j)}$. In (7.6), we prove that $\|\boldsymbol{\Delta}\|_\infty/\sqrt{n} \le Csk\log d/n$ with probability larger than $1 - \exp(-ckn) - d^{-c/2} \ge 1 - c_1/d$ for some constant c_1. Since $\widehat{\boldsymbol{\beta}}^d$ is a special case of $\overline{\boldsymbol{\beta}}^d$ when $k = 1$, we also have $\sqrt{n}(\widehat{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*) = \boldsymbol{Z} + \boldsymbol{\Delta}_1$, where (7.6) gives $\|\boldsymbol{\Delta}\|_\infty/\sqrt{n} \le Cs\log d/n$. Therefore, we have $\|\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d\|_\infty \le Csk\log d/n$ with high probability.

It only remains to bound the rate of $\|\boldsymbol{Z}\|_\infty/\sqrt{n}$. By Condition 3.2, conditioning on $\{\boldsymbol{X}_i\}_{i=1}^n$, we have for any $\ell = 1, \ldots, d$,

$$\mathbb{P}\Big(|Z_\ell|/\sqrt{n} > t \,\Big|\, \{\boldsymbol{X}_i\}_{i=1}^n\Big) = \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^{k} \boldsymbol{M}_\ell^{(j)T} X^{(j)T} \boldsymbol{\varepsilon}^{(j)}\Big| > t \,\Big|\, \{\boldsymbol{X}_i\}_{i=1}^n\Big) \le 2\exp\Big(-\frac{cnt^2}{\kappa^2 Q_\ell}\Big), \quad (7.12)$$

where κ is the variance proxy of ε defined in Condition 3.2 and

$$Q_\ell = \frac{1}{n}\sum_{j=1}^{k} \|X^{(j)} \boldsymbol{M}_\ell^{(j)T}\|_2^2.$$

Let $Q_{\max} = \max_{1\le\ell\le d} Q_\ell$. Applying the union bound to (7.12), we have

$$\begin{aligned}\mathbb{P}\Big(\|\boldsymbol{Z}\|_\infty/\sqrt{n} > t \,\Big|\, \{\boldsymbol{X}_i\}_{i=1}^n\Big) &\le \mathbb{P}\Big(\max_{1\le\ell\le d} |Z_\ell|/\sqrt{n} > t \,\Big|\, \{\boldsymbol{X}_i\}_{i=1}^n\Big) \\ &\le \sum_{\ell=1}^{d} \mathbb{P}\Big(|Z_\ell|/\sqrt{n} > t \,\Big|\, \{\boldsymbol{X}_i\}_{i=1}^n\Big) \le 2d\exp\Big(-\frac{cnt^2}{\kappa^2 Q_{\max}}\Big).\end{aligned}$$

Let $t = \sqrt{2\kappa^2 Q_{\max} \log d/(cn)}$, then with conditional probability $1 - 2/d$,

$$\|\boldsymbol{Z}\|_\infty/\sqrt{n} \le \sqrt{\kappa^2 Q_{\max} \log d/(cn)}. \tag{7.13}$$

The last step is to bound $Q_{\max}$. By the definition of Q_ℓ, we have

$$Q_\ell = \frac{1}{k}\sum_{j=1}^k \boldsymbol{M}_\ell^{(j)T} \widehat{\Sigma}^{(j)} \boldsymbol{M}_\ell^{(j)} \le \frac{1}{k}\sum_{j=1}^k [\boldsymbol{\Omega}]_\ell^T \widehat{\Sigma}^{(j)} [\boldsymbol{\Omega}]_\ell = \frac{1}{k}\sum_{j=1}^k \frac{1}{n_k}\sum_{i\in\mathcal{D}_j} (\boldsymbol{X}_i^T [\boldsymbol{\Omega}]_\ell)^2 = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i^T [\boldsymbol{\Omega}]_\ell)^2, \tag{7.14}$$

where $\Omega = \Sigma^{-1}$. The inequality is due to the fact that $M_\ell^{(j)}$ is the minimizer in (3.4). By condition (3.2) and the connection between subgaussian and subexponential distributions, the random variable $(\boldsymbol{X}_i^T \boldsymbol{\Omega}_\ell)^2$ satisfies

$$\sup_{q\ge 1} q^{-1} \big(\mathbb{E}|(\boldsymbol{X}_i^T \boldsymbol{\Omega}_\ell)^2|^q\big)^{1/q} \le 4\kappa^2 \Omega_{\ell\ell}.$$

Therefore, by Bernstein's inequality for subexponential random variables, we have

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i^T [\boldsymbol{\Omega}]_\ell)^2 - \mathbb{E}[\boldsymbol{X}_1^T [\boldsymbol{\Omega}]_\ell]^2\Big| > t\Big) \le 2\exp\Big(-c\Big(\frac{nt^2}{16\kappa^4\Omega_{\ell\ell}^2}\Big) \bigwedge \Big(\frac{nt}{4\kappa^2\Omega_{\ell\ell}}\Big)\Big).$$

Applying the union bound again, we have

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\le\ell\le d}\Big|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i^T [\boldsymbol{\Omega}]_\ell)^2 - \mathbb{E}[\boldsymbol{X}_1^T [\boldsymbol{\Omega}]_\ell]^2\Big| > 8\kappa^2\Omega_{\ell\ell}\sqrt{\frac{\log d}{cn}}\Big) \\
&\le \sum_{j=1}^d \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i^T [\boldsymbol{\Omega}]_\ell)^2 - \mathbb{E}[\boldsymbol{X}_1^T [\boldsymbol{\Omega}]_\ell]^2\Big| > 8\kappa^2\Omega_{\ell\ell}\sqrt{\frac{\log d}{cn}}\Big) \le 2/d.
\end{aligned}$$

Therefore, with probability $1 - 2/d$, there exist a constant C_1 such that

$$Q_{\max} = \max_{1\le\ell\le d} Q_\ell \le \max_{1\le\ell\le d}\Big|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i^T \boldsymbol{\Omega}_\ell)^2 - \mathbb{E}[\boldsymbol{X}_1^T \boldsymbol{\Omega}_\ell]^2\Big| + \mathbb{E}[\boldsymbol{X}_1^T \boldsymbol{\Omega}_\ell]^2 \le 8\kappa^2 \boldsymbol{\Omega}_{jj}\sqrt{\frac{\log d}{cn}} + \Omega_{jj} \le C_1,$$

where the last inequality is due to Condition 3.1. By (7.13), we have with probability $1 - 4/d$, $\|\boldsymbol{Z}\|_\infty/\sqrt{n} \le \sqrt{\kappa^2 C_1 \log d/(cn)}$. Combining this with the result on $\|\boldsymbol{\Delta}\|_\infty$ delivers the rate in the lemma. □

Proof of Theorem 4.3. By Lemma 4.1 and $k = O(\sqrt{n/(s^2 \log d)})$, there exists a sufficiently large C_0 such that for the event $\mathcal{E} := \{\|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty \le C_0\sqrt{\log d/n}\}$, we have $\mathbb{P}(\mathcal{E}) \ge 1 - c/d$. We choose $\nu = C_0\sqrt{\log d/n}$, which implies that, under $\mathcal{E}$, we have $\nu \ge \|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty$.

Let $\mathcal{S}$ be the support of $\boldsymbol{\beta}^*$. The derivations in the remainder of the proof hold on the event $\mathcal{E}$. Observe $\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d_{\mathcal{S}^c}) = \boldsymbol{0}$ as $\|\overline{\boldsymbol{\beta}}^d_{\mathcal{S}^c}\|_\infty \le \nu$. For $j \in \mathcal{S}$, if $|\boldsymbol{\beta}^*_j| \ge 2\nu$, we have $|\overline{\boldsymbol{\beta}}^d_j| \ge |\beta^*_j| - \nu \ge \nu$ and thus $|\mathcal{T}_\nu(\overline{\beta}^d_j) - \beta^*_j| = |\overline{\beta}^d_j - \beta^*_j| \le \nu$. While if $|\beta^*_j| < 2\nu$, $|\mathcal{T}_\nu(\overline{\beta}^d_j) - \beta^*_j| \le |\beta^*_j| \vee |\overline{\beta}^d_j - \beta^*_j| \le 2\nu$. Therefore, on the event $\mathcal{E}$,

$$\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2 = \|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d_{\mathcal{S}}) - \boldsymbol{\beta}^*_{\mathcal{S}}\|_2 \le 2\sqrt{s}\nu \text{ and } \|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_\infty = \|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d_{\mathcal{S}}) - \boldsymbol{\beta}^*_{\mathcal{S}}\|_\infty \le 2\nu.$$

The statement of the theorem follows because $\nu = C_0\sqrt{\log d/n}$ and $\mathbb{P}(\mathcal{E}) \geq 1 - c/d$. Following the same reasoning, on the event $\mathcal{E}' := \mathcal{E} \cup \{\|\widehat{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty \leq C_0\sqrt{\log d/n}\} \cup \{\|\widehat{\boldsymbol{\beta}}^d - \overline{\boldsymbol{\beta}}^d\|_\infty \leq C_0 sk\log d/n\}$, we have

$$\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)\|_2 = \|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d_{\mathcal{S}}) - \mathcal{T}_\nu(\widehat{\boldsymbol{\beta}}^d_{\mathcal{S}})\|_2 \leq \|\overline{\boldsymbol{\beta}}^d_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}^d_{\mathcal{S}}\|_2 \leq \sqrt{s}\|\overline{\boldsymbol{\beta}}^d_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}^d_{\mathcal{S}}\|_\infty \leq Cs^{3/2}k\log d/n.$$

As Lemma 4.1 also gives $\mathbb{P}(\mathcal{E}') \geq 1 - c/d$, the proof is complete. □

Proof of Lemma 4.6. The strategy of proving this lemma is similar to the proof of Lemma 4.1. In the proof of Lemma A.7 and Theorem 3.8, we have shown that

$$(\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*) = \underbrace{-\frac{1}{k}\sum_{j=1}^k \widehat{\Theta}^{(j)T}\nabla\ell^{(j)}_{n_k}(\boldsymbol{\beta}^*)}_{\mathbf{T}} + \frac{1}{k}\sum_{j=1}^k \boldsymbol{\Delta}_j,$$

where the remainder term for each j is

$$\boldsymbol{\Delta}_j = \left(I - \widehat{\Theta}^{(j)T}\frac{1}{n_k}\sum_{i\in\mathcal{I}_j} b''(\widetilde{\eta}_i)\boldsymbol{X}_i\boldsymbol{X}_i^T\right)(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)$$

and $\widetilde{\eta}_i = t\boldsymbol{X}_i^T\boldsymbol{\beta}^* + (1-t)\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)$ for some $t \in (0,1)$. We bound $\boldsymbol{\Delta}_j$ by decomposing it into three terms:

$$\begin{aligned}
\|\boldsymbol{\Delta}_j\|_\infty \leq\ & \underbrace{\left\|\left(I - \Theta^*\frac{1}{n_k}\sum_{i\in\mathcal{I}_j} b''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\boldsymbol{X}_i\boldsymbol{X}_i^T\right)(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)\right\|_\infty}_{I_1} \\
&+ \underbrace{\left\|\Theta^*\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}(b''(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)) - b''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*))\boldsymbol{X}_i\boldsymbol{X}_i^T\right)(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)\right\|_\infty}_{I_2} \\
&+ \underbrace{\left\|(\widehat{\Theta}^{(j)} - \Theta^*)^T\frac{1}{n_k}\sum_{i\in\mathcal{I}_j} b''(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j))\boldsymbol{X}_i\boldsymbol{X}_i^T\right)(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)\right\|_\infty}_{I_3}.
\end{aligned}$$

By Hoeffding's inequality and Condition 3.3, the first term is bounded by

$$|I_1| \leq \left\|I - \Theta^*\frac{1}{n_k}\sum_{i\in\mathcal{I}_j} b''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\boldsymbol{X}_i\boldsymbol{X}_i^T\right\|_{\max}\left\|\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*\right\|_1 \leq C\frac{sk\log d}{n}, \tag{7.15}$$

with probability $1 - c/d$. By Condition 3.6 (iii), Condition 3.7 (iv) and Lemma A.4, we have with probability $1 - c/d$,

$$|I_2| \leq \max_i \left\|\boldsymbol{\Theta}^*\boldsymbol{X}_i\right\|_\infty\frac{1}{n_k}\sum_{i\in\mathcal{I}_j} U_3[\boldsymbol{X}_i(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)]^2 \leq C\frac{sk\log d}{n}. \tag{7.16}$$

Finally, we bound I_3 by with probability $1 - c/d$,

$$\begin{aligned} |I_3| &\leq \left(U_2 \frac{1}{n_k} \sum_{i \in \mathcal{I}_j} b''(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j))[\boldsymbol{X}_i^T(\widehat{\Theta}^{(j)} - \Theta^*)]^2 \right)^{1/2} \left(\frac{1}{n_k} \sum_{i \in \mathcal{I}_j} [\boldsymbol{X}_i(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*)]^2 \right)^{1/2} \\ &\leq C \frac{(s_1 \vee s) k \log d}{n}, \end{aligned} \tag{7.17}$$

where the last inequality is due to Lemma A.4 and Lemma C.4 of Ning and Liu (2014).

