Journal of Machine Learning Research 23 (2022) 1-77 Submitted 6/21; Revised 5/22; Published 6/22

Distributed Bootstrap for Simultaneous Inference Under High Dimensionality

Yang Yu YUYANG930930@GMAIL.COM
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

Shih-Kang Chao SKCHAO74@GMAIL.COM
Department of Statistics
University of Missouri
Columbia, MO 65211, USA

Guang Cheng[*] GUANGCHENG@UCLA.EDU
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095, USA

Editor: Victor Chernozhukov

Abstract

We propose a distributed bootstrap method for simultaneous inference on high-dimensional massive data that are stored and processed with many machines. The method produces an ℓ_∞-norm confidence region based on a communication-efficient de-biased lasso, and we propose an efficient cross-validation approach to tune the method at every iteration. We theoretically prove a lower bound on the number of communication rounds $\tau_{\min}$ that warrants the statistical accuracy and efficiency. Furthermore, $\tau_{\min}$ only increases logarithmically with the number of workers and the intrinsic dimensionality, while nearly invariant to the nominal dimensionality. We test our theory by extensive simulation studies, and a variable screening task on a semi-synthetic dataset based on the US Airline On-Time Performance dataset. The code to reproduce the numerical results is available in Supplementary Material.

Keywords: Distributed Learning, High-dimensional Inference, Multiplier Bootstrap, Simultaneous Inference, De-biased Lasso

1. Introduction

Modern massive datasets with enormous sample size and tremendous dimensionality are usually impossible to be processed with a single machine. For remedy, a master-worker architecture is often adopted, e.g., Hadoop (Singh and Kaur, 2014), which operates on a cluster of nodes for data storage and processing, where the master node also contains a portion of the data; see Figure 1. An inherent problem of this architecture is that inter-node communication can be over a thousand times slower than intra-node computation due to the inter-node communication protocol, which unfortunately always comes with significant

[*]. Part of this manuscript was completed while Cheng was at Purdue.

overhead (Lan et al., 2018; Fan et al., 2019a). Hence, communication efficiency is usually a top concern for algorithm development in distributed learning.

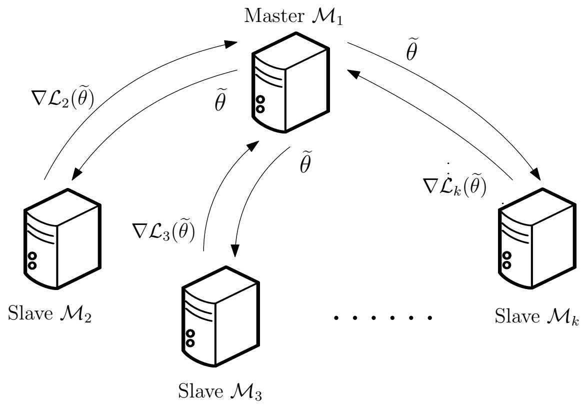

Figure 1: Master-worker architecture for storing and processing distributed data.

Classical statistical methods are usually not communication-efficient as some of them require hundreds or even thousands passes over the entire dataset. In the last few years, active research has greatly advanced our ability to perform distributed statistical optimization and inference in, e.g., maximum likelihood estimation (Zhang et al., 2012; Li et al., 2013; Chen and Xie, 2014; Battey et al., 2018; Jordan et al., 2019; Huang and Huo, 2019; Chen et al., 2018; Zhu et al., 2020), Lasso (Lee et al., 2017; Wang et al., 2017; Wang and Zhang, 2017), partially linear models (Zhao et al., 2016), nonstandard regression (Shi et al., 2018; Banerjee et al., 2019), quantile regression (Volgushev et al., 2019; Chen et al., 2019), principal component analysis (Fan et al., 2019b; Chen et al., 2020), just to name a few. However, solutions for many other problems in the distributed framework, for example the statistical inference for high-dimensional models, are still elusive.

Simultaneous inference for high-dimensional statistical models has been widely considered in many applications where datasets can be handled with a standalone computer (Cai and Sun, 2017), and many recent papers focus on bootstrap as an effective way to implement simultaneous inference (Dezeure et al., 2017; Zhang and Cheng, 2017; Belloni et al., 2018, 2019; Yu et al., 2020a). These existing methods typically use the well-celebrated de-biased Lasso (van de Geer et al., 2014; Zhang and Zhang, 2014; Javanmard and Montanari, 2014a,b), where the de-biased score results from the KKT condition of the Lasso optimization problem. However, de-biased Lasso is not directly applicable in a distributed computational framework. For one thing, the implementation of de-biased Lasso requires expensive subroutines such as nodewise Lasso (van de Geer et al., 2014), which has to be replaced by a more communication-efficient method. For another, the quality of the de-biased score, which is essential to the validity of the bootstrap, is generally worse in a distributed computational framework than that in a centralized computational framework. In particular, it is heavily biased so the asymptotic normality fails. However, it can possibly be improved with sufficient rounds of communication between the master and worker nodes. The bootstrap validity therefore critically hinges on the interplay between the dimensionality of the model and the sparsity level, as well as the rounds of communication, the number of worker nodes and the size of local sample that are specific to the distributed computational framework.

In this paper, we tackle the challenges discussed above and propose a communication-efficient simultaneous inference method for high-dimensional models. The main component at the core of our method is a novel way to improve the quality of the de-biased score with a carefully selected number of rounds of communication while relaxing the constraint on the number of machines. Our method is motivated by Wang et al. (2017), who proposed an iterative procedure for computing the estimator but no statistical inference was provided. Note that the de-biased Lasso has been applied by Lee et al. (2017) to obtain a communication-efficient $\sqrt{N}$-consistent estimator, but their method restricts the number of worker nodes to be less than the local sample size. Next, we apply communicate-efficient multiplier bootstrap methods `k-grad` and `n+k-1-grad`, which are originally proposed in Yu et al. (2020b) for low dimensional models. These bootstrap methods prevent repeatedly refitting the models and relax the constraint on the number of machines that plague the methods proposed earlier (Kleiner et al., 2014; Sengupta et al., 2016). A key challenge in implementation is that cross-validation, which is a popular method for selecting tuning parameters, usually requires multiple passes of the entire dataset and is typically inefficient in the distributed computational framework. We propose a new cross-validation that only requires the master node for implementation without needing to communicate with the worker nodes.

Our theoretical study focuses on the explicit lower bounds on the rounds of communication that warrant the validity of the bootstrap method for high-dimensional generalized linear models; see Section 3.1 for an overview. In short, the greater the number of worker nodes and/or the intrinsic dimensionality, the greater the rounds of communication required for the bootstrap validity. The bootstrap validity and efficiency are corroborated by an extensive simulation study.

We further demonstrate the merit of our method on variable screening with a semi-synthetic dataset, based on the large-scale US Airline On-Time Performance dataset. By performing a pilot study on an independently sampled subset of data, we take four key explanatory variables for flight delay, which correspond to the dummy variables of the four years after the September 11 attacks. On another independently sampled subset of data, we combine the dummy variables of the four years with artificial high-dimensional spurious variables to create a design matrix. We perform our method on this artificial dataset, and find that the relevant variables are correctly identified as the number of iteration increases. In particular, we visualize the effect of these four years by confidence intervals.