Combining (7.15) - (7.17) and applying the union bound, we have

$$\Big\| \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{\Delta}_j \Big\|_\infty \leq \max_j \|\boldsymbol{\Delta}_j\|_\infty = O_P\Big(\frac{(s_1 \vee s) k \log d}{n} \Big).$$

Therefore, we only need to bound the infinity norm of the leading term $\mathbf{T}$. By Condition 3.7 and equation (7.10), we have with probability $1 - c/d$,

$$\max_{1 \leq j \leq k} \max_{1 \leq v \leq d} \|\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*\|_1 \leq C s_1 \sqrt{\log d / n} \text{ and } \Big\| \frac{1}{k} \sum_{j=1}^{k} \nabla \ell_{n_k}^{(j)}(\boldsymbol{\beta}^*) \Big\|_\infty \leq C \sqrt{\log d / n}. \tag{7.18}$$

This, together with Condition 3.6 and Condition 3.7 give the bound,

$$\|\mathbf{T}\|_\infty \leq \Big(M \max_{v,j} \|\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*\|_1 + \max_i \|\boldsymbol{X}_i^T \boldsymbol{\Theta}^*\|_\infty \Big) \Big\| \frac{1}{k} \sum_{j=1}^{k} \nabla \ell_{n_k}^{(j)}(\boldsymbol{\beta}^*) \Big\|_\infty \leq C \Big(\sqrt{\frac{\log d}{n}} + \frac{s_1 \log d}{n} \Big),$$

with probability $1 - c/d$. Since $\widehat{\boldsymbol{\beta}}^d$ is a special case of $\overline{\boldsymbol{\beta}}^d$ when $k = 1$, the proof of the lemma is complete. □

Proof of Corollary 4.9. By an analogous proof strategy to that of Theorem 4.7, $\big|[\mathcal{T}_\zeta(\overline{\Theta})]_{vv} - \Theta_{vv}^*\big| = O_p\big(\sqrt{n^{-1} \log d}\big) = o_{\mathbb{P}}(1)$ under the conditions of the Corollary provided $k = o\big(((s \vee s_1) \log d)^{-1} \sqrt{n}\big)$. □

Proof of Theorem 4.10.

$$\begin{aligned} \overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}} &= \frac{1}{k} \sum_{j=1}^{k} ((X^{(j)})^T X^{(j)})^{-1} (X^{(j)})^T \boldsymbol{Y}^{(j)} - (X^T X)^{-1} X^T \boldsymbol{Y} \\ &= \frac{1}{k} \sum_{j=1}^{k} \left(\left(X^{(j)T} X^{(j)} / n_k \right)^{-1} - (X^T X / n)^{-1} \right) X^{(j)T} \boldsymbol{\varepsilon}^{(j)} / n_k \\ &= \frac{1}{k} \sum_{j=1}^{k} \left(\left(X^{(j)T} X^{(j)} / n_k \right)^{-1} - \Sigma^{-1} \right) X^{(j)T} \boldsymbol{\varepsilon}^{(j)} / n_k + \left(\Sigma^{-1} - (X^T X / n)^{-1} \right) X^T \boldsymbol{\varepsilon} / n. \end{aligned} \tag{7.19}$$

For simplicity, denote $X^{(j)T} X^{(j)} / n_k$ by $S_X^{(j)}$, $X^T X / n$ by S_X, $(S_X^{(j)})^{-1} - (\Sigma)^{-1}$ by $D_1^{(j)}$ and $(\Sigma)^{-1} - {S_X}^{-1}$ by D_2. For any $\tau \in \mathbb{R}$, define an event $\mathcal{E}^{(j)} = \{\|(S_X^{(j)})^{-1}\|_2 \leq 2/C_{\min}\} \cap \{\|S_X^{(j)} - \Sigma\|_2 \leq$

$(\delta_1 \vee \delta_1^2)\}$ for all $j = 1, \ldots, k$, where $\delta_1 = C_1\sqrt{d/n_k} + \tau/\sqrt{n_k}$, and an event $\mathcal{E} = \{\|(S_X)^{-1}\|_2 \leq 2/C_{\min}\} \cap \{\|S_X - \Sigma\|_2 < (\delta_2 \vee \delta_2^2)\}$, where $\delta_2 = C_1\sqrt{d/n} + \tau/\sqrt{n}$. Note that by Lemma B.1 and B.4, the probability of both $(\mathcal{E}^{(j)})^c$ and $\mathcal{E}^c$ are very small. In particular

$$\mathbb{P}(\mathcal{E}^c) \leq \exp(-cn) + \exp(-c_1\tau^2) \text{ and } \mathbb{P}((\mathcal{E}^{(j)})^c) \leq \exp(-cn/k) + \exp(-c_1\tau^2).$$

Then, letting $\mathcal{E}_0 := \bigcap_{j=1}^{k} \mathcal{E}^{(j)}$, an application of the union bound and Lemma B.8 delivers

$$\begin{aligned}
\mathbb{P}\left(\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 > t\right) &\leq \mathbb{P}\left(\left\{\left\|\frac{1}{k}\sum_{j=1}^{k}(X^{(j)}D_1^{(j)})^T\boldsymbol{\varepsilon}^{(j)}/n_k\right\|_2 > t/2\right\} \cap \mathcal{E}_0\right) \\
&\quad + \mathbb{P}\left(\left\{\|(XD_2)^T\boldsymbol{\varepsilon}/n\|_2 > t/2\right\} \cap \mathcal{E}\right) + \mathbb{P}(\mathcal{E}_0^c) + \mathbb{P}(\mathcal{E}^c) \\
&\leq 2\exp\left(d\log(6) - \frac{t^2C_{\min}^3 n}{32C_3\sigma_1^2\delta_1^2}\right) + k\exp(-cn/k) + (k+1)\exp(-c_1\tau^2).
\end{aligned}$$

When $d \to \infty$ and $\log n = o(d)$, choose $\tau = \sqrt{d/c_1}$ and $\delta_1 = O(\sqrt{kd/n})$. Then there exists a constant C such that

$$\mathbb{P}\left(\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 > C\frac{\sqrt{k}d}{n}\right) \leq (k+3)\exp(-d) + k\exp(-\frac{cn}{k}).$$

Otherwise choose $\tau = \sqrt{\log n/c_1}$ and $\delta_1 = O(\sqrt{k\log n/n})$. Then there exists a constant C such that

$$\mathbb{P}\left(\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 > C\frac{\sqrt{k}\log n}{n}\right) \leq \frac{k+3}{n} + k\exp(-\frac{cn}{k}).$$

Overall, we have

$$\mathbb{P}\left(\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2 > C\frac{\sqrt{k}(d \vee \log n)}{n}\right) \leq ck\exp(-(d \vee \log n)) + k\exp(-cn/k),$$

which leads to the final conclusion. □

Proof of Corollary 4.12. Define an event $\mathcal{E} = \{\|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty \leq 2C\sqrt{\log d/n}\}$, then by the condition on the minimal signal strength and Lemma 4.1, for some constant C' we have

$$\begin{aligned}
\mathbb{P}\left(\|\overline{\boldsymbol{\beta}}^r - \widehat{\boldsymbol{\beta}}^o\|_2 > C'\frac{\sqrt{k}(s \vee \log n)}{n}\right) &\leq \mathbb{P}\left(\left\{\|\overline{\boldsymbol{\beta}}^r - \widehat{\boldsymbol{\beta}}^o\|_2 > C'\frac{\sqrt{k}(s \vee \log n)}{n}\right\} \cap \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c) \\
&\leq \mathbb{P}\left(\left\{\|\overline{\boldsymbol{\beta}}^o - \widehat{\boldsymbol{\beta}}^o\|_2 > C'\frac{\sqrt{k}(s \vee \log n)}{n}\right\} \cap \mathcal{E}\right) + c/d \\
&\leq ck\exp(-(s \vee \log n)) + k\exp(-cn/k) + c/d.
\end{aligned}$$

where $\overline{\boldsymbol{\beta}}^o = \frac{1}{k}\sum_{j=1}^{k}(X_S^{(j)T}X_S^{(j)})^{-1}X_S^{(j)T}\boldsymbol{Y}^{(j)}$, which is the average of the oracle estimators on the subsamples. Then the conclusion can be easily validated. □

Proof of Theorem 4.13. The following notation is used throughout the proof.

$$S(\boldsymbol{\beta}) := \nabla^2 \ell_n(\boldsymbol{\beta}) = \frac{1}{n} X^T D(X\boldsymbol{\beta}) X, \quad S^{(j)}(\boldsymbol{\beta}) := \nabla^2 \ell_{n_k}^{(j)}(\boldsymbol{\beta}) = \frac{1}{n_k} X^{(j)T} D(X^{(j)}\boldsymbol{\beta}) X^{(j)},$$
$$S_X := \frac{1}{n} X^T X, \quad S_X^{(j)} := \frac{1}{n_k} X^{(j)T} X^{(j)}$$

For any $j = 1, \ldots, k$, $\widehat{\boldsymbol{\beta}}^{(j)}$ satisfies

$$\nabla \ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)}) = \frac{1}{n_k} X^{(j)T}(\boldsymbol{Y}^{(j)} - \boldsymbol{\mu}(X^{(j)}\widehat{\boldsymbol{\beta}}^{(j)})) = 0.$$

Through a Taylor expansion of the left hand side at the point $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, we have

$$\frac{1}{n_k} X^{(j)T}(\boldsymbol{Y}^{(j)} - \boldsymbol{\mu}(X^{(j)}\boldsymbol{\beta}^*)) - S^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^*) - \mathbf{r}^{(j)} = 0,$$

where the remainder term $\mathbf{r}^{(j)}$ is a d dimensional vector with g^{th} component

$$\begin{aligned} r_g^{(j)} &= \frac{1}{6n_k}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^*)^T \nabla^2_{\boldsymbol{\beta}}[(\boldsymbol{X}_g^{(j)})^T \boldsymbol{\mu}(X^{(j)}\boldsymbol{\beta})](\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^*) \\ &= \frac{1}{6n_k}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^*)^T X^{(j)T} \mathrm{diag}\{\boldsymbol{X}_g^{(j)} \circ \boldsymbol{\mu}''((X^{(j)}\widetilde{\boldsymbol{\beta}}^{(j)}))\} X^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^*), \end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}^{(j)}$ is in a line segment between $\widehat{\boldsymbol{\beta}}^{(j)}$ and $\boldsymbol{\beta}^*$. It therefore follows that

$$\widehat{\boldsymbol{\beta}}^{(j)} = \boldsymbol{\beta}^* + (S^{(j)})^{-1}[X^{(j)T}(\boldsymbol{Y}^{(j)} - \boldsymbol{\mu}(X^{(j)}\boldsymbol{\beta}^*)) + n_k \mathbf{r}^{(j)}].$$

A similar equation holds for the global MLE $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + S^{-1}[X^T(\boldsymbol{Y} - \boldsymbol{\mu}(X\boldsymbol{\beta}^*)) + n\mathbf{r}],$$

where for $g = 1, \ldots, d$,

$$r_g = \frac{1}{6n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T X^T \mathrm{diag}\{\boldsymbol{X}_g \circ \boldsymbol{\mu}''((X\widetilde{\boldsymbol{\beta}}^{(j)}))\} X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*).$$

Therefore we have

$$\begin{aligned} \frac{1}{k}\sum_{j=1}^k \widehat{\boldsymbol{\beta}}^{(j)} - \widehat{\boldsymbol{\beta}} &= \frac{1}{k}\sum_{j=1}^k \left\{(S^{(j)})^{-1} - \Sigma^{-1}\right\} X^{(j)T}(\boldsymbol{Y}^{(j)} - \boldsymbol{\mu}(X^{(j)}\boldsymbol{\beta}^*)) \\ &\quad - \left\{S^{-1} - \Sigma^{-1}\right\} X^T(\boldsymbol{Y} - \boldsymbol{\mu}(X\boldsymbol{\beta}^*)) + \boldsymbol{R} = \boldsymbol{B} + \boldsymbol{R}, \end{aligned}$$

where $\boldsymbol{R} = (1/k)\sum_{j=1}^k (S^{(j)})^{-1}\mathbf{r}^{(j)} - S^{-1}\mathbf{r}$. We next derive stochastic bounds for $\|\boldsymbol{B}\|_2$ and $\|\boldsymbol{R}\|_2$ respectively, but to study the appropriate threshold, we introduce the following events with probability that approaches one under appropriate scaling. For $j = 1, \ldots, k$ and $\kappa, \tau, t > 0$,

$$\begin{aligned} \mathcal{E}^{(j)} &:= \{\|(S^{(j)})^{-1}\|_2 \le 2/C_{\min}\} \cap \left\{\|S^{(j)} - \Sigma\|_2 \le (\delta_1 \vee \delta_1^2)\right\} \cap \{\|S_X^{(j)}\|_2 \le 2C_{\max}\}, \\ \mathcal{E} &:= \{\|S^{-1}\|_2 \le 2/L_{\min}\} \cap \left\{\|S - \Sigma\|_2 \le (\delta_2 \vee \delta_2^2)\right\} \cap \{\|S_X\|_2 \le 2C_{\max}\}, \\ \mathcal{F}^{(j)} &:= \left\{\|\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^*\|_2 > t\right\}, \quad \mathcal{F} := \left\{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > t\right\}, \end{aligned}$$

where $\delta_1 = C_1\sqrt{d/n_k}+\tau/\sqrt{n_k}$ and $\delta_2 = C_1\sqrt{d/n_k}+\tau/\sqrt{n}$. Denote the intersection of all the above events by $\mathcal{A}$. Note that Condition 3.6 implies that $\sqrt{b''(\boldsymbol{X}_i^T\boldsymbol{\beta})}\boldsymbol{X}_i$ are i.i.d. sub-gaussian vectors, so by Lemmas B.1, B.4, B.3 and B.10, we have

$$\mathbb{P}(\mathcal{A}^c) \leq (2k+1)\exp\left(-\frac{cn}{k}\right) + (k+1)\exp(-c_1\tau^2) + 2k\exp\left(d\log 6 - \frac{nC_{\min}^2L_{\min}^2t^2}{2^{11}C_{\max}U_2\phi k}\right).$$

We first consider the bounded design, i.e., Condition 3.6 (ii). In order to bound $\|\boldsymbol{R}\|_2$, we first derive an upper bound for $r_g^{(j)}$. Under the event $\mathcal{A}$, by Lemma A.5 we have

$$\max_{1\leq g\leq d, 1\leq j\leq k} r_g^{(j)} \leq \frac{1}{3}MU_3C_{\max}t^2 \text{ and } \max_{1\leq g\leq d} r_g \leq \frac{1}{3}MU_3C_{\max}t^2.$$

It follows that, under $\mathcal{A}$,

$$\|\boldsymbol{R}\|_2 \leq \frac{2}{3}M\sqrt{d}U_3C_{\max}t^2. \tag{7.20}$$

Note that $\boldsymbol{B}$ is very similar to the RHS of Equation (7.19). Now we use essentially the same proof strategy as in the OLS part to bound $\|\boldsymbol{B}\|_2$. Following similar notations as in OLS, we denote $(S^{(j)})^{-1}-\Sigma^{-1}$ by $D_1^{(j)}$, $S^{-1}-\Sigma^{-1}$ by D_2, $\boldsymbol{Y}^{(j)}-\boldsymbol{\mu}(X^{(j)}\boldsymbol{\beta}^*)$ by $\boldsymbol{\varepsilon}^{(j)}$ and $\boldsymbol{Y}-\boldsymbol{\mu}(X\boldsymbol{\beta}^*)$ by $\boldsymbol{\varepsilon}$. For concision, we relegate the details of the proof to Lemma B.9, which delivers the following stochastic bound on $\|\boldsymbol{B}\|_2$.

$$\mathbb{P}(\{\|\boldsymbol{B}\|_2 > t_1\}\cap\mathcal{A}) \leq 2\exp\left(d\log(6) - \frac{C_{\min}^4L_{\min}^2nt_1^2}{128\phi U_2C_{\max}(\delta_1\vee\delta_1^2)^2}\right). \tag{7.21}$$