We go beyond our previous publication Yu et al. (2020b) in two major aspects: (1) In this paper we focus on high-dimensional models. In particular, the dimensionality of the model can exceed the sample size in each computing node. We handle high dimensionality using ℓ_1 penalization, and consider de-biased Lasso under the distributed computational framework. (2) We tune the ℓ_1 penalized problem with a carefully designed cross-validation method, which can be applied under distributed computational framework.

The rest of the paper is organized as follows. In Section 2, we introduce the problem formulation of distributed high-dimensional simultaneous inference and present the main bootstrap algorithm as well as the cross-validation algorithm for hyperparameter tuning. Theoretical guarantees of bootstrap validity for high-dimensional (generalized) linear models are provided in Section 3. Section 4 presents simulation results that corroborate our theoretical findings. Section 5 showcases an application on variable screening for high-dimensional logistic regression with a big real dataset using our new method. Finally, Section 6 concludes the paper. Technical details are in Appendices. The proofs of the theoretical results and the code to reproduce the numerical results are in Supplementary Material.

Notations. We denote the ℓ_p-norm ($p \geq 1$) of any vector $v = (v_1, \dots, v_n)$ by $\|v\|_p = (\sum_{i=1}^n |v_i|^p)^{1/p}$ and $\|v\|_\infty = \max_{1 \leq i \leq n} |v_i|$. The induced p-norm and the max-norm of any matrix $M \in \mathbb{R}^{m \times n}$ (with element M_{ij} at i-th row and j-th column) are denoted by $|||M|||_p = \sup_{x \in \mathbb{R}^n; \|x\|_p = 1} \|Mx\|_p$ and $|||M|||_{\max} = \max_{1 \leq i \leq m; 1 \leq j \leq n} |M_{i,j}|$. We write $a \lesssim b$ if $a = O(b)$, and $a \ll b$ if $a = o(b)$.

2. Distributed Bootstrap for High-Dimensional Simultaneous Inference

In this section, we introduce the distributed computational framework and present a novel bootstrap algorithm for high-dimensional simultaneous inference under this framework. A communication-efficient cross-validation method is proposed for tuning.

2.1 Distributed Computation Framework

Suppose data $\{Z_i\}_{i=1}^N$ are i.i.d., and $\mathcal{L}(\theta; Z)$ is a twice-differentiable convex loss function arising from a statistical model, where $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$. Suppose that the parameter of interest θ^* is the minimizer of an expected loss:

$$\theta^* = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathcal{L}^*(\theta), \text{ where } \mathcal{L}^*(\theta) := \mathbb{E}_Z[\mathcal{L}(\theta; Z)].$$

We consider a high-dimensional setting where $d > N$ is possible, and θ^* is sparse, i.e., the support of θ^* is small.

We consider a distributed computation framework, in which the entire data are stored distributedly in k machines, and each machine has data size n. Denote by $\{Z_{ij}\}_{i=1,\dots,n; j=1,\dots,k}$ the entire data, where Z_{ij} is i-th datum on the j-th machine $\mathcal{M}_j$, and $N = nk$. Without loss of generality, assume that the first machine $\mathcal{M}_1$ is the master node; see Figure 1. Define the local and global loss functions as

$$\begin{aligned} \text{global loss: } \mathcal{L}_N(\theta) &= \frac{1}{k} \sum_{j=1}^k \mathcal{L}_j(\theta), \quad \text{where} \\ \text{local loss: } \mathcal{L}_j(\theta) &= \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\theta; Z_{ij}), \quad j = 1, \dots, k. \end{aligned} \tag{1}$$

A great computational overhead occurs when the master and worker nodes communicate. In order to circumvent the overhead, the rounds of communications between the master and worker nodes should be minimized, and the algorithms with reduced communication overheads are "communication-efficient".

2.2 High-Dimensional Simultaneous Inference

In this paper, we focus on the simultaneous confidence region for θ^* in a high-dimensional model, which is one of the effective ways for variable selection and inference that are immune to the well-known multiple testing problem. In particular, given an estimator $\widehat{\theta}$ that is $\sqrt{N}$-consistent, simultaneous confidence intervals can be found with confidence α, for large $\alpha \in (0,1)$, by finding the quantile

$$c(\alpha) := \inf\{t \in \mathbb{R} : P(\widehat{T} \leq t) \geq \alpha\} \quad \text{where} \tag{2}$$

$$\widehat{T} := \left\| \sqrt{N}\big(\widehat{\theta} - \theta^*\big) \right\|_\infty. \tag{3}$$

where $\widehat{\theta}$ may be computed through the de-biased Lasso (van de Geer et al., 2014; Zhang and Zhang, 2014; Javanmard and Montanari, 2014a,b):

$$\widehat{\theta} = \widehat{\theta}_{Lasso} - \widehat{\Theta}\nabla\mathcal{L}_N(\widehat{\theta}_{Lasso}), \tag{4}$$

where

$$\widehat{\theta}_{Lasso} = \underset{\theta\in\mathbb{R}^d}{\arg\min}\, \mathcal{L}_N(\theta) + \lambda\|\theta\|_1$$

is the Lasso estimator with some hyperparameter $\lambda > 0$, $\widehat{\Theta}$ is a surrogate inverse Hessian matrix and $\mathcal{L}_N(\theta) = N^{-1}\sum_{i=1}^N \mathcal{L}(\theta; Z_i)$ is the empirical loss.

Implementing the simultaneous inference based on $\widehat{\theta}$ and $\widehat{T}$ in distributed computational framework inevitably faces some computational challenges. Firstly, computing $\widehat{\theta}$ usually involves some iterative optimization routines that can accumulate a large communication overhead without a careful engineering. Next, some bootstrap methods have been proposed for estimating $c(\alpha)$, e.g., the multiplier bootstrap (Zhang and Cheng, 2017), but they cannot be straightforwardly implemented within a distributed computational framework due to excessive resampling and communication. Even though some communication-efficient bootstrap methods have been proposed, e.g., Kleiner et al. (2014); Sengupta et al. (2016); Yu et al. (2020b), they either require a large number of machines or are inapplicable to high-dimensional models.

Because of the above-mentioned difficulties, inference based on $\widehat{T}$ is inapplicable in the distributed computational framework and is regarded as an "oracle" in this paper. Our goal is to provide a method that is communication-efficient while entertaining the same statistical accuracy as that based on the oracle $\widehat{T}$.

2.3 High-Dimensional Distributed Bootstrap

In order to adapt (4) to the distributed computational setting, we first need to find a good substitute $\widetilde{\theta}$ for $\widehat{\theta}_{Lasso}$ that is communication-efficient, while noting that standard algorithms for Lasso are not communication-efficient. Fortunately, $\widetilde{\theta}$ can be computed by the communication-efficient surrogate likelihood (CSL) algorithm with the ℓ_1-norm regularization (Wang et al., 2017; Jordan et al., 2019), which iteratively generates a sequence of estimators $\widetilde{\theta}^{(t)}$ with regularization parameters $\lambda^{(t)}$ at each iteration $t = 0, \ldots, \tau - 1$. See Remark 1 for model tuning and Lines 1-16 of Algorithm 1 for the exact implementation. Under regularity conditions, if t is sufficiently large, it is warranted that $\widetilde{\theta}$ is close to $\widehat{\theta}_{Lasso}$.

Typical algorithms for computing $\widehat{\Theta}$, e.g., the nodewise Lasso (van de Geer et al., 2014), cannot be extended straightforwardly to the distributed computational framework due to the same issue of communication inefficiency. We overcome this by performing the nodewise Lasso using only $\mathcal{M}_1$ without accessing the entire dataset. This simple approach does not sacrifice accuracy as long as a sufficient amount of communication brings $\widetilde{\theta}$ sufficiently close to θ^*.

Lastly, given the surrogate estimators $\widetilde{\theta}$ for $\widehat{\theta}_{Lasso}$ and $\widetilde{\Theta}$ for $\widehat{\Theta}$, we estimate the asymptotic quantile $c(\alpha)$ of $\widehat{T}$ by bootstrapping $\|\widetilde{\Theta}\sqrt{N}\nabla\mathcal{L}_N(\widetilde{\theta})\|_\infty$ using the `k-grad` or `n+k-1-grad` bootstrap originally proposed by Yu et al. (2020b) for low-dimensional models. However, the number of communication rounds between master and worker nodes has to be carefully fine-tuned for high-dimensional models. In particular, the `k-grad` algorithm computes

$$\overline{W}^{(b)} := \Bigg\| \underbrace{-\widetilde{\Theta}\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\epsilon_j^{(b)}\sqrt{n}(\mathbf{g}_j - \bar{\mathbf{g}})}_{=:\overline{A}} \Bigg\|_\infty, \tag{5}$$

where $\epsilon_j^{(b)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ independent from the data, $\mathbf{g}_j = \nabla\mathcal{L}_j(\widetilde{\theta})$ and $\bar{\mathbf{g}} = k^{-1}\sum_{j=1}^{k}\mathbf{g}_j$. However, it is known that `k-grad` does not perform well when k is small (Yu et al., 2020b). The improved algorithm `n+k-1-grad` computes

$$\widetilde{W}^{(b)} := \Bigg\| \underbrace{-\widetilde{\Theta}\frac{1}{\sqrt{n+k-1}}\Bigg(\sum_{i=1}^{n}\epsilon_{i1}^{(b)}(\mathbf{g}_{i1} - \bar{\mathbf{g}}) + \sum_{j=2}^{k}\epsilon_j^{(b)}\sqrt{n}(\mathbf{g}_j - \bar{\mathbf{g}})\Bigg)}_{=:\widetilde{A}} \Bigg\|_\infty, \tag{6}$$

where $\epsilon_{i1}^{(b)}$ and $\epsilon_j^{(b)}$ are i.i.d. $\mathcal{N}(0,1)$ multipliers, and $\mathbf{g}_{i1} = \nabla\mathcal{L}(\widetilde{\theta}; Z_{i1})$ is based on a single datum Z_{i1} in the master. The key advantage of `k-grad` or `n+k-1-grad` is that once the master has the gradients $\mathbf{g}_j$ from the worker nodes, the quantile of $\{\overline{W}^{(b)}\}_{b=1}^{B}$ can be computed in the master node only, without needing to communicate with worker nodes. See Algorithm 3 in the Appendix for the pseudocode of `k-grad` and `n+k-1-grad`.

Algorithm 1 presents the complete statistical inference procedure. There are two key innovative steps in Algorithm 1 that facilitate the statistical inference for high dimensional model with a big dataset. First, we introduce de-biased Lasso in distributed inference, which goes beyond high dimensional model estimation considered in Jordan et al. (2019); Wang et al. (2017). Second, we use nodewise Lasso to provide a sparse estimation of the high-dimensional inverse Hessian matrix instead of the empirical Hessian used in Yu et al. (2020b).

Algorithm 1 can achieve high computational efficiency due to two reasons. First, we initialize Algorithm 1 with a warm start. Namely, we warm start with the Lasso estimator estimated with dataset in the master node, which provides a good initializer. Second, because the nodewise Lasso is computationally expensive, we perform it only once at the very beginning and freeze it through the iterations of the algorithm without updating it.

The number of iterations τ in Algorithm 1 steers the trade-off between statistical accuracy and communication efficiency. In particular, a larger τ leads to a more accurate

Algorithm 1 k-grad/n+k-1-grad with de-biased ℓ_1-CSL estimator

```
Require: $\tau \geq 1$ rounds of communication; hyperparameters $\{\lambda^{(t)}\}_{t=0}^{\tau-1}$ , nodewise Lasso
    procedure Node$(\cdot,\cdot)$ with hyperparameters $\{\lambda_l\}_{l=1}^{d}$ (see Section B)
$\widetilde{\theta}^{(0)} \leftarrow \arg\min_\theta \mathcal{L}_1(\theta) + \lambda^{(0)}\|\theta\|_1$ at $\mathcal{M}_1$
Compute $\widetilde{\Theta}$ by running Node$(\nabla^2\mathcal{L}_1(\widetilde{\theta}^{(0)}), \{\lambda_l\}_{l=1}^{d})$ at $\mathcal{M}_1$
for $t = 1, \dots, \tau$ do
    Transmit $\widetilde{\theta}^{(t-1)}$ to $\{\mathcal{M}_j\}_{j=2}^{k}$
    Compute $\nabla\mathcal{L}_1(\widetilde{\theta}^{(t-1)})$ at $\mathcal{M}_1$
    for $j = 2, \dots, k$ do
        Compute $\nabla\mathcal{L}_j(\widetilde{\theta}^{(t-1)})$ at $\mathcal{M}_j$
        Transmit $\nabla\mathcal{L}_j(\widetilde{\theta}^{(t-1)})$ to $\mathcal{M}_1$
    end for
    $\nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) \leftarrow k^{-1}\sum_{j=1}^{k}\nabla\mathcal{L}_j(\widetilde{\theta}^{(t-1)})$ at $\mathcal{M}_1$
    if $t < \tau$ then
        $\widetilde{\theta}^{(t)} \leftarrow \arg\min_\theta \mathcal{L}_1(\theta) - \theta^\top\left(\nabla\mathcal{L}_1(\widetilde{\theta}^{(t-1)}) - \nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)})\right) + \lambda^{(t)}\|\theta\|_1$ at $\mathcal{M}_1$
    else
        $\widetilde{\theta}^{(\tau)} \leftarrow \widetilde{\theta}^{(\tau-1)} - \widetilde{\Theta}\nabla\mathcal{L}_N(\widetilde{\theta}^{(\tau-1)})$ at $\mathcal{M}_1$
    end if
end for
Run DistBoots('k-grad' or 'n+k-1-grad', $\widetilde{\theta} = \widetilde{\theta}^{(\tau)}, \{\mathbf{g}_j = \nabla\mathcal{L}_j(\widetilde{\theta}^{(\tau-1)})\}_{j=1}^{k}$,
                    $\widetilde{\Theta} = \widetilde{\Theta}$) at $\mathcal{M}_1$
```

coverage of the simultaneous confidence interval, but it also induces a higher communication cost. Therefore, studying the minimal τ that warrants the bootstrap accuracy is crucial, which is done in Section 3.