Combining Equation (7.21) with (7.20) leads us to the following inequality.

$$\begin{aligned}\mathbb{P}\left(\|\overline{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\|_2 > \frac{2}{3}M\sqrt{d}U_3C_{\max}t^2+t_1\right) &\leq (2k+1)\exp\left(-\frac{cn}{k}\right)+(k+1)\exp(-c_1\tau^2)\\ &+(k+1)\exp\left(d\log 6-\frac{C_{\min}^2L_{\min}^2nt^2}{2^{11}C_{\max}U_2\phi k}\right)+2\exp\left(d\log 6-\frac{C_{\min}^4L_{\min}^2nt_1^2}{128\phi U_2C_{\max}(\delta_1\vee\delta_1^2)^2}\right).\end{aligned}$$

Choose $t=t_1=\sqrt{d/n_k}$ and, when $d\gg\log n$, choose $\tau=\sqrt{d/c_1}$ and $\delta_1=O(\sqrt{kd/n})$. Then there exists a constant $C>0$ such that

$$\mathbb{P}\left(\|\overline{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\|_2 > C\frac{kd^{3/2}}{n}\right) \leq (2k+1)\exp(-\frac{cn}{k})+2(k+2)\exp(-d).$$

When it is not true that $d\gg\log n$, choose $\tau=\sqrt{\log n/c_1}$ and $\delta=O(\sqrt{k\log n/n})$. Then there exists a constant $C>0$ such that

$$\mathbb{P}\left(\|\overline{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\|_2 > C\frac{k\sqrt{d}\log n}{n}\right) \leq (2k+1)\exp(-\frac{cn}{k})+\frac{k+3}{n}.$$

Overall, we have

$$\mathbb{P}\left(\|\overline{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\|_2 > C\frac{k\sqrt{d}(d\vee\log n)}{n}\right) \leq ck\exp(-cn/k)+ck\exp(-c\max(d,\log n)),$$

which leads to the final conclusion. □

Proof of Corollary 4.15. Define an event $\mathcal{E} = \{\|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty \leq 2C\sqrt{\log d/n}\}$, then by the conditions of Corollary 4.15 and results of Lemma 4.6 and Theorem 4.13,

$$\begin{aligned}
\mathbb{P}\left(\|\overline{\boldsymbol{\beta}}^r - \widehat{\boldsymbol{\beta}}^o\|_2 > C'\frac{k\sqrt{s}(s \vee \log n)}{n}\right) &\leq \mathbb{P}\left(\left\{\|\overline{\boldsymbol{\beta}}^r - \widehat{\boldsymbol{\beta}}^o\|_2 > C'\frac{k\sqrt{s}(s \vee \log n)}{n}\right\} \cap \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c) \\
&\leq \mathbb{P}\left(\left\{\|\overline{\boldsymbol{\beta}}^o - \widehat{\boldsymbol{\beta}}^o\|_2 > C'\frac{k\sqrt{s}(s \vee \log n)}{n}\right\} \cap \mathcal{E}\right) + c/d \\
&\leq ck\exp(-(s \vee \log n)) + k\exp(-cn/k) + c/d.
\end{aligned}$$

where $\overline{\boldsymbol{\beta}}^o = \frac{1}{k}\sum_{j=1}^{k}\widehat{\boldsymbol{\beta}}^o(\mathcal{D}_j)$, $\widehat{\boldsymbol{\beta}}^o(\mathcal{D}_j) = \operatorname{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^d, \boldsymbol{\beta}_{S^c}=0}\ell^{(j)}(\boldsymbol{\beta})$ and C' is a constant. Then it is not hard to see that the final conclusion is true. □

Acknowledgements: The authors thank Weichen Wang, Jason Lee and Yuekai Sun for helpful comments.

References

Bickel, P. J. (1975). One-step huber estimates in the linear model. *Journal of the American Statistical Association* **70** 428–434.

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

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

Chen, X. and Xie, M. (2012). A split and conquer approach for analysis of extraordinarily large data. Tech. Rep. 2012-01, Department of Statistics, Rutgers University.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical statistics*. Chapman and Hall, London.

de la Peña, V. H., Lai, T. L. and Shao, Q.-M. (2009). *Self-normalized processes*. Probability and its Applications (New York), Springer-Verlag, Berlin. Limit theory and statistical applications.

Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 37–65.

Fan, J., Han, F. and Liu, H. (2014). Challenges of big data analysis. *National Sci. Rev.* **1** 293–314.

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

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on* **57** 5467–5484.

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

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* **15** 2869–2909.

Kallenberg, O. (1997). *Foundations of modern probability*. Probability and its Applications (New York), Springer-Verlag, New York.

Kleiner, A., Talwalkar, A., Sarkar, P. and Jordan, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 795–816.

Lee, J. D., Sun, Y., Liu, Q. and Taylor, J. E. (2015). Communication-efficient sparse regression: a one-shot approach. *ArXiv 1503.04337* .

Loh, P.-L. and Wainwright, M. J. (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems 26* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Weinberger, eds.). 476–484.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462.

Negahban, S., Yu, B., Wainwright, M. J. and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*.

Ning, Y. and Liu, H. (2014). A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models. *ArXiv 1412.8765* .

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

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

Wang, Z., Liu, H. and Zhang, T. (2014a). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Statist.* **42** 2164–2201.

Wang, Z., Liu, H. and Zhang, T. (2014b). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Statist.* **42** 2164–2201.

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

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science* **27** 576–593.

Zhang, Y., Duchi, J. C. and Wainwright, M. J. (2013). Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm with Minimax Optimal Rates. *ArXiv e-prints* .

Zhao, T., Cheng, G. and Liu, H. (2014a). A Partially Linear Framework for Massive Heterogeneous Data. *ArXiv 1410.8570* .

Zhao, T., Kolar, M. and Liu, H. (2014b). A General Framework for Robust Testing and Confidence Regions in High-Dimensional Quantile Regression. *ArXiv 1412.8724* .

Supplementary material to

Distributed Estimation and Inference with Statistical Guarantees

Heather Battey*† Jianqing Fan* Han Liu* Junwei Lu* Ziwei Zhu*

Abstract

This document contains the supplementary material to the paper "Distributed Estimation and Inference with Statistical Guarantees". In Appendix A, we provide the proofs of technical results required for the analysis of divide and conquer inference. Appendix B collects the proofs of lemmas for the estimation part.

A Auxiliary Lemmas for Inference

In this section, we provide the proofs of the technical lemmas for the divide and conquer inference.

Lemma A.1. Under Condition 3.2, $\left(\mathbf{m}_v^{(j)T}\widehat{\Sigma}\mathbf{m}_v^{(j)}\right)^{-1/2} \geq c_{n_k}$ for any $j \in \{1,\ldots,k\}$ and for any $v \in \{1,\ldots,d\}$, where c_{n_k} satisfies $\liminf_{n_k\to\infty} c_{n_k} = c_\infty > 0$.

Proof. The proof appears in the proof of Lemma B1 of Zhao et al. (2014b). □

Lemma A.2. Under the GLM (2.7), we have

$$\mathbb{E}\exp(t(Y-\mu(\theta))) = \exp(\phi^{-1}(b(\theta+t\phi)-b(\theta)-\phi t b'(\theta))),$$

and typically when there exists $U>0$ such that $b''(\theta)<U$ for all $\theta\in\mathbb{R}$, we will have

$$\mathbb{E}\exp(t(Y-\mu(\theta))) \leq \exp\left(\frac{\phi U t^2}{2}\right),$$

which implies that Y is a sub-Gaussian random variable with variance proxy ϕU.

Proof.

$$\begin{aligned}
\mathbb{E}\exp\left(t(Y-\mu(\theta))\right) &= \int_{-\infty}^{+\infty} c(y)\exp\left(\frac{y\theta-b(\theta)}{\phi}\right)\exp(t(y-\mu(\theta)))dy \\
&= \int_{-\infty}^{+\infty} c(y)\exp\left(\frac{(\theta+t\phi)y-(b(\theta)+\phi t b'(\theta))}{\phi}\right)dy \\
&= \int_{-\infty}^{+\infty} c(y)\exp\left(\frac{(\theta+t\phi)y-b(\theta+t\phi)+b(\theta+t\phi)-(b(\theta)+\phi t b'(\theta))}{\phi}\right)dy \\
&= \exp\left(\phi^{-1}(b(\theta+t\phi)-b(\theta)-\phi t b'(\theta))\right).
\end{aligned}$$

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540; Email: {hbattey,jqfan,hanliu,junweil,ziweiz} @princeton.edu;
†Department of Mathematics, Imperial College London, London, SW7 2AZ; Email: h.battey@imperial.ac.uk

When $b''(\theta) < U$. the mean value theorem gives

$$\mathbb{E}\exp\left(t(Y-\mu(\theta))\right) = \exp\left(\frac{b''(\widetilde{\theta})\phi^2 t^2}{2\phi}\right) \leq \exp\left(\frac{\phi U t^2}{2}\right).$$

□

Lemma A.3. Under Condition 3.6, we have for any $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^d$ and any $i = 1, \ldots, n$, $\left|\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}) - \ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}')\right| \leq K_i|\boldsymbol{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}')|$, where $0 < K_i < \infty$.

Proof. By the canonical form of the generalized linear model (equation (2.8)),

$$\left|\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}) - \ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}')\right| = \left|b''(\boldsymbol{X}_i^T\boldsymbol{\beta}) - b''(\boldsymbol{X}_i^T\boldsymbol{\beta}')\right| \leq |b'''(\widetilde{\eta})||\boldsymbol{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}')|$$

by the mean value theorem, where $\widetilde{\eta}$ lies in a line segment between $\boldsymbol{X}_i^T\boldsymbol{\beta}$ and $\boldsymbol{X}_i^T\boldsymbol{\beta}'$. $|b'''(\eta)| < U_3 < \infty$ by Condition 3.6 for any η, hence the conclusion follows with $K_i = U_3$ for all i. □

Lemma A.4. Under Conditions 2.6 and 2.1 (i), we have for any $\delta \in (0,1)$ such that $\delta^{-1} \ll d$,

$$\mathbb{P}\Big(\frac{1}{n}\big\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*)\big\|_2^2 \gtrsim s\frac{\log(d/\delta)}{n}\Big) < \delta$$

Proof. Decompose the object of interest as

$$\begin{aligned}\frac{1}{n}\big\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*)\big\|_2^2 &= (\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*)^T(\widehat{\Sigma}-\Sigma)(\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*) + (\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*)^T\Sigma(\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*) \\ &\leq \|\widehat{\Sigma}-\Sigma\|_{\max}\|\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*\|_1^2 + \lambda_{\max}(\Sigma)\|\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*\|_2^2.\end{aligned}$$

This gives rise to the tail probability bound

$$\mathbb{P}\Big(\frac{1}{n}\big\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*)\big\|_2^2 > t\Big) \leq \mathbb{P}\Big(\|\widehat{\Sigma}-\Sigma\|_{\max}\|\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*\|_1^2 > \frac{t}{2}\Big) + \mathbb{P}\Big(\lambda_{\max}(\Sigma)\|\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*\|_2^2 > \frac{t}{2}\Big). \quad \text{(A.1)}$$

Let $\mathcal{M} := \big\{\|\widehat{\Sigma}-\Sigma\|_\infty \leq M\big\}$. Since $\{\boldsymbol{X}_i\}_{i=1}^n$ is bounded, it is sub-Gaussian as well. Suppose $\|\boldsymbol{X}_i\|_{\psi_2} < \kappa$, then by Lemma B.2 we have,

$$\begin{aligned}\mathbb{P}(\mathcal{M}^c) &\leq \sum_{p,q=1}^d \mathbb{P}(|\widehat{\Sigma}_{pq}^{(j)} - \Sigma_{pq}| > M) \\ &\leq d^2\exp\left(-Cn\cdot\min\Big\{\frac{M^2}{\kappa^4}, \frac{M}{\kappa^2}\Big\}\right),\end{aligned}$$

where C is a constant. Hence taking $M = n^{-1}\log(d/\delta)$,

$$\mathbb{P}(\mathcal{M}^c) \leq d^2\exp\left\{-Cn\min\Big\{\frac{(\log(d/\delta))^2}{\kappa^4 n^2}, \frac{(\log(d/\delta))^2}{\kappa^2 n}\Big\}\right\}$$

and the right hand side is less than δ for $\delta^{-1} \ll d$. Thus by Condition 2.1, the first term on the right hand side of equation (A.1) is

$$\mathbb{P}\Big(\big\|\widehat{\Sigma}-\Sigma\big\|_{\max}\big\|\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*\big\|_1^2 \gtrsim \frac{s\log(d/\delta)}{n}\Big) < 2\delta.$$

Furthermore, by Condition 3.6 (i), the second term on the right hand side of equation (A.1) is

$$\mathbb{P}\Big(\lambda_{\max}(\Sigma)\big\|\widehat{\boldsymbol{\beta}}^{\lambda}-\boldsymbol{\beta}^{*}\big\|_2^2 \gtrsim C_{\max}\frac{s\log(d/\delta)}{n}\Big)<\delta.$$

Taking t as the dominant term, $t \asymp C_{\max}n^{-1}s\log(d/\delta)$, yields the result. □

Lemma A.5. Under Condition 3.6, we have for any $i=1,\ldots,n$,

$$|b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)-b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_2)| \le MU_3\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_1,$$

and if we consider the sub-Gaussian design instead, we have

$$\mathbb{P}\left(|b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)-b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_2)| \ge hU_3\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_1\right) \le nd\exp\left(1-\frac{Ch^2}{s_1^2}\right).$$

Proof. For the bounded design, by Condition 3.6 (iii), we have

$$|b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)-b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_2)| \le U_3|\boldsymbol{X}_i^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)| \le U_3\|\boldsymbol{X}_i\|_{\max}\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_1 \le MU_3\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_1.$$

For the sub-Gaussian design, denote the event $\{\max_{1\le i\le n,1\le j\le d}|X_{ij}| \le h\}$ by $\mathcal{C}$, where κ is a positive constant. Then it follows that,

$$\mathbb{P}\left(\mathcal{C}^c\right) \le nd\exp\left(1-\frac{Ch^2}{s_1^2}\right),$$

where C is a constant. Since on the event $\mathcal{C}$, $|b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)-b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_2)| \le hU_3\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_1$, we reach the conclusion. □

Remark A.6. For the sub-Gaussian design, in order to let the tail probability go to zero, $h \gg \log((n\vee d))$.

Lemma A.7. Suppose, for any $k \ll d$ satisfying $k=o\big(((s\vee s_1)logd)^{-1}\sqrt{n}\big)$, the following conditions are satisfied. (A1) $\mathbb{P}\left(n_k^{-1}\big\|X^{(j)}\widehat{\boldsymbol{\Theta}}^{(j)}\big\|_{\max} \ge H\right) \le \xi$, where H is a constant and $\xi=o(k^{-1})$. (A2) For any $\boldsymbol{\beta},\boldsymbol{\beta}' \in \mathbb{R}^d$ and for any $i\in\{1,\ldots,n\}$, $\big|\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta})-\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}')\big| \le K_i\big|\boldsymbol{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}')\big|$ with $\mathbb{P}(K_i>h)\le\psi$ for $\psi=o(k^{-1})$ and $h=O(1)$. (A3) $\mathbb{P}\Big(n_k^{-1}\big\|X^{(j)}(\widehat{\boldsymbol{\beta}}^{\lambda}-\boldsymbol{\beta}^{*})\big\|_2^2 \gtrsim n^{-1}sk\log(d/\delta)\Big)<\delta$. (A4) $\mathbb{P}\Big(\max_{1\le v\le d}\Big|\big(\widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla^2\ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j))-\boldsymbol{e}_v\big)\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)-\boldsymbol{\beta}^{*}\big)\Big| \gtrsim n^{-1}sk\log(d/\delta)\Big)<\delta$. Then

$$\overline{\beta}_v^d-\beta_v^* = -\frac{1}{k}\sum_{j=1}^{k}\widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla\ell_{n_k}^{(j)}(\boldsymbol{\beta}^*)+o_{\mathbb{P}}(n^{-1/2}).$$

for any $1\le v\le d$.

Proof of Lemma A.7. $\overline{\beta}_v^d-\beta_v^* = k^{-1}\sum_{j=1}^{k}\big(\widehat{\beta}_v(\mathcal{D}_j)-\beta_v^*\big)$. By the definition of $\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$,

$$\widehat{\beta}_v^d(\mathcal{D}_j)-\beta_v^* = \widehat{\beta}_v^{\lambda}(\mathcal{D}_j)-\beta_v^*-\widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla\ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)).$$

Consider a mean value expansion of $\nabla \ell_{n_k}^{(j)}\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)\big)$ around $\boldsymbol{\beta}^*$:

$$\nabla \ell_{n_k}^{(j)}\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)\big) = \nabla \ell_{n_k}^{(j)}(\boldsymbol{\beta}^*) + \nabla^2 \ell_{n_k}^{(j)}(\boldsymbol{\beta}_\alpha)\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\big),$$

where $\boldsymbol{\beta}_\alpha = \alpha\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) + (1-\alpha)\boldsymbol{\beta}^*$, $\alpha \in [0,1]$. So

$$\frac{1}{k}\sum_{j=1}^{k} \widehat{\beta}_v^d(\mathcal{D}_j) - \beta_v^* = -\frac{1}{k}\sum_{j=1}^{k} \widehat{\boldsymbol{\Theta}}_v^{(j)T} \nabla \ell_{n_k}^{(j)}(\boldsymbol{\beta}^*) - \underbrace{\frac{1}{k}\sum_{j=1}^{k}\big(\widehat{\boldsymbol{\Theta}}_v^{(j)T} \nabla^2 \ell_{n_k}^{(j)}(\boldsymbol{\beta}_\alpha) - \boldsymbol{e}_v\big)\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)}_{\Delta}$$

and $|\Delta| \le \frac{1}{k}\sum_{j=1}^{k}\big(|\Delta_1^{(j)}| + |\Delta_2^{(j)}|\big)$ where

$$\big|\Delta_1^{(j)}\big| = \Big|\big(\widehat{\boldsymbol{\Theta}}_v^{(j)T} \nabla^2 \ell_{n_k}^{(j)}\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)\big) - \boldsymbol{e}_v\big)\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\Big|.$$

By (A4) of the lemma, for $t \asymp n^{-1}sk\log(d/\delta)$,

$$\mathbb{P}\Big(|\sum_{j=1}^{k}\Delta_1^{(j)}| > kt\Big) \le \mathbb{P}\Big(\cup_{j=1}^{k}|\Delta_1^{(j)}| > t\Big) \le \sum_{j=1}^{k}\mathbb{P}(|\Delta_1^{(j)}| > t) < k\delta.$$

Substituting $\delta = o(k^{-1})$ in the expression for t and noting that $k \ll d$, we obtain $k^{-1}\sum_{j=1}^{k}\Delta_1^{(j)} = o_{\mathbb{P}}(n^{-1/2})$ for $k = o\big((s\log d)^{-1}\sqrt{n}\big)$. By (A2),

$$\begin{aligned}
\big|\Delta_2^{(j)}\big| &= \Big|\widehat{\boldsymbol{\Theta}}_v^{(j)T}\big(\nabla^2 \ell_{n_k}^{(j)}(\boldsymbol{\beta}_\alpha) - \nabla^2 \ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j))\big)\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\Big| \\
&= \Big|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\widehat{\boldsymbol{\Theta}}_v^{(j)T}\boldsymbol{X}_i\boldsymbol{X}_i^T\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\big(\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}_\alpha) - \ell_i''(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j))\big)\Big| \\
&\le \Big(\max_{1\le i\le n} K_i\Big)\Big(\frac{1}{n_k}\|X^{(j)}\widehat{\Theta}^{(j)}\|_{\max}\Big)\Big\|\frac{1}{n_k}X^{(j)}\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\big)\Big\|_2^2,
\end{aligned}$$

therefore by (A1) and (A3) of the lemma, for $t \asymp n^{-1}sk\log(d/\delta)$,

$$\mathbb{P}\Big(|\sum_{j=1}^{k}\Delta_2^{(j)}| > kt\Big) \le \mathbb{P}\Big(\cup_{j=1}^{k}|\Delta_2^{(j)}| > t\Big) \le \sum_{j=1}^{k}\mathbb{P}(|\Delta_2^{(j)}| > t) < k(\psi+\delta+\xi).$$

Substituting $\delta = o(k^{-1})$ in the expression for t and noting that $k \ll d$, we obtain $k^{-1}\sum_{j=1}^{k}\Delta_2^{(j)} = o_{\mathbb{P}}(n^{-1/2})$ for $sk\log(d/\delta) = o(\sqrt{n})$, i.e. for $k = o\big((s\log d)^{-1}\sqrt{n}\big)$. Combining these two results delivers $\Delta = o_{\mathbb{P}}(n^{-1/2})$ for $k = o\big((s\log d)^{-1}\sqrt{n}\big)$. □

Lemma A.8. Suppose, in addition to Conditions (A1)-(A5) of Lemma A.7, (A5) $\big|\widetilde{\Theta}_{vv} - \Theta_{vv}^*\big| = o_{\mathbb{P}}(1)$ for all $v \in \{1,\dots,d\}$; (A6) $1/\Theta_{vv}^* = O(1)$ for all $v \in \{1,\dots,d\}$; (A7) $\|\sum_{1\le j\le k}\sum_{i\in\mathcal{I}_j}\nabla\ell_i(\boldsymbol{\beta}^*)\|_\infty = O_{\mathbb{P}}(\sqrt{n\log d})$; (A8) For each $v \in \{1,\dots,d\}$, letting $\xi_{iv}^{(j)} = \boldsymbol{\Theta}_v^{*T}\nabla\ell_i^{(j)}(\boldsymbol{\beta}^*)/\sqrt{n\Theta_{vv}^*}$, $\mathbb{E}\big[\xi_{iv}^{(j)}\big] = 0$, $\mathrm{Var}\big(\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\xi_{iv}^{(j)}\big) = 1$ and, for all $\varepsilon > 0$,

$$\lim_{k\to\infty}\lim_{n_k\to\infty}\sum_{j=1}^{k}\sum_{i\in\mathcal{D}_j}\mathbb{E}\big[(\xi_{iv}^{(j)})^2\,\mathbb{1}\{|\xi_{iv}^{(j)}| > \varepsilon\}\big] = 0. \tag{A.2}$$

Then under $H_0 : \beta_v^* = \beta_v^H$, taking $k = o(((s \vee s_1)\log d)^{-1}\sqrt{n})$ delivers $\overline{S}_n \rightsquigarrow N(0,1)$, where $\overline{S}_n$ is defined in equation (3.13).

Proof. Rewrite equation (3.13) as

$$\begin{aligned}\overline{S}_n &= \sqrt{n}\frac{1}{k}\sum_{j=1}^{k}\left[\frac{\widehat{\beta}_v^d - \beta_v^H}{(\Theta_{vv}^*)^{1/2}} + \frac{\widehat{\beta}_v^d - \beta_v^H}{(\Theta_{vv}^*)^{1/2}}\left(\frac{(\Theta_{vv}^*)^{1/2}}{\left[\widehat{\Theta}^{(j)}\widehat{H}^{(j)}\widehat{\Theta}^{(j)T}\right]_{vv}^{1/2}} - 1\right)\right] \\ &= \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}(\Delta_{1,i}^{(j)} + \Delta_{2,i}^{(j)}), \qquad \text{where} \end{aligned} \tag{A.3}$$

$$\Delta_{1,i}^{(j)} = \frac{\widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla\ell_i^{(j)}(\boldsymbol{\beta}^*)}{(n\Theta_{vv}^*)^{1/2}}, \qquad \Delta_{2,i}^{(j)} = \frac{\widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla\ell_i^{(j)}(\boldsymbol{\beta}^*)}{(n\Theta_{vv}^*)^{1/2}}\left(\frac{(\Theta_{vv}^*)^{1/2}}{\overline{\Theta}_{vv}^{1/2}} - 1\right).$$

Further decomposing the first term, we have

$$\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\Delta_{1,i}^{(j)} = \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\xi_{i,v}^{(j)} + \Delta, \quad \text{where} \quad \Delta = \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\left(\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*\right)^T\frac{\nabla\ell_i(\boldsymbol{\beta}^*)}{(n\Theta_{vv}^*)^{1/2}}$$

and $\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\xi_{i,v}^{(j)} \rightsquigarrow N(0,1)$ by the Lindeberg-Feller central limit theorem. Then by Hölder's inequality, Condition 3.7 and Assumption (A6) and (A7),

$$\begin{aligned}|\Delta| &\leq \max_{1\leq j\leq k}\left\|\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*\right\|_1\frac{\left\|\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\nabla\ell_i(\boldsymbol{\beta}^*)\right\|_\infty}{(n\Theta_{vv}^*)^{1/2}} \\ &= O_{\mathbb{P}}\left(s_1\sqrt{\frac{k\log d}{n}}\right)O_{\mathbb{P}}(\sqrt{\log d}) = o_{\mathbb{P}}(1),\end{aligned}$$

where the last equation holds with the choice of $k = o((s_1\log d)^{-1}\sqrt{n})$. Letting $\overline{\Delta}^{(j)} = (\Theta_{vv}^*)^{1/2} - \overline{\Theta}_{vv}^{1/2}$ we have

$$\begin{aligned}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\Delta_{2,i}^{(j)} &= \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\frac{\boldsymbol{\Theta}_v^{*T}\nabla\ell_i^{(j)}(\boldsymbol{\beta}^*)}{(\Theta_{vv}^*)^{1/2}}\overline{\Delta}^{(j)} + \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\left(\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*\right)^T\frac{\nabla\ell_i(\boldsymbol{\beta}^*)}{(\boldsymbol{\Theta}_{vv}^*)^{1/2}}\overline{\Delta}^{(j)} \\ &= \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}(\Delta_{21,i}^{(j)} + \Delta_{22,i}^{(j)}), \text{ where} \\ &\left|\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\Delta_{21,i}^{(i)}\right| \leq \left|\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\xi_{i,v}^{(j)}\right|\left|\overline{\Theta}_{vv}^{1/2} - (\Theta_{vv}^*)^{1/2}\right|.\end{aligned}$$

Since $\Theta_{vv}^* \geq 0$, $\overline{\Theta}_{vv}^{1/2} = |\overline{\Theta}_{vv}|^{1/2} = |\overline{\Theta}_{vv} - \Theta_{vv}^* + \Theta_{vv}^*|^{1/2} \leq |\overline{\Theta}_{vv} - \Theta_{vv}^*|^{1/2} + (\Theta_{vv}^*)^{1/2}$. Similarly

$$(\Theta_{vv}^*)^{1/2} = |\Theta_{vv}^*|^{1/2} = |\Theta_{vv}^* - \overline{\Theta}_{vv} + \overline{\Theta}_{vv}|^{1/2} \leq |\Theta_{vv}^* - \overline{\Theta}_{vv}|^{1/2} + \overline{\Theta}_{vv}^{1/2},$$

yielding $|\overline{\Theta}_{vv}^{1/2} - (\Theta_{vv}^*)^{1/2}| \leq |\overline{\Theta}_{vv} - \Theta_{vv}^*|^{1/2}$ and consequently, by assumption (A5),

$$\left|\overline{\Delta}^{(j)}\right| = \left|\overline{\Theta}_{vv}^{1/2} - (\Theta_{vv}^*)^{1/2}\right| = o_{\mathbb{P}}(1).$$

Invoking (A9) and the Lindeberg-Feller CLT, $\left|\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\Delta_{21,i}^{(i)}\right| = O_{\mathbb{P}}(1)o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$. Similarly

$$\left|\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\Delta_{22,i}^{(j)}\right| \leq \max_{1\leq j\leq k}\|\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*\|_1\left|\overline{\Delta}^{(j)}\right|\left|\left(\boldsymbol{\Theta}_v^{*T}\boldsymbol{\Theta}_v^*\right)^{-1/2}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\xi_{iv}^{(j)}\right| = o_{\mathbb{P}}(1).$$

Combining all terms in the decomposition (A.3) delivers the result. □

(B1)-(B5) of Condition A.9 are used in the proofs of subsequent lemmas.