Remark 1 *Two groups of hyperparameters need to be chosen in Algorithm 1: $\{\lambda^{(t)}\}_{t=0}^{\tau-1}$ for regularization in CSL estimation, and $\{\lambda_l\}_{l=1}^{d}$ for regularization in nodewise Lasso (see Algorithm 4). In Section 2.4, we propose a cross-validation method for tuning $\{\lambda^{(t)}\}_{t=0}^{\tau-1}$. As to $\{\lambda_l\}_{l=1}^{d}$, while van de Geer et al. (2014) suggests to choose the same value for all λ_l by cross-validation, a potentially better way may be to allow λ_l to be different across l and select each λ_l via cross-validation for the corresponding nodewise Lasso, which is the approach we take for a distributed variable screening task in Section 5.*

Remark 2 *There exist other options than CSL for $\widetilde{\theta}$ such as the averaging de-biased estimator (Lee et al., 2017), but an additional round of communication may be needed to compute the local gradients. More importantly, their method may be inaccurate when $n < k$.*

2.4 Communication-Efficient Cross-Validation

We propose a communication-efficient cross-validation method for tuning the hyperparameters $\{\lambda^{(t)}\}_{t=0}^{\tau-1}$ in Algorithm 1. Wang et al. (2017) proposes to hold out a validation set on each node for selecting $\lambda^{(t)}$. However, this method requires fitting the model for each candidate value of $\lambda^{(t)}$, which uses the same communication cost as the complete CSL estimation procedure.

We propose a communication-efficient K-fold cross-validation method that chooses $\lambda^{(t)}$ for the CSL estimation at every iteration t. At iteration t, the master uses the gradients already communicated from the worker nodes at iteration $t-1$. Hence, the cross-validation needs only the master node, which circumvents costly communication between the master and the worker nodes.

Specifically, notice that the surrogate loss (see Line 12 in Algorithm 1) is constructed using n observations $\mathcal{Z} = \{Z_{i1}\}_{i=1}^{n}$ in the master node and $k-1$ gradients $\mathcal{G} = \{\nabla \mathcal{L}_j(\widetilde{\theta}^{(t-1)})\}_{j=2}^{k}$ from the worker nodes. We then create K (approximately) equal-size partitions to both $\mathcal{Z}$ and $\mathcal{G}$. The objective function for training is formed using $K-1$ partitions of $\mathcal{Z}$ and $\mathcal{G}$. In terms of the measure of fit, instead of computing the original likelihood or loss, we calculate the unregularized surrogate loss using the last partition of $\mathcal{Z}$ and $\mathcal{G}$, still in the master node. See Algorithm 2 for the pseudocode.

Algorithm 2 Distributed K-fold cross-validation for t-step CSL

```
Require: $(t-1)$-step CSL estimate $\widetilde{\theta}^{(t-1)}$, set $\Lambda$ of candidate values for $\lambda^{(t)}$, partition of master data $\mathcal{Z} = \bigcup_{q=1}^{K} \mathcal{Z}_q$, partition of worker gradients $\mathcal{G} = \bigcup_{q=1}^{K} \mathcal{G}_q$
for $q = 1, \ldots, K$ do
    $\mathcal{Z}_{train} \leftarrow \bigcup_{r \neq q} \mathcal{Z}_r$;    $\mathcal{Z}_{test} \leftarrow \mathcal{Z}_q$
    $\mathcal{G}_{train} \leftarrow \bigcup_{r \neq q} \mathcal{G}_r$;    $\mathcal{G}_{test} \leftarrow \mathcal{G}_q$
    $g_{1,train} \leftarrow \mathrm{Avg}_{Z \in \mathcal{Z}_{train}}\Big(\nabla \mathcal{L}(\widetilde{\theta}^{(t-1)}; Z)\Big)$;    $g_{1,test} \leftarrow \mathrm{Avg}_{Z \in \mathcal{Z}_{test}}\Big(\nabla \mathcal{L}(\widetilde{\theta}^{(t-1)}; Z)\Big)$
    $\bar{g}_{train} \leftarrow \mathrm{Avg}_{g \in \{g_{1,train}\} \cup \mathcal{G}_{train}}(g)$;    $\bar{g}_{test} \leftarrow \mathrm{Avg}_{g \in \{g_{1,test}\} \cup \mathcal{G}_{test}}(g)$
    for $\lambda \in \Lambda_t$ do
        $\beta \leftarrow \arg\min_{\theta} \mathrm{Avg}_{Z \in \mathcal{Z}_{train}}\big(\mathcal{L}(\theta; Z)\big) - \theta^{\top}(g_{1,train} - \bar{g}_{train}) + \lambda \|\theta\|_1$
        $Loss(\lambda, q) \leftarrow \mathrm{Avg}_{Z \in \mathcal{Z}_{test}}\big(\mathcal{L}(\beta; Z)\big) - \beta^{\top}(g_{1,test} - \bar{g}_{test})$
    end for
end for
Return $\lambda^{(t)} = \arg\min_{\lambda \in \Lambda} K^{-1} \sum_{q=1}^{K} Loss(\lambda, q)$
```

3. Theoretical Analysis

Section 3.1 provides an overview of the theoretical results. Sections 3.2 and 3.3 presents the rigorous statements for linear models and generalized linear models (GLMs) respectively.

3.1 An Overview of Theoretical Results

As discussed in Section 2.3, τ has to be large enough to ensure the bootstrap accuracy, yet it also induces a great communication cost. Hence, our main goal is to pin down the minimal number of iterations $\tau_{\min}$ (communication rounds) sufficient for the bootstrap validity in Algorithm 1. An overview of the theoretical results is provided in Figure 2.

As an overall trend in Figure 2, $\tau_{\min}$ is increasing logarithmically in k and decreasing in n for both `k-grad` and `n+k-1-grad` in (generalized) linear models; in addition, $\tau_{\min}$ is increasing in $\overline{s}$ logarithmically, where $\overline{s}$ is the maximum of the sparsity of the true coefficient vector and the inverse population Hessian matrix to be formally defined later.

By comparing the left and right panels of Figure 2 under a fixed tuple $(n, k, \overline{s})$, the $\tau_{\min}$ for `k-grad` is always greater or equal to that for `n+k-1-grad`, which indicates a greater communication efficiency of `n+k-1-grad`. For very small k, `n+k-1-grad` can still provably work, while `k-grad` cannot. Particularly, $\tau_{\min} = 1$ can work for certain instances of `n+k-1-grad` but is always too small for `k-grad`.