Condition A.9 . (B1) $\|\boldsymbol{w}^*\|_1 \lesssim s_1$, $\|J^*\|_{\max} < \infty$ and for any $\delta \in (0,1)$,

$$\mathbb{P}\Big(\|\widehat{\boldsymbol{\beta}}_{-v}^{\lambda} - \boldsymbol{\beta}_{-v}^*\|_1 \gtrsim n^{-1/2}s\sqrt{\log(d/\delta)}\Big) < \delta \quad \text{and} \quad \mathbb{P}\Big(\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\|_1 \gtrsim n^{-1/2}s_1\sqrt{\log(d/\delta)}\Big) < \delta.$$

(B2) For any $\delta \in (0,1)$,

$$\mathbb{P}\Big(\|\nabla_{-v}\ell_n(\beta_v^*, \boldsymbol{\beta}_{-v}^*)\|_\infty \gtrsim n^{-1/2}\sqrt{\log(d/\delta)}\Big) < \delta.$$

(B3) Suppose $\widehat{\boldsymbol{\beta}}_{-v}^{\lambda}$ satisfies (B1). Then for $\boldsymbol{\beta}_{-v,\alpha} = \alpha\boldsymbol{\beta}_{-v}^* + (1-\alpha)\widehat{\boldsymbol{\beta}}_{-v}^{\lambda}$ and for any $\delta \in (0,1)$,

$$\mathbb{P}\left(\sup_{\alpha\in[0,1]}\left|\left(\nabla_{v,-v}^2\ell_n(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha}) - \widehat{\boldsymbol{w}}^T\nabla_{-v,-v}^2\ell_n(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha})\right)(\widehat{\boldsymbol{\beta}}_{-v}^{\lambda} - \boldsymbol{\beta}_{-v}^*)\right| \gtrsim s_1 s\frac{\log(d/\delta)}{n}\right) < \delta.$$

(B4) There exists a constant $C > 0$ such that $C < I_{\theta|\boldsymbol{\gamma}}^* < \infty$, and for $\boldsymbol{v}^* = (1, -\boldsymbol{w}^{*T})^T$, it holds that

$$\frac{\sqrt{n}\boldsymbol{v}^{*T}\nabla\ell_n(\beta_v^*, \boldsymbol{\beta}_{-v}^*)}{\sqrt{\boldsymbol{v}^{*T}J^*\boldsymbol{v}^*}} \rightsquigarrow N(0,1).$$

(B5) For any δ, if there exists an estimator $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_v^T, \widetilde{\boldsymbol{\beta}}_{-v}^T)^T$ satisfying $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq Cs\sqrt{n^{-1}\log(d/\delta)}$ with probability $> 1-\delta$, then

$$\mathbb{P}\Big(\left\|\nabla^2\,\ell_n(\widetilde{\boldsymbol{\beta}}) - J^*\right\|_{\max} \gtrsim n^{-1/2}\sqrt{\log(d/\delta)}\Big) < \delta.$$

The proof of Theorem 3.11 is an application of Lemma A.13. To apply this Lemma, we must first verify (B1) to (B4) of Condition A.9. We do this in Lemma A.10.

Lemma A.10. Under the requirements of Theorem 3.11, (B1) - (B4) of Condition A.9 are fulfilled.

Proof. **Verification of (B1).** As stated in Theorem 3.11, $\|\boldsymbol{w}^*\|_1 = O(s_1)$ and $\|J^*\|_{\max} < \infty$ by part (i) of Condition 3.6. The rest of (B1) follows from the proof of Lemma C.3 of Ning and Liu (2014).

Verification of (B2). Let $\boldsymbol{X}_i = (Q_i, \boldsymbol{Z}_i^T)^T$. Since $\| \nabla_{\boldsymbol{\gamma}} \ell_n(\boldsymbol{\beta}^*)\|_\infty = \big\| -\frac{1}{n}\sum_{i=1}^n \big(Y_i - b'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\big)\boldsymbol{Z}_i \big\|_\infty$, since the product of a subgaussian random variable and a bounded random variable is subgaussian, and since $\mathbb{E}[\nabla_{\boldsymbol{\gamma}} \ell_n(\boldsymbol{\beta}^*)] = 0$, we have by Condition 3.6, Bernstein's inequality and the union bound

$$\mathbb{P}\big(\| \nabla_{\boldsymbol{\gamma}} \ell_n(\boldsymbol{\beta}^*)\|_\infty > t\big) < (d-1)\exp\{-nt^2/M^2\sigma_b^2\}.$$

Setting $2(d-1)\exp\{-nt^2/M^2\sigma_b^2\} = \delta$ and solving for t delivers the result.

Verification of (B3) Let $\boldsymbol{\beta}_\alpha^* = (\theta^*, \boldsymbol{\gamma}_\alpha)$ and decompose the object of interest as

$$\big|\big(\nabla^2_{v,-v}\ell_n(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha}) - \widehat{\boldsymbol{w}}^T\nabla^2_{-v,-v}\ell_n(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha})\big)(\widehat{\boldsymbol{\beta}}^\lambda_{-v} - \boldsymbol{\beta}^*_{-v})\big| \le \sum_{t=1}^5 |\Delta_t|, \tag{A.4}$$

where the terms Δ_1 - Δ_5 are given by $\Delta_1 = \nabla^2_{v,-v}\ell_n(\boldsymbol{\beta}^*_\alpha) - \nabla^2_{v,-v}\ell_n(\boldsymbol{\beta}^*)$,

$$\begin{aligned}
\Delta_2 &= \nabla^2_{v,-v}\ell_n(\boldsymbol{\beta}^*) - \boldsymbol{w}^{*T}J^*_{-v,-v}, & \Delta_3 &= \boldsymbol{w}^{*T}\big(J^*_{-v,-v} - \nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*)\big),\\
\Delta_4 &= \boldsymbol{w}^{*T}\big(\nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*) - \nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*_\alpha)\big), & \Delta_5 &= (\boldsymbol{w}^{*T} - \widehat{\boldsymbol{w}}^T)\nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*_\alpha).
\end{aligned}$$

We have the following bounds

$$\begin{aligned}
|\Delta_1| &= \Big|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T(\widehat{\boldsymbol{\beta}}^\lambda_{-v} - \boldsymbol{\beta}^*_{-v})\big(\ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*_\alpha) - \ell_i''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\big)\Big|\\
&\le \max_{1\le i\le n} K_i \max_{1\le i\le n}\|\boldsymbol{X}_i\|_\infty \big\|\frac{1}{n}\boldsymbol{Z}(\widehat{\boldsymbol{\beta}}_{-v} - \boldsymbol{\beta}^*_{-v})\big\|_2^2,
\end{aligned}$$

$$|\Delta_2| \le \big\|\nabla^2_{v,-v}\ell_n(\boldsymbol{\beta}^*) - J^*_{v,-v}\big\|_\infty \|\widehat{\boldsymbol{\beta}}^\lambda_{-v} - \boldsymbol{\beta}^*_{-v}\|_1,\ |\Delta_3| \le \|\boldsymbol{w}\|_1\big\|J^*_{-v,-v} - \nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*)\big\|_{\max}\|\widehat{\boldsymbol{\beta}}^\lambda_{-v} - \boldsymbol{\beta}^*_v\|_1,$$

$$\begin{aligned}
|\Delta_4| &= \big|\boldsymbol{w}^{*T}\big(\nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*) - \nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}^*_v)\big)(\widehat{\boldsymbol{\gamma}}^\lambda - \boldsymbol{\lambda}^*)\big|\\
&\le \max_{1\le i\le n} K_i\|\boldsymbol{w}^*\|_1\big\|\frac{1}{n}\boldsymbol{Z}(\widehat{\boldsymbol{\beta}}^\lambda_{-v} - \boldsymbol{\beta}^*_{-v})\big\|_2^2,
\end{aligned}$$

and $|\Delta_5| \le \|\boldsymbol{w}^* - \widehat{\boldsymbol{w}}\|_1\big\|\nabla_{-v,-v}\ell_n(\boldsymbol{\beta}^*_v)\big\|_{\max}\|\widehat{\boldsymbol{\beta}}^\lambda_{-v} - \boldsymbol{\beta}^*_{-v}\|_1$. Let $\varepsilon = \delta/5$. Then by Condition 3.6 and Lemma A.4

$$\mathbb{P}\Big(|\Delta_1| \gtrsim s\frac{\log(d/\varepsilon)}{n}\Big) < \varepsilon \quad \text{and} \quad \mathbb{P}\Big(|\Delta_4| \gtrsim ss_1\frac{\log(d/\varepsilon)}{n}\Big) < \varepsilon.$$

Noting the $\boldsymbol{\beta}^*$ itself satisfies the requirements on $\widetilde{\boldsymbol{\beta}}$ in (B5), Lemma A.11 and Condition 2.1 together give

$$\mathbb{P}\Big(|\Delta_2| \gtrsim s_1\frac{\log(d/\varepsilon)}{n}\Big) < \varepsilon \quad \text{and} \quad \mathbb{P}\Big(|\Delta_3| \gtrsim s_1 s\frac{\log(d/\varepsilon)}{n}\Big) < \varepsilon.$$

By (B1) verified above and noting that

$$\big\|\nabla_{-v,-v}\ell_n(\boldsymbol{\beta}^*_v)\big\|_{\max} \le \big\|\nabla_{-v,-v}\ell_n(\boldsymbol{\beta}^*_v) - \nabla_{-v,-v}\ell_n(\boldsymbol{\beta}^*)\big\|_{\max} + \big\|\nabla_{-v,-v}\ell_n(\boldsymbol{\beta}^*)\big\|_{\max},$$

the proof of Lemma A.11 delivers $\mathbb{P}\Big(|\Delta_5| \gtrsim s_1 s\log(d/\varepsilon)/n\Big) < \varepsilon$. Combining the bounds, we finally have

$$\mathbb{P}\Big(\sup_{\alpha\in[0,1]}\big|\big(\nabla^2_{v,-v}\ell_n(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha}) - \widehat{\boldsymbol{w}}^T\nabla^2_{-v,-v}\ell_n(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha})\big)(\widehat{\boldsymbol{\beta}}^\lambda_{-v} - \boldsymbol{\beta}^*_{-v})\big| \gtrsim s_1 s\frac{\log(d/\delta)}{n}\Big) < \delta$$

Verification of (B4). See Ning and Liu (2014), proof of Lemma C.2. □

In the following lemma, we verify (B5) under the same conditions.

Lemma A.11. Under Conditions 3.6 and 2.1, (B5) of Condition A.9 is fulfilled.

Proof. We obtain a tail probability bound for Δ_1 and Δ_2 in the decomposition

$$\| \nabla^2 \ell_n(\widetilde{\boldsymbol{\beta}}) - J^* \|_{\max} \leq \| \nabla^2 \ell_n(\widetilde{\boldsymbol{\beta}}) - \nabla^2 \ell_n(\boldsymbol{\beta}^*) \|_{\max} + \| \nabla^2 \ell_n(\boldsymbol{\beta}^*) - J^* \|_{\max} = \Delta_1 + \Delta_2.$$

For the control over Δ_1, note that by Condition 3.6 (ii) and (iii),

$$\big| [\nabla^2 \ell_n(\boldsymbol{\beta}^*)]_{jk} \big| \leq \big| b''(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) \big| \big| X_{ij} X_{ik} \big| \leq U_2 M^2.$$

Hence Hoeffding's inequality and the union bound deliver

$$\mathbb{P}(\Delta_2 > t) = \mathbb{P}\Big(\| \nabla^2 \ell_n(\boldsymbol{\beta}^*) - J^* \|_{\max} > t \Big) \leq 2d^2 \exp\Big\{ -\frac{nt^2}{8U_2^2 M^4} \Big\}. \tag{A.5}$$

For the control over Δ_1, we have by Lemma A.5,

$$\begin{aligned} \big| [\nabla^2 \ell_n(\widetilde{\boldsymbol{\beta}}) - \nabla^2 \ell_n(\boldsymbol{\beta}^*)]_{jk} \big| &= \big| \big(b''(\boldsymbol{X}_i^T \widetilde{\boldsymbol{\beta}}) - b''(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) \big) X_{ij} X_{ik} \big| \\ &\leq M^3 U_3 \| \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \|_1 \leq M^3 U_3 s \sqrt{n^{-1} \log(d/\delta)} \end{aligned}$$

with probability $> 1 - \delta$. Hoeffding's inequality and the union bound again deliver

$$\mathbb{P}(\Delta_1 > t) = \mathbb{P}\Big(\| \nabla^2_{\eta\eta} \ell_n(\widetilde{\boldsymbol{\beta}}) - \nabla^2_{\eta\eta} \ell_n(\boldsymbol{\beta}^*) \|_{\max} > t \Big) \leq 2d^2 \exp\Big\{ -\frac{n^2 t^2}{8U_3^2 M^6 s^2 \log(d/\delta)} \Big\}. \tag{A.6}$$

Combining the bounds from equations (A.5) and (A.6) we have

$$\mathbb{P}\Big(\| \nabla^2 \ell(\widetilde{\boldsymbol{\beta}}) - J^* \|_{\max} > t \Big) \leq 2d^2 \Big(\exp\Big\{ -\frac{nt^2}{8U_3^2 M^4} \Big\} + \exp\Big\{ -\frac{n^2 t^2}{8U_3^2 M^6 s^2 \log(d/\delta)} \Big\} \Big).$$

Setting each term equal to $\delta/2$, solving for t and ignoring the relative magnitude of constants, we have $t = U_3 \max\big\{ n^{-1} s \log(d/\delta), n^{-1/2} \sqrt{\log(d/\delta)} \big\} = U_3 n^{-1/2} \log(d/\delta)$, thus verifying (B5). □

Lemma A.12. For each $j \in \{1, \ldots, k\}$, let $\boldsymbol{\beta}_{-v,\alpha_j} = \alpha_j \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j) + (1 - \alpha_j) \boldsymbol{\beta}^*_{-v}$, for some $\alpha_j \in [0, 1]$, where $\widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)$ is defined in equation (2.2). Define

$$\begin{aligned} \Delta_1^{(j)} &= (\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*)^T \nabla_{-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}^*_{-v}) \\ \Delta_2^{(j)} &= \big(\nabla^2_{v,-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}_{-v,\alpha_j}) - \widehat{\boldsymbol{w}}^T \nabla_{-v,-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}_{-v,\alpha_j}) \big) (\widehat{\boldsymbol{\beta}}^{\lambda}_{-v} - \boldsymbol{\beta}^*_{-v}). \end{aligned}$$

Under (B1) - (B3) of Condition A.9, $\Big| k^{-1} \sum_{j=1}^k \Delta_1^{(j)} \Big| = o_{\mathbb{P}}\big(n^{-1/2} \big)$ and $\Big| k^{-1} \sum_{j=1}^k \Delta_2^{(j)} \Big| = o_{\mathbb{P}}\big(n^{-1/2} \big)$ whenever $k \ll d$ is chosen to satisfy $k = o\big((s_1 \log d)^{-1} \sqrt{n} \big)$.