Regarding the comparison between high-dimensional sparse linear models (top panels) and GLMs (bottom panels), GLMs typically require a greater n than sparse linear models, which ensures that the error between $\widetilde{\theta}^{(t)}$ and θ^* decreases in a short transient phase; see Section C in the Appendix for details.

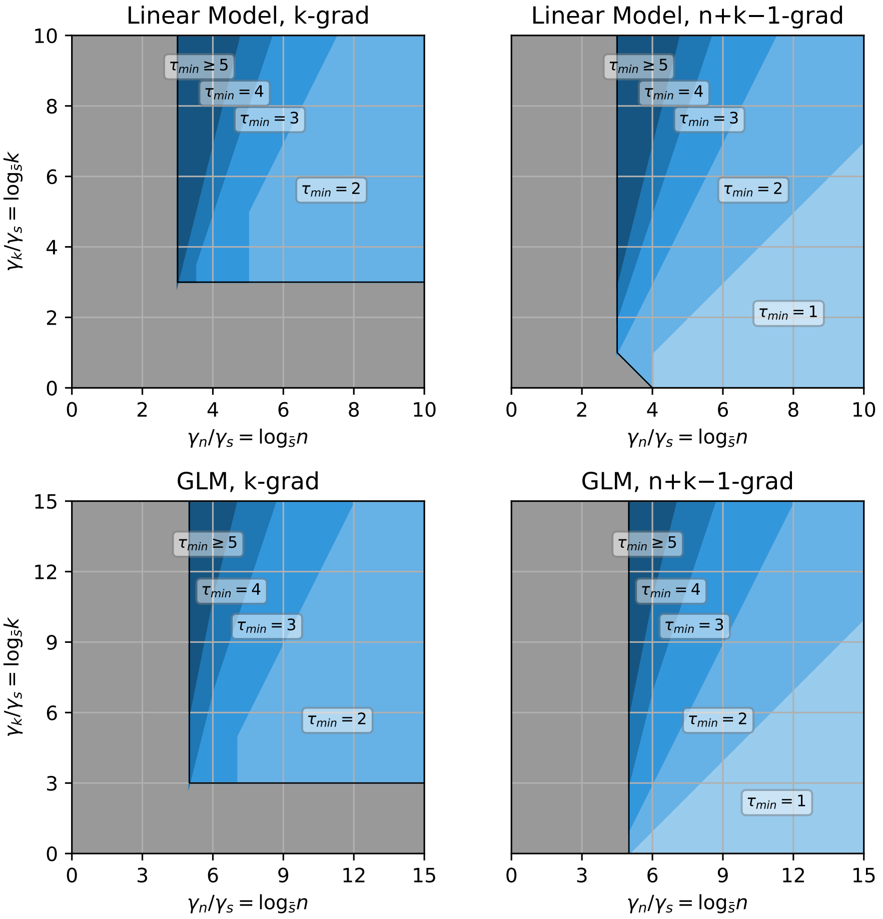

Figure 2: Illustration of Theorems 3-11. Gray region are where the bootstrap validity are not warranted by our theory, and the other area is colored blue with varying lightness according to the lower bound of iteration τ. $\gamma_n = \log_d n$, $\gamma_k = \log_d k$ and $\gamma_{\bar{s}} = \log_d \bar{s}$ are the orders of the local sample size n, number of machines k and the sparsity $\bar{s}$.

3.2 Linear Model

Suppose that N i.i.d. observations are generated by a linear model $y = x^\top \theta^* + e$ with an unknown coefficient vector $\theta^* \in \mathbb{R}^d$, covariate random vector $x \in \mathbb{R}^d$, and noise $e \in \mathbb{R}$ independent of x with zero mean and variance of σ^2. We consider the least-squares loss $\mathcal{L}(\theta; z) = \mathcal{L}(\theta; x, y) = (y - x^\top \theta)^2/2$.

We impose the following assumptions on the linear model.

(A1) x is sub-Gaussian, i.e.,

$$\sup_{\|w\|_2 \le 1} \mathbb{E}\big[\exp((w^\top x)^2/L^2)\big] = O(1),$$

for some absolute constant $L > 0$. Moreover, $1/\lambda_{\min}(\Sigma) \le \mu$ for some absolute constant $\mu > 0$, where $\Sigma = \mathbb{E}[xx^\top]$.

(A2) e is sub-Gaussian, i.e.,

$$\mathbb{E}\big[\exp(e^2/L'^2)\big] = O(1),$$

for some absolute constant $L' > 0$. Moreover, $\sigma > 0$ is an absolute constant.

(A3) θ^* and $\Theta_{l,\cdot}$ are sparse for $l = 1, \cdots, d$, where $\Theta := \Sigma^{-1} = \mathbb{E}[xx^\top]^{-1}$. Specifically, we denote by $S := \{l : \theta^*_l \neq 0\}$ the active set of covariates and its cardinality by $s_0 := |S|$. Also, we define $s_l := |\{l' \neq l : \Theta_{l,l'} \neq 0\}|$, $s^* := \max_l s_l$, and $\overline{s} = s_0 \vee s^*$.

Assumption (A1) ensures a restricted eigenvalue condition when $n \gtrsim \overline{s} \log d$ by Rudelson and Zhou (2013). Under the assumptions, we first investigate the theoretical property of Algorithm 1, where we apply `k-grad` with the de-biased ℓ_1-CSL estimator with τ communications. Define

$$T := \big\|\sqrt{N}\big(\widetilde{\theta}^{(\tau)} - \theta^*\big)\big\|_\infty, \tag{7}$$

where $\widetilde{\theta}^{(\tau)}$ is an output of Algorithm 1.

Theorem 3 (`k-grad`, sparse linear model) *Suppose (A1)-(A3) hold, and that we run Algorithm 1 with* ***`k-grad`*** *method in linear models. Let*

$$\lambda_l \asymp \sqrt{\frac{\log d}{n}} \quad \text{and} \quad \lambda^{(t)} \asymp \sqrt{\frac{\log d}{nk}} + \sqrt{\frac{\log d}{n}}\left(s_0\sqrt{\frac{\log d}{n}}\right)^t, \tag{8}$$

for $l = 1, \ldots, d$ and $t = 0, \ldots, \tau - 1$. Assume $n = d^{\gamma_n}$, $k = d^{\gamma_k}$, $\overline{s} = d^{\gamma_s}$ for some constants $\gamma_n, \gamma_k, \gamma_s > 0$. If $\gamma_n > 3\gamma_s$, $\gamma_k > 3\gamma_s$, and $\tau \ge \tau_{\min}$, where

$$\tau_{\min} = 1 + \left\lfloor \max\left\{\frac{\gamma_k + \gamma_s}{\gamma_n - 2\gamma_s}, 1 + \frac{3\gamma_s}{\gamma_n - 2\gamma_s}\right\}\right\rfloor,$$

then for T defined in (7), we have

$$\sup_{\alpha \in (0,1)} |P(T \le c_{\overline{W}}(\alpha)) - \alpha| = o(1). \tag{9}$$

where $c_{\overline{W}}(\alpha) := \inf\{t \in \mathbb{R} : P_\epsilon(\overline{W} \le t) \ge \alpha\}$, in which $\overline{W}$ is the ***`k-grad`*** *bootstrap statistics with the same distribution as $\overline{W}^{(b)}$ in (5) and P_ϵ denotes the probability with respect to the randomness from the multipliers.*

In addition, (9) also holds if T is replaced by $\widehat{T}$ defined in (3).