Proof. By Hölder's inequality,

$$\big| \Delta_1^{(j)} \big| = \big| \big(\boldsymbol{w}^* - \widehat{\boldsymbol{w}}(\mathcal{D}_j) \big)^T \nabla_{-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}^*_{-v}) \big| \leq \| \widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^* \|_1 \| \nabla_{-v} \ell^{(j)}_{n_k}(\beta^*_v, \boldsymbol{\beta}^*_{-v}) \|_\infty,$$

hence, for any t,

$$\big\{\big|\Delta_1^{(j)}\big| > t\big\} \subseteq \big\{\|\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*\|_1 \|\nabla_{-v}\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*)\|_\infty > t\big\}.$$

Taking $t = vq$ where $v = Cn^{-1/2}s_1\sqrt{k\log(d/\delta)}$ and $q = Cn^{-1/2}\sqrt{k\log(d/\delta)}$, we have

$$\begin{aligned}
&\mathbb{P}\Big(\big\{\|\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*\|_1 \|\nabla_{-v}\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*)\|_\infty > vq\big\}\Big) \\
= \;& \mathbb{P}\Big(\big\{\|\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*\|_1 \|\nabla_{-v}\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*)\|_\infty > vq\big\} \cap \Big\{\frac{\|\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*\|_1}{v} \le 1\Big\}\Big) \\
+ \;& \mathbb{P}\Big(\big\{\|\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*\|_1 \|\nabla_{-v}\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*)\|_\infty > vq\big\} \cap \Big\{\frac{\|\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*\|_1}{v} > 1\Big\}\Big) \le 2\delta
\end{aligned}$$

by (B1) and (B2) of Condition A.9. Hence the union bound delivers

$$\mathbb{P}\Big(\big|\sum_{j=1}^k \Delta_1^{(j)}\big| > kvq\Big) \le \mathbb{P}\Big(\cup_{j=1}^k\big\{\big|\Delta_1^{(j)}\big| > vq\big\}\Big) \le \sum_{j=1}^k \mathbb{P}\Big(\big|\Delta_1^{(j)}\big| > vq\Big) \le 2k\delta = o(1)$$

for $\delta = o(k^{-1})$. Taking $\delta = k^{-1}$ for $\alpha > 0$ arbitrarily small in the definition of v and q, the requirement is $ks_1\log d = o\big(\sqrt{n}\big)$ and $ks_1\log k = o(\sqrt{n})$ for $\alpha > 0$ arbitrarily small. Since $k \ll d$, $k^{-1}\sum_{j=1}^k \Delta_1^{(j)} = o_{\mathbb{P}}\big(n^{-1/2}\big)$ with $k = o\big((s_1\log d)^{-1}\sqrt{n}\big)$. Next, consider

$$\big|\Delta_2^{(j)}\big| \le \sup_{\alpha\in[0,1]}\Big|\big(\nabla^2_{v,-v}\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha}) - \widehat{\boldsymbol{w}}^T\nabla^2_{-v,-v}\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha})\big)(\widehat{\boldsymbol{\beta}}^\lambda_{-v}(\mathcal{D}_j) - \boldsymbol{\beta}^*_{-v})\Big|.$$

By (B3) of Condition A.9, $\mathbb{P}\big(\big|\Delta_2^{(j)}\big| \ge t\big) < \delta$ for $t \asymp s_1 s n^{-1}k\log(d/\delta)$, hence, proceeding in an analogous fashion to in the control over $k^{-1}\sum_{j=1}^k \Delta_1^{(j)}$, we obtain

$$\mathbb{P}\Big(\Big|\sum_{j=1}^k \Delta_2^{(j)}\Big| > kt\Big) \le \mathbb{P}\left(\cup_{j=1}^k\big|\Delta_2^{(j)}\big| > t\right) \le \sum_{j=1}^k \mathbb{P}\left(\big|\Delta_2^{(j)}\big| > t\right) \le k\delta = o(1)$$

for $\delta = o(k^{-1})$. Hence $k^{-1}\sum_{j=1}^k \Delta_2^{(j)} = o_{\mathbb{P}}\big(n^{-1/2}\big)$ with $k = o\big((s_1 s\log d)^{-1}n^{3/2}\big)$. Since $(s_1\log d)^{-1}\sqrt{n} = o\big((s_1 s\log d)^{-1}n^{3/2}\big)$, $k^{-1}\sum_{j=1}^k(\Delta_1^{(j)} + \Delta_2^{(j)}) = o_{\mathbb{P}}\big(n^{-1/2}\big)$ requires $k = o\big((s_1\log d)^{-1}\sqrt{n}\big)$. □

Lemma A.13. Under (B1) - (B4) of Condition A.9, with $k \ll d$ chosen to satisfy the scaling $k = o\big(((s \vee s_1)\log d)^{-1}\sqrt{n}\big)$,

$$\frac{1}{k}\sum_{j=1}^k \widehat{S}^{(j)}(\beta_v^*, \widehat{\boldsymbol{\gamma}}^\lambda(\mathcal{D}_{\mathrm{j}})) = \frac{1}{k}\sum_{j=1}^k S^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*) + o_{\mathbb{P}}(n^{-1/2}) \;\text{ and}$$

$$\lim_{n\to\infty}\sup_t\big|\mathbb{P}((J^*_{v|-v})^{-1/2}\sqrt{n}\frac{1}{k}\sum_{j=1}^k S^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*) < t) - \Phi(t)\big| \to 0.$$

Proof. Recall

$$S^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*) = \nabla_v\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*) - \boldsymbol{w}^{*T}\nabla_{-v}\ell_{n_k}^{(j)}(\beta_v^*, \boldsymbol{\beta}_{-v}^*).$$

Through a mean value expansion of $\widehat{S}^{(j)}(\beta_v^*, \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j))$ around $\boldsymbol{\beta}^*_{-v}$, we have for each $j \in \{1, \ldots, k\}$,

$$\begin{aligned}\widehat{S}^{(j)}\big(\beta_v^*, \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)\big) &= \nabla_v \ell^{(j)}_{n_k}\big(\beta_v^*, \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)\big) - \widehat{\boldsymbol{w}}(\mathcal{D}_j)^T \nabla_{-v} \ell^{(j)}_{n_k}\big(\beta_v^*, \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)\big) \\ &= S^{(j)}(\beta_v^*, \boldsymbol{\beta}^*_{-v}) + \Delta_1^{(j)} + \Delta_2^{(j)},\end{aligned}$$

for some $\boldsymbol{\beta}_{-v,\alpha} = \alpha \widehat{\boldsymbol{\beta}}_{-v}(\mathcal{D}_j) + (1-\alpha)\boldsymbol{\beta}^*_{-v}$, where

$$\begin{aligned}\Delta_1^{(j)} &= \big(\boldsymbol{w}^* - \widehat{\boldsymbol{w}}(\mathcal{D}_j)\big)^T \nabla_{-v} \ell^{(j)}_{n_k}(\beta_v^*, \boldsymbol{\beta}^*_{-v}) \\ \Delta_2^{(j)} &= \Big[\nabla^2_{v,-v} \ell^{(j)}_{n_k}(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha}) - \widehat{\boldsymbol{w}}(\mathcal{D}_j)^T \nabla^2_{-v,-v} \ell^{(j)}_{n_k}(\beta_v^*, \boldsymbol{\beta}_{-v,\alpha})\Big](\widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j) - \boldsymbol{\beta}^*_{-v}).\end{aligned}$$

It follows that

$$\frac{1}{k}\sum_{j=1}^k \widehat{S}^{(j)}\big(\beta_v^*, \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)\big) = \frac{1}{k}\sum_{j=1}^k S^{(j)}(\beta_v^*, \boldsymbol{\beta}^*_{-v}) + \frac{1}{k}\sum_{j=1}^k \big(\Delta_1^{(j)} + \Delta_2^{(j)}\big) = \frac{1}{k}\sum_{j=1}^k S^{(j)}(\theta^*, \boldsymbol{\gamma}^*) + o_{\mathbb{P}}(n^{-1/2}) \tag{A.7}$$

by Lemma A.12 whenever $k = o\big((s_1 \log d)^{-1}\sqrt{n}\big)$. Observe

$$\sqrt{n}\Big(k^{-1}\sum_{j=1}^k S^{(j)}(\beta_v^*, \boldsymbol{\beta}^*_{-v})\Big) = \sqrt{n}(1, -\boldsymbol{w}^{*T})\Big(\frac{1}{k}\sum_{j=1}^k \nabla\, \ell^{(j)}_{n_k}(\beta_v^*, \boldsymbol{\beta}^*_{-v})\Big) \quad \text{and}$$

$$J^*_{v|-v} = (1, -\boldsymbol{w}^{*T}) J^* (1, -\boldsymbol{w}^{*T})^T.$$

So $\sqrt{n}\frac{1}{k}\sum_{j=1}^k S^{(j)}(\beta_v^*, \boldsymbol{\beta}^*_{-v}) \rightsquigarrow N(0, J^*_{v|-v})$ by Condition (B4). Similar to Corollary 3.9, we apply the Berry-Essen inequality to show that $\sup_t |\mathbb{P}(\sqrt{n}\frac{1}{k}\sum_{j=1}^k S^{(j)}(\beta_v^*, \boldsymbol{\beta}^*_{-v}) < t) - \Phi(t)| \to 0$. □

Lemma A.14. Under Condition (B1), for any $\delta \in (0,1)$,

$$\mathbb{P}\Big(\|\overline{\boldsymbol{w}} - \boldsymbol{w}^*\|_1 > Cn^{-1/2} s_1 \sqrt{k \log(d/\delta)}\Big) < k\delta \quad \text{and} \quad \mathbb{P}\Big(\|\overline{\boldsymbol{\beta}}_{-v} - \boldsymbol{\beta}^*_{-v}\|_1 > Cn^{-1/2} s \sqrt{k \log(d/\delta)}\Big) < k\delta.$$

Proof. Set $t = Cs_1\sqrt{n^{-1}(k\log(d/\delta))}$ and note

$$\mathbb{P}\big(\|\sum_{j=1}^k (\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*)\|_1 > kt\big) \le \mathbb{P}\big(\cup_{j=1}^k \|\widehat{\boldsymbol{w}}(\mathcal{D}_j) - \boldsymbol{w}^*\|_1 > t\big) \le \sum_{j=1}^k \mathbb{P}\big(\|\overline{\boldsymbol{w}} - \boldsymbol{w}^*\|_1 > t\big)$$

by the union bound. Then by Condition (B1), $\mathbb{P}\Big(\|\overline{\boldsymbol{w}} - \boldsymbol{w}^*\|_1 > Cn^{-1/2} s_1 \sqrt{k \log(d/\delta)}\Big) < k\delta$. The proof of the second bound is analogous, setting $t = Cs\sqrt{n^{-1}(k\log(d/\delta))}$. □

Lemma A.15. Suppose (B5) of Condition A.9 is satisfied. For any δ, if there exists an estimator $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_v^T, \widetilde{\boldsymbol{\beta}}_{-v}^T)^T$ satisfying $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le Cs\sqrt{n^{-1}\log(d/\delta)}$ with probability $1-\delta$, then

$$\mathbb{P}\Big(\Big\|\frac{1}{k}\sum_{j=1}^k \nabla^2\, \ell^{(j)}_{n_k}(\widetilde{\boldsymbol{\beta}}) - J^*\Big\|_{\max} > Cn^{-1/2}\sqrt{k\log(d/\delta)}\Big) < k\delta.$$

Proof. The proof follows from (B5) in Condition A.9 via an analogous argument to that of Lemma A.14, taking $t = C\sqrt{n^{-1}(k\log(d/\delta))}$. □

Lemma A.16. Suppose (B1)-(B5) of Condition A.9 are fulfilled. Then for any $k \ll d$ satisfying $k = o\big(((s \vee s_1)\log d)^{-1}\sqrt{n}\big)$, $|\overline{J}_{\theta|\gamma} - J^*_{v|-v}| = o_{\mathbb{P}}(1)$.

Proof. Recall that $J^*_{v|-v} = J^*_{v,v} - \boldsymbol{J}^*_{v,-v}J^{*-1}_{-v,-v}\boldsymbol{J}^*_{-v,v}$ and

$$\overline{J}_{v|-v} = \frac{1}{k}\sum_{j=1}^{k}\big(\nabla_{v,v}\,\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v}) - \overline{w}^T\nabla^2_{-v,v}\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v}),\ \text{so}$$

$$\big|\overline{J}_{v|-v} - J^*_{v|-v}\big| = \underbrace{\big|\frac{1}{k}\sum_{j=1}^{k}\nabla_{v,v}\,\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v}) - J^*_{v,v}\big|}_{\Delta_1} + \underbrace{\big|\overline{\boldsymbol{w}}^T\big(\frac{1}{k}\sum_{j=1}^{k}\nabla^2_{-v,v}\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v}) - \boldsymbol{w}^{*T}\boldsymbol{J}^*_{-v,v}\big)\big|}_{\Delta_2}.$$

Let $\widetilde{\boldsymbol{\beta}} = (\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v})$ and note that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$ satisfies the clause in (B5) of Condition A.9 by Lemma A.14 when $k = o\big(((s \vee s_1)\log d)^{-1}\sqrt{n}\big)$. Hence $\Delta_1 = o_{\mathbb{P}}(1)$ by Lemma A.15.

$$\begin{aligned}\Delta_2 \;\le\; &\underbrace{\Big|(\overline{\boldsymbol{w}} - \boldsymbol{w}^*)^T\Big(\frac{1}{k}\sum_{j=1}^{k}\nabla^2_{-v,v}\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v}) - \boldsymbol{J}^*_{-v,v}\Big)\Big|}_{\Delta_{21}} \\ &+ \underbrace{\big|(\overline{\boldsymbol{w}} - \boldsymbol{w}^*)^T\boldsymbol{J}^*_{-v,v}\big|}_{\Delta_{22}} + \underbrace{\Big|\boldsymbol{w}^{*T}\Big(\frac{1}{k}\sum_{j=1}^{k}\nabla^2_{-v,v}\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v}) - \boldsymbol{J}^*_{-v,v}\Big)\Big|}_{\Delta_{23}}.\end{aligned}$$

By the fact that $\|J^*\|_{\max} < \infty$ and $\|\boldsymbol{w}^*\|_1 \le Cs_1$ by (B1) of Condition A.9, an application of Lemmas A.14 and A.15 delivers

$$\begin{aligned}\Delta_{21} &\le \|\overline{\boldsymbol{w}} - \boldsymbol{w}^*\|_1\big\|\frac{1}{k}\sum_{j=1}^{k}\nabla^2_{-v,v}\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v}) - \boldsymbol{J}^*_{-v,v}\big\|_\infty = o_{\mathbb{P}}(1),\\ \Delta_{22} &\le \|\overline{\boldsymbol{w}} - \boldsymbol{w}^*\|_1\|\boldsymbol{J}^*_{-v,v}\|_\infty = o_{\mathbb{P}}(1),\\ \Delta_{23} &\le \big\|\frac{1}{k}\sum_{j=1}^{k}\nabla^2_{-v,v}\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v}) - \boldsymbol{J}^*_{-v,v}\big\|_\infty\|\boldsymbol{w}^*\|_1 = o_{\mathbb{P}}(1)\end{aligned}$$

for $k = o\big((s_1\log d)^{-1}n\big)$, a fortiori for $k = o\big(((s\vee s_1)\log d)^{-1}\sqrt{n}\big)$. Hence $\big|\overline{J}_{v|-v} - J^*_{v|-v}\big| = o_{\mathbb{P}}(1)$. □

B Auxiliary Lemmas for Estimation

In this section, we provide the proofs of the technical lemmas for the divide and conquer estimation.