Theorem 3 warrants the bootstrap validity for the simultaneous confidence intervals produced by Algorithm 1 with the `k-grad`. Furthermore, it also suggests that the bootstrap quantile can approximates the quantile of the oracle statistics T; that is, our distributed bootstrap procedure is as statistically efficient as the oracle centralized method.

Next, we show that the same distributed bootstrap validity and the efficiency of the `k-grad` also hold for the `n+k-1-grad` in Algorithm 1.

Theorem 4 (`n+k-1-grad`, sparse linear model) *Suppose (A1)-(A3) hold, and that we run Algorithm 1 with `n+k-1-grad` method. Let λ_l and $\lambda^{(t)}$ be as in* (8) *for $l = 1, \ldots, d$ and $t = 0, \ldots, \tau - 1$. Assume $n = d^{\gamma_n}$, $k = d^{\gamma_k}$, $\overline{s} = d^{\gamma_s}$ for some constants $\gamma_n, \gamma_k, \gamma_s > 0$. If $\gamma_n > 3\gamma_s$, $\gamma_n + \gamma_k > 4\gamma_s$, and $\tau \geq \tau_{\min}$, where*

$$\tau_{\min} = 1 + \left\lfloor \frac{(\gamma_k \vee \gamma_s) + \gamma_s}{\gamma_n - 2\gamma_s} \right\rfloor,$$

then for T defined in (7)*, we have*

$$\sup_{\alpha \in (0,1)} |P(T \leq c_{\widetilde{W}}(\alpha)) - \alpha| = o(1). \tag{10}$$

where

$$c_{\widetilde{W}}(\alpha) := \inf\{t \in \mathbb{R} : P_\epsilon(\widetilde{W} \leq t) \geq \alpha\},$$

in which $\widetilde{W}$ is the `n+k-1-grad` bootstrap statistics with the same distribution as $\widetilde{W}^{(b)}$ in (6) *and P_ϵ denotes the probability with respect to the randomness from the multipliers.*

In addition, (10) *also holds if T is replaced by $\widehat{T}$ defined in* (3)*.*

Note by Theorem 2.4 of van de Geer et al. (2014) that $\widehat{T}$ is well approximated by $\|\widehat{\Theta}\sqrt{N}\nabla\mathcal{L}_N(\theta^*)\|_\infty$, which is further approximated by the ℓ_∞-norm of the oracle score

$$A = -\Theta \frac{1}{\sqrt{N}} \sum_{i=1}^{n} \sum_{j=1}^{k} \nabla\mathcal{L}(\theta^*; Z_{ij}),$$

given that $\widehat{\Theta}$ only deviates from Θ up to order $O_P(s^*(\log d)^{1/2} N^{-1/2})$ in ℓ_∞-norm. To gain a deeper look into the efficiency of `k-grad` and `n+k-1-grad`, we compare the difference between the covariance of A and the conditional covariance of $\overline{A}$ (for `k-grad`, defined in (5)), and $\widetilde{A}$ (for `n+k-1-grad`, defined in (6)). In particular, conditioning on the data Z_{ij}, we have

$$\begin{aligned}\left\|\!\left|\mathrm{cov}_\epsilon(\overline{A}) - \mathrm{cov}(A)\right|\!\right\|_{\max} \leq & s^* \|\widetilde{\theta}^{(\tau-1)} - \theta^*\|_1 + n s^* \|\widetilde{\theta}^{(\tau-1)} - \theta^*\|_1^2 \\ & + O_P\left(\sqrt{\frac{s^{*2}}{k}} + \sqrt{\frac{s^*}{n}}\right),\end{aligned} \tag{11}$$

$$\begin{aligned}\left\|\!\left|\mathrm{cov}_\epsilon(\widetilde{A}) - \mathrm{cov}(A)\right|\!\right\|_{\max} \leq & s^* \|\widetilde{\theta}^{(\tau-1)} - \theta^*\|_1 + (n \wedge k) s^* \|\widetilde{\theta}^{(\tau-1)} - \theta^*\|_1^2 \\ & + O_P\left(\sqrt{\frac{s^{*2}}{n+k}} + \sqrt{\frac{s^*}{n}}\right),\end{aligned} \tag{12}$$

up to some logarithmic terms in d, n or k. Overall, `n+k-1-grad` in (12) has a smaller error term than that of `k-grad` in (11). In particular, `k-grad` requires both n and k to be large, while `n+k-1-grad` requires a large n but not necessarily a large k. In addition, $\tau = 1$ could be enough for `n+k-1-grad`, but not for `k-grad`. To see it, if $\|\widetilde{\theta}^{(0)} - \theta^*\|_1$ is of order $O_P(s^*/\sqrt{n})$, the right-hand side of (11) can grow with s^*, while the error in (12) still shrinks to zero as long as $k \ll n$.

Remark 5 *Note in both Theorems 3 and 4 that the expression of $\tau_{\min}$ does not depend on d, because the direct effect of d only enters through an iterative logarithmic term $\log\log d$ which is dominated by $\log \overline{s} \asymp \log d$.*

Remark 6 *The rates of $\{\lambda^{(t)}\}_{t=0}^{\tau-1}$ and $\{\lambda_l\}_{l=1}^{d}$ in Theorems 3 and 4 are motivated by those in Wang et al. (2017) and van de Geer et al. (2014), which, unfortunately, are not useful in practice. We therefore provide a practically useful cross-validation method in Section 2.4.*

Remark 7 *The main result (Theorem 2.2) in Zhang and Cheng (2017) can be seen as a justification of multiplier bootstrap for high-dimensional linear models with data being processed in a centralized manner. Theorem 4 compliments it by justifying a distributed multiplier bootstrap with at least one round of communication ($\tau \geq 1$).*

Remark 8 *A rate of $\sup_{\alpha\in(0,1)} \left|P(T \leq c_{\overline{W}}(\alpha)) - \alpha\right|$ may be shown to be polynomial in n and k with a more careful analysis, which is faster than the order obtained by the extreme value distribution approach (Chernozhukov et al., 2013; Zhang and Cheng, 2017) that is at best logarithmic.*

Remark 9 *We have not addressed the question of whether the conditions for $\tau_{\min}$ in Theorem 3 and 4 can be improved in a minimax sense. This is left for future research. On the other hand, we remark that the total communication cost in our algorithm is of order $\Omega(\tau_{\min}kd)$, because in each iteration we communicate d-dimensional vectors between the master node and $k-1$ worker nodes, and $\tau_{\min}$ only grows logarithmically with k. Our order matches those in the existing communication-efficient statistical inference literature e.g., Jordan et al. (2019); Wang et al. (2017).*

3.3 Generalized Linear Model

In this section, we consider GLMs, which generate i.i.d. observations $(x, y) \in \mathbb{R}^d \times \mathbb{R}$. We assume that the loss function $\mathcal{L}$ is of the form $\mathcal{L}(\theta; z) = g(y, x^\top\theta)$ for $\theta, x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ with $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and $g(a, b)$ is three times differentiable with respect to b, and denote $\frac{\partial}{\partial b} g(a, b)$, $\left(\frac{\partial}{\partial b}\right)^2 g(a, b)$, $\left(\frac{\partial}{\partial b}\right)^3 g(a, b)$ by $g'(a, b)$, $g''(a, b)$, $g'''(a, b)$ respectively. We let θ^* be the unique minimizer of the expected loss $\mathcal{L}^*(\theta)$.

We let $X_1 \in \mathbb{R}^{n\times d}$ be the design matrix in the master node $\mathcal{M}_1$ and $X_1^* := P^* X_1$ be the weighted design matrix with a diagonal $P^* \in \mathbb{R}^{n\times n}$ with elements $\{g''(y_{i1}, x_{i1}^\top \theta^*)^{1/2}\}_{i=1,\ldots,n}$. We further let $(X_1^*)_{-l}\varphi_l^*$ be the L_2 projection of $(X_1^*)_l$ on $(X_1^*)_{-l}$, for $l = 1, \ldots, d$. Equivalently, for $l = 1, \ldots, d$, we define $\varphi_l^* := \arg\min_{\varphi\in\mathbb{R}^{d-1}} \mathbb{E}[\|(X_1^*)_l - (X_1^*)_{-l}\varphi\|_2^2]$.

We impose the following assumptions on the GLM.

(B1) For some $\Delta > 0$, and $\Delta' > 0$ such that $|x^\top \theta^*| \leq \Delta'$,

$$\sup_{|b| \vee |b'| \leq \Delta + \Delta'} \sup_a \frac{|g''(a,b) - g''(a,b')|}{|b - b'|} \leq 1,$$

$$\max_{|b_0| \leq \Delta} \sup_a |g'(a, b_0)| = O(1), \quad \text{and} \quad \max_{|b| \leq \Delta + \Delta'} \sup_a |g''(a,b)| = O(1).$$

(B2) $\|x\|_\infty = O(1)$. Moreover, $x^\top \theta^* = O(1)$ and $\max_l \left| g''(y, x^\top \theta^*)^{1/2} x_{-l}^\top \varphi_l^* \right| = O(1)$, where x_{-l} consists of all but the l-th coordinate of x.

(B3) The least and the greatest eigenvalues of $\nabla^2 \mathcal{L}^*(\theta^*)$ and $\mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top\right]$ are bounded away from zero and infinity respectively.

(B4) For some constant $L > 0$,

$$\max_l \max_{q=1,2} \mathbb{E}[|\mathbf{h}_l^{2+q}|/L^q] + \mathbb{E}[\exp(|\mathbf{h}_l|/L)] = O(1), \quad \text{or}$$

$$\max_l \max_{q=1,2} \mathbb{E}[|\mathbf{h}_l^{2+q}|/L^q] + \mathbb{E}[(\max_l |\mathbf{h}_l|/L)^4] = O(1),$$

where $\mathbf{h} = \nabla^2 \mathcal{L}^*(\theta^*)^{-1} \nabla \mathcal{L}(\theta^*; Z)$ and $\mathbf{h}_l$ is the l-th coordinate.

(B5) θ^* and $\Theta_{l,\cdot}$ are sparse, where the inverse population Hessian matrix $\Theta := \nabla^2 \mathcal{L}^*(\theta^*)^{-1}$, i.e., $S := \{l : \theta_l^* \neq 0\}$, $s_0 := |S|$, $s_l := |\{l' \neq l : \Theta_{l,l'} \neq 0\}|$, $s^* := \max_l s_l$, and $\overline{s} = s_0 \vee s^*$.

Assumption (B1) imposes smoothness conditions on the loss function, which is satisfied by, for example, the logistic regression. In particular, logistic regression has $g(a,b) = -ab + \log(1 + \exp(b))$, and it can be easily seen that $|g'(a,b)| \leq 2$, $|g''(a,b)| \leq 1$, $|g'''(a,b)| \leq 1$. Assumption (B2) imposes some boundedness conditions required for the validity of the nodewise Lasso (Algorithm 4; van de Geer et al. (2014)) in the master node. Assumption (B3) is a standard assumption in the GLM literature. Assumption (B4) is required for proving the validity of multiplier bootstrap (Chernozhukov et al., 2013).

Analogously to Theorem 3 and 4 that focus on the distributed bootstrap validity and the efficiency of Algorithm 1 using `k-grad`/ `n+k-1-grad` for linear models, here we extend them to the high-dimensional de-biased GLMs. See Figure 2 for a comparison between the results of high-dimensional linear models and GLMs.

Theorem 10 (`k-grad`, sparse GLM) *Suppose (B1)-(B5) hold, and that we run Algorithm 1 with* `k-grad` *method in GLMs. Let* $\lambda_l \asymp \sqrt{\log d / n}$ *for* $l = 1, \ldots, d$, *and* $\lambda^{(t)}$ *be as*

$$\lambda^{(t)} \asymp \begin{cases} \sqrt{\frac{\log d}{nk}} + \frac{1}{s_0^2}\left(s_0^2 \sqrt{\frac{\log d}{n}}\right)^{2^t}, & t \leq \tau_0, \\ \sqrt{\frac{\log d}{nk}} + \frac{1}{s_0^2}\left(s_0^2 \sqrt{\frac{\log d}{n}}\right)^{2^{\tau_0}} \left(s_0 \sqrt{\frac{\log d}{n}}\right)^{t - \tau_0}, & t > \tau_0 + 1, \end{cases} \tag{13}$$

for $t = 0, \ldots, \tau - 1$, *where*

$$\tau_0 = 1 + \left\lfloor \log_2 \frac{\gamma_n - 2\gamma_s}{\gamma_n - 4\gamma_s} \right\rfloor. \tag{14}$$

Assume $n = d^{\gamma_n}$, $k = d^{\gamma_k}$, $\overline{s} = d^{\gamma_s}$ for some constants $\gamma_n, \gamma_k, \gamma_s > 0$. If $\gamma_n > 5\gamma_s$, $\gamma_k > 3\gamma_s$, and $\tau \geq \tau_{\min}$, where

$$\tau_{\min} = \max\left\{\tau_0 + \left\lfloor \frac{\gamma_k + \gamma_s}{\gamma_n - 2\gamma_s} + \nu_0 \right\rfloor, 2 + \left\lfloor \log_2 \frac{\gamma_n - \gamma_s}{\gamma_n - 4\gamma_s} \right\rfloor \right\},$$

$$\nu_0 = 2 - \frac{2^{\tau_0}(\gamma_n - 4\gamma_s)}{\gamma_n - 2\gamma_s} \in (0, 1], \tag{15}$$

then we have (9)*. In addition,* (9) *also holds if T is replaced by $\widehat{T}$ defined in* (3)*.*

The τ_0 in (14) is the preliminary communication rounds needed for the CSL estimator to go through the regions which are far from θ^*. As $\overline{s}$ grows, the time spent in these regions can increase. However, when n is large, e.g., $n \gg \overline{s}^6$, the loss function is more well-behaved so that the preliminary communication round can reduce to $\tau_0 = 1$. See Section C in the Appendix for more details.