Lemma B.1. Suppose X is a $n \times d$ matrix that has independent sub-gaussian rows $\{\boldsymbol{X}_i\}_{i=1}^n$. Denote $\mathbb{E}(\boldsymbol{X}_i \boldsymbol{X}_i^T)$ by Σ, then we have

$$\mathbb{P}\left(\|\frac{1}{n} X^T X - \Sigma_X\|_2 \geq (\delta \vee \delta^2) \right) \leq \exp(-c_1 t^2),$$

where $t \geq 0$, $\delta = C_1\sqrt{d/n} + t/\sqrt{n}$ and C_1 and c_1 are both constants depending only on $\|\boldsymbol{X}_i\|_{\psi_2}$.

Proof. See Vershynin (2010). □

Lemma B.2. (Bernstein-type inequality) Let $X_1, \ldots, X_n$ be independent centered sub-exponential random variables, and $M = \max_{1 \leq i \leq n} \|X_i\|_{\psi_1}$. Then for every $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and every $t \geq 0$, we have

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \geq t \right) \leq \exp\left(-C_2 \min\left(\frac{t^2}{M^2 \|a\|_2^2}, \frac{t}{M\|a\|_\infty} \right) \right).$$

Proof. See Vershynin (2010). □

Lemma B.3. Suppose X is a $n \times d$ matrix that has independent sub-gaussian rows $\{\mathbf{x}_i\}_{i=1}^n$. If $\lambda_{\max}(\Sigma) \leq C_{\max}$ and $d \ll n$, then for all $M > C_{\max}$, there exists a constant $c > 0$ such that when n and d are sufficiently large,

$$\mathbb{P}\left(\left\| \frac{1}{n} X^T X \right\|_2 \geq M \right) \leq \exp(-cn).$$

Proof. Apply Lemma B.1 with $t = \sqrt{cn/c_1}$, where $(\sqrt{c/c_1} \vee c/c_1) < M - C_{\max}$, and it follows that

$$\mathbb{P}\left(\left\| \frac{1}{n} X^T X - \Sigma \right\|_2 \geq (\delta \vee \delta^2) \right) \leq \exp(-cn).$$

Since $d \ll n$, we obtain $(\delta \vee \delta^2) \to \sqrt{c/c_1}$, which completes the proof. □

Lemma B.4. Suppose X is a $n \times d$ matrix that has independent sub-gaussian rows $\{\boldsymbol{X}_i\}_{i=1}^n$. $\mathbb{E}\boldsymbol{X}_i = \mathbf{0}$, $\lambda_{\min}(\Sigma) \geq C_{\min} > 0$ and $d \ll n$. For all $m < C_{\min}$, there exists a constant $c > 0$ such that when n and d are sufficiently large,

$$\mathbb{P}\left(\left\| \left(\frac{1}{n} X^T X \right)^{-1} \right\|_2 \geq \frac{1}{m} \right) = \mathbb{P}\left(\lambda_{\min}\left(\frac{1}{n} X^T X \right) \leq m \right) \leq \exp(-cn).$$

Proof. It is easy to check the following inequality. For any two symmetric and semi-definite $d \times d$ matrices A and B, we have

$$\lambda_{\min}(A) \geq \lambda_{\min}(B) - \|A - B\|_2 \ ,$$

because for any vector $\mathbf{x}$ satisfying $\|\mathbf{x}\|_2 = 1$, we have $\|A\boldsymbol{x}\|_2 = \|B\boldsymbol{x} + (A-B)\boldsymbol{x}\|_2 \geq \|B\boldsymbol{x}\|_2 - \|(A-B)\boldsymbol{x}\|_2 \geq \lambda_{\min}(B) - \|A - B\|_2$. Then it follows that

$$\begin{aligned} \mathbb{P}\left(\left\| \left(\frac{1}{n} X^T X \right)^{-1} \right\|_2 \geq \frac{1}{m} \right) &= \mathbb{P}\left(\lambda_{\min}\left(\frac{1}{n} X^T X \right) \leq m \right) \leq \mathbb{P}\left(\lambda_{\min}(\Sigma) - \left\| \frac{1}{n} X^T X - \Sigma \right\|_2 \geq m \right) \\ &\leq \mathbb{P}\left(\left\| \frac{1}{n} X^T X - \Sigma_X \right\|_2 \geq C_{\min} - m \right) \leq \exp(-cn), \end{aligned}$$

where c satisfies $(\sqrt{c/c_1} \vee c/c_1) < C_{\min} - m$ and the last inequality is an application of Lemma B.1 with $t = \sqrt{cn/c_1}$. □

Lemma B.5. (Hoeffding-type Inequality). Let $X_1,\ldots,X_n$ be independent centered sub-gaussian random variables, and let $K = \max_i \|X_i\|_{\psi_2}$. Then for every $a = (a_1,\ldots,a_n) \in \mathbb{R}^n$ and every $t > 0$, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| \geq t\right) \leq e \cdot \exp\left(-\frac{ct^2}{K^2\|a\|_2^2}\right).$$

Lemma B.6. (Sub-exponential is sub-gaussian squared). A random variable X is a sub-gaussian if and only if X^2 is sub-exponential. Moreover,

$$\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2.$$

Lemma B.7. Let $X_1,\ldots,X_n$ be independent centered sub-gaussian random variables. Let $\kappa = \max_i \|X_i\|_{\psi_2}$ and $\sigma^2 = \max_i \mathbb{E}X_i^2$. Suppose $\sigma^2 > 1$, then we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 > 2\sigma^2\right) \leq \exp\left(-C_2\frac{\sigma^2 n}{\kappa^2}\right).$$

Proof. Combining Lemma B.2 and Lemma B.6 yields the result. □

Lemma B.8. Following the same notation as in the beginning of Proof of Theorem 4.10,

$$\mathbb{P}\left(\left\{\|\frac{1}{k}\sum_{j=1}^k (X^{(j)}D_1^{(j)})^T\boldsymbol{\varepsilon}^{(j)}/n_k\|_2 > t/2\right\} \cap \mathcal{E}_0\right) \leq \exp\left(d\log(6) - \frac{t^2 C_{\min}^3 n}{32C_3 s_1^2(\delta_1 \vee \delta_1^2)^2}\right)$$

and

$$\mathbb{P}\left(\left\{\|(XD_2)^T\boldsymbol{\varepsilon}/n\|_2 > t/2\right\} \cap \mathcal{E}\right) \leq \exp\left(d\log(6) - \frac{t^2 C_{\min}^3 n}{32C_3 s_1^2(\delta_2 \vee \delta_2^2)^2}\right).$$

Proof.

$$\begin{aligned}\mathbb{E}\left(\exp\left(\lambda(D_1^{(j)}\mathbf{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k)\right) \mid X^{(j)}\right) &= \prod_{i=1}^{n_k}\mathbb{E}\left(\exp\left((\lambda \boldsymbol{X}_i^{(j)}/n_k)^T(D^{(j)}\mathbf{v})\varepsilon_i\right) \mid X^{(j)}\right) \\ &\leq \exp\left(C_3\lambda^2 s_1^2\sum_{i=1}^n (A_i^{(j)})^2/n_k^2\right),\end{aligned} \tag{B.1}$$

$$\begin{aligned}\mathbb{E}\left(\exp\left(\lambda(D_2\mathbf{v})^T(X^T\boldsymbol{\varepsilon}/n)\right) \mid X\right) &= \prod_{i=1}^{N}\mathbb{E}\left(\exp\left((\lambda \boldsymbol{X}_i/N)^T(D_2\mathbf{v})\varepsilon_i\right) \mid X\right) \\ &\leq \exp\left(C_3\lambda^2 s_1^2\sum_{i=1}^N A_i^2/n^2\right),\end{aligned} \tag{B.2}$$

where we write $A_i^{(j)}$ and A_i in place of $(\boldsymbol{X}_i^{(j)})^T D_1^{(j)}\mathbf{v}$ and $(\boldsymbol{X}_i)^T D_2\mathbf{v}$ respectively C_3 is an absolute constant, and the last inequality holds because ε_i are sub-gaussian. Next we provide an upper bound on $\sum_{i=1}^{n_k}(A_i^{(j)})^2$ and $\sum_{i=1}^{n} A_i^2$. Note that

$$\begin{aligned}\sum_{i=1}^{n}(A_i^{(j)})^2 &= \mathbf{v}^T D_1^{(j)} X^T X D_1^{(j)}\mathbf{v} = \mathbf{v}^T((S_X^{(j)})^{-1} - (\Sigma)^{-1})n_k S_X^{(j)}((S_X^{(j)})^{-1} - (\Sigma)^{-1})\mathbf{v} \\ &= n_k \mathbf{v}^T\Sigma^{-1}(\Sigma - S_X^{(j)})(S_X^{(j)})^{-1}(\Sigma - S_X^{(j)})\Sigma^{-1}\mathbf{v},\end{aligned}$$

and similarly,

$$\sum_{i=1}^{n} A_i^2 = n\mathbf{v}^T\Sigma^{-1}(\Sigma - S_X)(S_X)^{-1}(\Sigma - S_X)\Sigma^{-1}\mathbf{v}.$$

For any $\tau \in \mathbb{R}$, define the event $\mathcal{E}^{(j)} = \{\|(S_X^{(j)})^{-1}\|_2 \le 2/C_{\min}\} \cap \{\|S_X^{(j)} - \Sigma\|_2 \le (\delta_1 \vee \delta_1^2)\}$ for all $j = 1, \ldots, k$, where $\delta_1 = C_1\sqrt{d/n_k} + \tau/\sqrt{n_k}$, and the event $\mathcal{E} = \{\|(S_X)^{-1}\|_2 \le 2/C_{\min}\} \cap \{\|S_X - \Sigma\|_2 < (\delta_2 \vee \delta_2^2)\}$, where $\delta_2 = C_1\sqrt{d/n} + \tau/\sqrt{n}$. On $\mathcal{E}^{(j)}$ and $\mathcal{E}$, we have respectively

$$\sum_{i=1}^{n_k}(A_i^{(j)})^2 \le \frac{2n_k}{C_{\min}^3}(\delta_1 \vee \delta_1^2)^2 \text{ and } \sum_{i=1}^{n} A_i^2 \le \frac{2n}{C_{\min}^3}(\delta_2 \vee \delta_2^2)^2.$$

Therefore from Equation (B.1) and (B.2) we obtain

$$\mathbb{E}\left(\exp(\lambda(D_1^{(j)}\mathbf{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k))\,\mathbb{1}\{\mathcal{E}^{(j)}\}\right) \le \exp\left(\frac{2C_3\lambda^2 s_1^2}{C_{\min}^3 n_k}(\delta_1 \vee \delta_1^2)^2\right)$$

and

$$\mathbb{E}\left(\exp(\lambda(D_2\mathbf{v})^T(X^T\boldsymbol{\varepsilon}/n))\,\mathbb{1}\{\mathcal{E}\}\right) \le \exp\left(\frac{2C_3\lambda^2 s_1^2}{C_{\min}^3 N}(\delta_2 \vee \delta_2^2)^2\right).$$

In addition, according to Lemma B.1 and B.4, the probability of both $(\mathcal{E}^{(j)})^c$ and $\mathcal{E}^c$ are very small. More specifically,

$$\mathbb{P}(\mathcal{E}^c) \le \exp(-cn) + \exp(-c_1\tau^2) \text{ and } \mathbb{P}((\mathcal{E}^{(j)})^c) \le \exp(-cn/k) + \exp(-c_1\tau^2).$$

Let $\mathcal{E}_0 := \bigcap_{j=1}^{k} \mathcal{E}^{(j)}$. An application of the Chernoff bound trick leads us to the following inequality.

$$\begin{aligned}&\mathbb{P}\left(\left\{\frac{1}{k}\sum_{j=1}^{k}(D_1^{(j)}\mathbf{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)})/n_k > t/2\right\} \cap \mathcal{E}_0\right) \\ &\le \exp(-\lambda t/2)\prod_{j=1}^{k}\mathbb{E}\left(\exp\left(\frac{\lambda}{k}(D_1^{(j)}\boldsymbol{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k)\right)\mathbb{1}\{\mathcal{E}^{(j)}\}\right) \\ &\le \exp\left(-\lambda t/2 + \frac{2C_3\lambda^2 s_1^2}{C_{\min}^3 n}(\delta_1 \vee \delta_1^2)^2\right).\end{aligned}$$

Minimize the right hand side by λ, then we have

$$\mathbb{P}\left(\left\{\frac{1}{k}\sum_{j=1}^{k}(D_1^{(j)}\boldsymbol{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)})/n_k > t/2\right\}\cap\mathcal{E}_0\right) \leq \exp\left(-\frac{t^2C_{\min}^3 n}{32C_3 s_1^2(\delta_1\vee\delta_1^2)^2}\right).$$

Consider the $1/2-$net of $\mathbb{R}^p$, denoted by $\mathcal{N}(1/2)$. Again it is known that $|N(1/2)| < 6^p$. Using the maximal inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left\|\frac{1}{k}\sum_{j=1}^{k}(X^{(j)}D_1^{(j)})^T\boldsymbol{\varepsilon}^{(j)}/n_k\right\|_2 > t/2\right\}\cap\mathcal{E}_0\right)\\
&= \sup_{\|\mathbf{v}\|_2=1}\mathbb{P}\left(\left\{\frac{1}{k}\sum_{j=1}^{k}(D_1^{(j)}\boldsymbol{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)})/n_k > t/2\right\}\cap\mathcal{E}_0\right)\\
&\leq \sup_{\mathbf{v}\in N(1/2)}\mathbb{P}\left(\left\{\frac{1}{k}\sum_{j=1}^{k}(D_1^{(j)}\boldsymbol{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)})/n_k > t/4\right\}\cap\mathcal{E}_0\right)\\
&\leq \exp\left(d\log(6) - \frac{t^2C_{\min}^3 n}{32C_3 s_1^2(\delta_1\vee\delta_1^2)^2}\right).
\end{aligned}$$

Proceeding in an analogous fashion, we obtain

$$\mathbb{P}\left(\left\{\|(XD_2)^T\boldsymbol{\varepsilon}/n\|_2 > t/2\right\}\cap\mathcal{E}\right) \leq \exp\left(d\log(6) - \frac{t^2C_{\min}^3 n}{32C_3 s_1^2(\delta_2\vee\delta_2^2)^2}\right).$$

□

Lemma B.9. Following the same notation as in the proof of Theorem 4.13,

$$\mathbb{P}(\{\|\boldsymbol{B}\|_2 > t_1\}\cap\mathcal{A}) \leq 2\exp\left(d\log(6) - \frac{C_{\min}^4 L_{\min}^2 n t_1^2}{128\phi U_2 C_{\max}(\delta_1\vee\delta_1^2)^2}\right).$$

Proof. By Lemma A.2, for any $\lambda\in\mathbb{R}$ and $\mathbf{v}$ such that $\|\mathbf{v}\|_2 = 1$, we have

$$\begin{aligned}
\mathbb{E}\left(\exp(\lambda(D_1^{(j)}\mathbf{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k)) \mid X^{(j)}\right) &= \prod_{i=1}^{n_k}\mathbb{E}\left(\exp((\lambda\boldsymbol{X}_i^{(j)}/n_k)^T(D^{(j)}\mathbf{v})\varepsilon_i) \mid X^{(j)}\right)\\
&\leq \exp\left(\phi U\lambda^2\sum_{i=1}^{n_k}(A_i^{(j)})^2/n_k^2\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left(\exp(\lambda(D_2\mathbf{v})^T(X^T\boldsymbol{\varepsilon}/n)) \mid X\right) &= \prod_{i=1}^{n}\mathbb{E}\left(\exp((\lambda\boldsymbol{X}_i/n)^T(D_2\mathbf{v})\varepsilon_i) \mid X\right)\\
&\leq \exp\left(\phi U\lambda^2\sum_{i=1}^{n}A_i^2/n^2\right),
\end{aligned}$$

where we write $A_i^{(j)}$ and A_i in place of $(X_i^{(j)})^T D_1^{(j)}\mathbf{v}$ and $(X_i)^T D_2\mathbf{v}$ respectively. Next we give a upper bound on $\sum_{i=1}^{n_k}(A_i^{(j)})^2$ and $\sum_{i=1}^{n} A_i^2$. Note that

$$\begin{aligned}\sum_{i=1}^{n_k}(A_i^{(j)})^2 &= \mathbf{v}^T D_1^{(j)} X^T X D_1^{(j)}\mathbf{v} \\ &= \mathbf{v}^T((S^{(j)})^{-1} - \Sigma^{-1})nS_X((S^{(j)})^{-1} - \Sigma^{-1})\mathbf{v} \\ &= n\mathbf{v}^T\Sigma^{-1}(\Sigma - S^{(j)})(S^{(j)})^{-1}S_X^{(j)}(S^{(j)})^{-1}(\Sigma - S^{(j)})\Sigma^{-1}\mathbf{v}.\end{aligned}$$

Similarly,

$$\sum_{i=1}^{n} A_i^2 = n\mathbf{v}^T\Sigma^{-1}(\Sigma - S)S^{-1}S_X S^{-1}(\Sigma - S)\Sigma^{-1}\mathbf{v}.$$

On $\mathcal{E}^{(j)}$ and $\mathcal{E}$, we have respectively

$$\sum_{i=1}^{n_k}(A_i^{(j)})^2 \le \frac{8C_{\max}n_k}{C_{\min}^4 L_{\min}^2}(\delta_1 \vee \delta_1^2)^2 \text{ and } \sum_{i=1}^{n} A_i^2 \le \frac{8C_{\max}n}{C_{\min}^4 L_{\min}^2}(\delta_2 \vee \delta_2^2)^2.$$

Then it follows that

$$\mathbb{E}\left(\exp(\lambda(D_1^{(j)}\boldsymbol{v})^T(X^{(j)T}\boldsymbol{\varepsilon}^{(j)}/n_k))\,\mathbb{1}\{\mathcal{E}^{(j)}\}\right) \le \exp\left(\frac{8\phi U C_{\max}\lambda^2}{C_{\min}^4 L_{\min}^2 n_k}(\delta_1 \vee \delta_1^2)^2\right)$$

and

$$\mathbb{E}\left(\exp(\lambda(D_2\mathbf{v})^T(X^T\boldsymbol{\varepsilon}/n))\,\mathbb{1}\{\mathcal{E}\}\right) \le \exp\left(\frac{8\phi U C_{\max}\lambda^2}{C_{\min}^4 L_{\min}^2 n}(\delta_2 \vee \delta_2^2)^2\right).$$

Now we follow exactly the same steps as in the OLS part. Denote $\cap_{j=1}^k \mathcal{E}_j$ by $\mathcal{E}_0$. An application of the Chernoff bound technique and the maximal inequality leads us to the following inequality.

$$\mathbb{P}\left(\left\{\|\frac{1}{k}\sum_{j=1}^{k}(X^{(j)}D_1^{(j)})^T\boldsymbol{\varepsilon}^{(j)}/n_k\|_2 > t/2\right\} \cap \mathcal{E}_0\right) \le \exp\left(d\log(6) - \frac{C_{\min}^4 L_{\min}^2 n t^2}{128\phi U_2 C_{\max}(\delta_1 \vee \delta_1^2)^2}\right)$$

and

$$\mathbb{P}\left(\left\{\|(XD_2)^T\boldsymbol{\varepsilon}/n\|_2 > t/2\right\} \cap \mathcal{E}\right) \le \exp\left(d\log(6) - \frac{C_{\min}^4 L_{\min}^2 n t^2}{128\phi U_2 C_{\max}(\delta_2 \vee \delta_2^2)^2}\right).$$

We have thus derived an upper bound for $\|\boldsymbol{B}\|_2$ that holds with high probability. Specifically,

$$\begin{aligned}\mathbb{P}(\{\|\boldsymbol{B}\|_2 > t_1\} \cap \mathcal{A}) \le\ & \mathbb{P}\left(\left\{\|\frac{1}{k}\sum_{j=1}^{k}(X^{(j)}D_1^{(j)})^T\boldsymbol{\varepsilon}^{(j)}/n_k\|_2 > \frac{t_1}{2}\right\} \cap \mathcal{E}_0\right) \\ &+ \mathbb{P}\left(\left\{\|(XD_2)^T\boldsymbol{\varepsilon}/n\|_2 > \frac{t_1}{2}\right\} \cap \mathcal{E}\right) \le 2\exp\left(d\log(6) - \frac{C_{\min}^4 L_{\min}^2 n t_1^2}{128\phi U_2 C_{\max}(\delta_1 \vee \delta_1^2)^2}\right).\end{aligned}$$

□

Lemma B.10. Under Condition 3.6, for $\tau \le L_{\min}/(8MC_{\max}U_3\sqrt{d})$ and sufficiently large n and d we have

$$\mathbb{P}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > \tau) \le \exp\left(d\log 6 - \frac{nC_{\min}^2 L_{\min}^2 \tau^2}{2^{11} C_{\max} U_2 \phi}\right) + 2\exp(-cn).$$

Proof. The notation is that introduced in the proof of Theorem 4.13. We further define $\Sigma(\boldsymbol{\beta}) := \mathbb{E}(b''(X^T\boldsymbol{\beta})XX^T)$ as well as the event $\mathcal{H} := \{\ell_n(\boldsymbol{\beta}^*) > \max_{\boldsymbol{\beta}\in\partial\mathcal{B}_\tau} \ell_n(\boldsymbol{\beta})\}$, where $\mathcal{B}_\tau = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le \tau\}$. Note that as long as the event $\mathcal{H}$ holds, the MLE falls in $\mathcal{B}_\tau$, therefore the proof strategy involves showing that $\mathbb{P}(\mathcal{H})$ approaches 1 at certain rate. By the Taylor expansion,

$$\begin{aligned}
\ell_n(\boldsymbol{\beta}) - \ell_n(\boldsymbol{\beta}^*) &= (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \boldsymbol{v} - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T S(\widetilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\
&= (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \boldsymbol{v} - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T S(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T (S(\widetilde{\boldsymbol{\beta}}) - S(\boldsymbol{\beta}^*))(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\
&= A_1 + A_2,
\end{aligned}$$

where $S(\boldsymbol{\beta}) = (1/n)X^T D(X\boldsymbol{\beta})X$, $\widetilde{\boldsymbol{\beta}}$ is some vector between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$, $\boldsymbol{v} = (1/n)X^T(\boldsymbol{Y} - \boldsymbol{\mu}(X\boldsymbol{\beta}^*))$, $A_1 = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T\boldsymbol{v} - (1/2)(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T S(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ and $A_2 = -(1/2)(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T(S(\widetilde{\boldsymbol{\beta}}) - S(\boldsymbol{\beta}^*))(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$.

Define the event $\mathcal{E} := \{\lambda_{\min}[S(\boldsymbol{\beta}^*)] \ge L_{\min}/2\}$, where $L_{\min}$ is the same constant in Condition 3.6. Note that by Condition 3.6 (ii), $\sqrt{b''(\boldsymbol{X}_i^T\boldsymbol{\beta})}\boldsymbol{X}_i$ is a sub-gaussian random vector. Then by Condition 3.6 (iii) and Lemma B.4, for sufficiently large n and d we have $\mathbb{P}(\mathcal{E}^c) \le \exp(-cn)$. Therefore on the event $\mathcal{E}$,

$$A_1 \le \tau(\|\boldsymbol{v}\|_2 - \frac{L_{\min}}{4}\tau).$$

We next show that, under an appropriate choice of τ, $|A_2| < L_{\min}\tau^2/8$ with high probability. We first consider Condition 3.6 (ii). Define $\mathcal{F} := \{\|X^TX/n\|_2 \le 2C_{\max}\}$. By Lemma B.3, we have $\mathbb{P}(\mathcal{F}^c) \le \exp(-cn)$. By Lemma A.5, on the event $\mathcal{F}$, we have

$$\begin{aligned}
A_2 &\le \max_{1\le i\le n} |b''(\boldsymbol{X}_i^T\widetilde{\boldsymbol{\beta}}) - b''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)| C_{\max}\tau^2 \\
&\le MU_3\sqrt{d}\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \cdot C_{\max}\tau^2 \\
&\le MC_{\max}U_3\sqrt{d}\tau^3 \le \frac{L_{\min}\tau^2}{8},
\end{aligned}$$

where the last inequality holds if we choose $\tau \le L_{\min}/(8MC_{\max}U_3\sqrt{d})$. Now we obtain the following probabilistic upper bound on $\mathcal{H}^c$, which we later prove to be negligible.

$$\begin{aligned}
\mathbb{P}(\mathcal{H}^c) &\le \mathbb{P}(\mathcal{H}^c \cap \mathcal{E} \cap \mathcal{F}) + \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{F}^c) \\
&\le \mathbb{P}\left(\left\{\|\boldsymbol{v}\|_2 \ge \frac{L_{\min}\tau}{8}\right\} \cap \mathcal{E} \cap \mathcal{F}\right) + \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{F}^c).
\end{aligned} \tag{B.3}$$

Since each component of $\mathbf{v}$ is a weighted average of i.i.d. random variables, the effect of concentration tends to make $\|\mathbf{v}\|_2$ very small with large probability, which inspires us to study the moment generating function and apply the Chernoff bound technique. By Lemma A.2, for any constant

$\mathbf{u} \in \mathbb{R}^d$, $\|\mathbf{u}\|_2 = 1$ and let $a_i = \mathbf{u}^T \boldsymbol{X}_i$, then we have for any $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left(\exp(t\langle \mathbf{u}, \mathbf{v}\rangle) \mid X\right) &= \prod_{i=1}^{n} \mathbb{E}\left(\exp\left(\frac{t a_i}{n}(Y_i - \mu(\boldsymbol{X}_i^T \boldsymbol{\beta}))\right) \mid X\right) \\
&\leq \exp\left(\frac{\phi U_2 t^2}{2n^2} \sum_{i=1}^{n} a_i^2\right) \\
&= \exp\left(\frac{\phi U_2 t^2}{2n} \cdot \frac{\mathbf{u}^T X^T X \mathbf{u}}{n}\right).
\end{aligned}$$

It follows that

$$\mathbb{E} \exp(t\langle \mathbf{u}, \mathbf{v}\rangle \mathbb{1}\{\mathcal{E} \cap \mathcal{F}\}) \leq \exp\left(\frac{\phi C_{\max} U_2 t^2}{2n}\right).$$

By the Chernoff bound technique, we obtain

$$\mathbb{P}(\{\langle \mathbf{u}, \mathbf{v}\rangle > \varepsilon\} \cap \mathcal{E} \cap \mathcal{F}) \leq \exp\left(-\frac{n\varepsilon^2}{8 C_{\max} U_2 \phi}\right).$$

Consider a $1/2-$net of $\mathbb{R}^d$, denoted by $N(1/2)$. Since

$$\|\mathbf{v}\|_2 = \max_{\|\mathbf{u}\|_2 = 1} \langle \mathbf{u}, \mathbf{v}\rangle \leq 2 \max_{\mathbf{u} \in N(1/2)} \langle \mathbf{u}, \mathbf{v}\rangle,$$

it follows that

$$\begin{aligned}
\mathbb{P}(\{\|\mathbf{v}\|_2 > \frac{L_{\min}\tau}{8}\} \cap \mathcal{E} \cap \mathcal{F}) &\leq \mathbb{P}\left(\left\{\max_{\mathbf{u} \in N(1/2)} \langle \mathbf{u}, \mathbf{v}\rangle > \frac{L_{\min}\tau}{16}\right\} \cap \mathcal{E} \cap \mathcal{F}\right) \\
&\leq 6^d \exp\left(-\frac{n L_{\min}^2 \tau^2}{2^{10} \phi C_{\max} U_2}\right) \\
&= \exp\left(d \log 6 - \frac{n C_{\min}^2 L_{\min}^2 \tau^2}{2^{11} C_{\max} U_2 \phi}\right).
\end{aligned}$$

Finally combining the result above with Equation (B.3) delivers the conclusion. □

Remark B.11. Simple calculation shows that when $d = o(\sqrt{n})$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}(\sqrt{d/n})$. When d is a fixed constant, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}(\sqrt{1/n})$.