Theorem 11 (`n+k-1-grad`, sparse GLM) *Suppose (B1)-(B5) hold, and that we run Algorithm 1 with `n+k-1-grad` method in GLMs. Let $\lambda_l \asymp \sqrt{\log d/n}$ for $l = 1, \ldots, d$, and $\lambda^{(t)}$ be as in* (13) *for $t = 0, \ldots, \tau - 1$. Assume $n = d^{\gamma_n}$, $k = d^{\gamma_k}$, $\overline{s} = d^{\gamma_s}$ for some constants $\gamma_n, \gamma_k, \gamma_s > 0$. If $\gamma_n > 5\gamma_s$ and $\tau \geq \tau_{\min}$, where*

$$\tau_{\min} = \begin{cases} \max\left\{2 + \left\lfloor \log_2 \frac{\gamma_k + \gamma_s}{\gamma_n - 4\gamma_s} \right\rfloor, 1\right\}, & \textit{if} \quad \gamma_k \leq \gamma_n - 3\gamma_s, \\ \tau_0 + \left\lfloor \frac{\gamma_k + \gamma_s}{\gamma_n - 2\gamma_s} + \nu_0 \right\rfloor, & \textit{otherwise}, \end{cases}$$

τ_0 and ν_0 defined as in (14) *and* (15) *respectively, then we have* (10)*. In addition,* (10) *also holds if T is replaced by $\widehat{T}$ defined in* (3)*.*

Remark 12 *The selection of $\{\lambda_l\}_{l=1}^d$ in Theorems 10 and 11 are motivated by those in van de Geer et al. (2014), $\{\lambda^{(t)}\}_{t=0}^{\tau-1}$ are motivated by Wang et al. (2017) and Jordan et al. (2019). We perform a more careful analysis for the two phases of model tuning as in* (13)*.*

4. Simulation Studies

We demonstrate the merits of our methods using synthetic data in this section. The code to reproduce the simulation results and plots is available in Supplementary Material.

We consider a Gaussian linear model and a logistic regression model. We fix total sample size $N = 2^{14}$ and the dimension $d = 2^{10}$, and choose the number of machines k from $\{2^2, 2^3, \ldots, 2^6\}$. The true coefficient θ^* is a d-dimensional vector in which the first s_0 coordinates are 1 and the rest is 0, where $s_0 \in \{2^2, 2^4\}$ for the linear model and $s_0 \in \{2^1, 2^3\}$ for the GLM. We generate covariate vector x independently from $\mathcal{N}(0, \Sigma)$, while considering two different specifications for Σ:

- Toeplitz: $\Sigma_{l,l'} = 0.9^{|l-l'|}$;
- Equi-correlation: $\Sigma_{l,l'} = 0.8$ for all $l \neq l'$, $\Sigma_{l,l} = 1$ for all l.

For linear model, we generate the model noise independently from $\mathcal{N}(0,1)$; for GLM, we obtain i.i.d. responses from $y \sim \mathrm{Ber}(1/(1+\exp[-x^\top \theta^*]))$. For each choice of s_0 and k, we run Algorithm 1 with `k-grad` and `n+k-1-grad` on 1,000 independently generated datasets, and compute the empirical coverage probability and the average width based on the results from these 1,000 replications. At each replication, we draw $B = 500$ bootstrap samples, from which we calculate the 95% empirical quantile to further obtain the 95% simultaneous confidence interval.

For the ℓ_1-CSL computation, we choose the initial $\lambda^{(0)}$ by a local K-fold cross-validation, where $K = 10$ for linear regression and $K = 5$ for logistic regression. For each iteration t, $\lambda^{(t)}$ is selected by Algorithm 2 in Section 2.4 with K' folds with $K' = \min\{k-1, 5\}$, which ensures that each partition of worker gradients is non-empty when k is small. For an efficient implementation of the nodewise Lasso, we select a $\bar{\lambda}$ at every simulation repetition and set $\lambda_l = \bar{\lambda}$ for all l. Specifically, for each simulated dataset, we select $\bar{\lambda} = 10^{-1}\sum_{l=1}^{10}\hat{\lambda}_l$, where each $\hat{\lambda}_l$ is obtained obtained by a cross-validation of nodewise Lasso regression of l-th variable on the remaining variables. Since the variables are homogeneous, these $\hat{\lambda}_l$'s only deviate by some random variations, which can be alleviated by an average.

The computation of the oracle width starts with fixing (N, d, s_0) and generating 500 independent datasets. For each dataset, we compute the centralized de-biased Lasso estimator $\widehat{\theta}$ as in (4). The oracle width is defined as two times the 95% empirical quantile of $\|\widehat{\theta} - \theta^*\|_\infty$ of the 500 samples. The average widths are compared against the oracle widths by taking the ratio of the two.

The empirical coverage probabilities and the average width ratios of `k-grad` and `n+k-1-grad` are displayed for the linear model in Figures 3 (Toeplitz design) and 4 (equi-correlation design), and for the logistic regression in Figures 5 (Toeplitz design) and 6 (equi-correlation design), respectively. Note that increase in k indicates decrease in n, given the fixed N.

For small k, `k-grad` tends to over-cover, whereas `n+k-1-grad` has a more accurate coverage. By contrast, the coverage of both algorithms fall when k gets too large (or n gets too small), since the estimator $\widetilde{\theta}^{(\tau)}$ deviates from $\widehat{\theta}$ and the deviation of the width from the oracle width, which reflects the discussion of (11) and (12). Moreover, as $s_0 = \|\theta^*\|_0$ increases, it becomes harder for both algorithms to achieve the accurate 95% coverage, and both algorithms start to fail at a smaller k (or larger n), which stems from the fact that the bootstrap cannot accurately approximate variance of the asymptotic distribution as shown in (11) and (12). Nevertheless, raising the number of iterations improves the coverage, which verifies our theory. We also observe an under-coverage of our bootstrap method in both the linear regression and the logistic regression at the early stage of increasing k. This is due to the loss of accuracy in estimating the inverse Hessian matrices using only the data in the master node when k increases (or n decreases).

5. Variable Screening with Distributed Simultaneous Inference

Having demonstrated the performance of our method on purely synthetic data using sparse models in the last section, in this section, we artificially create spurious variables and mix them with the variables obtained from a real big dataset. We check if our method can successfully select the relevant variables associated with the response variable from the real

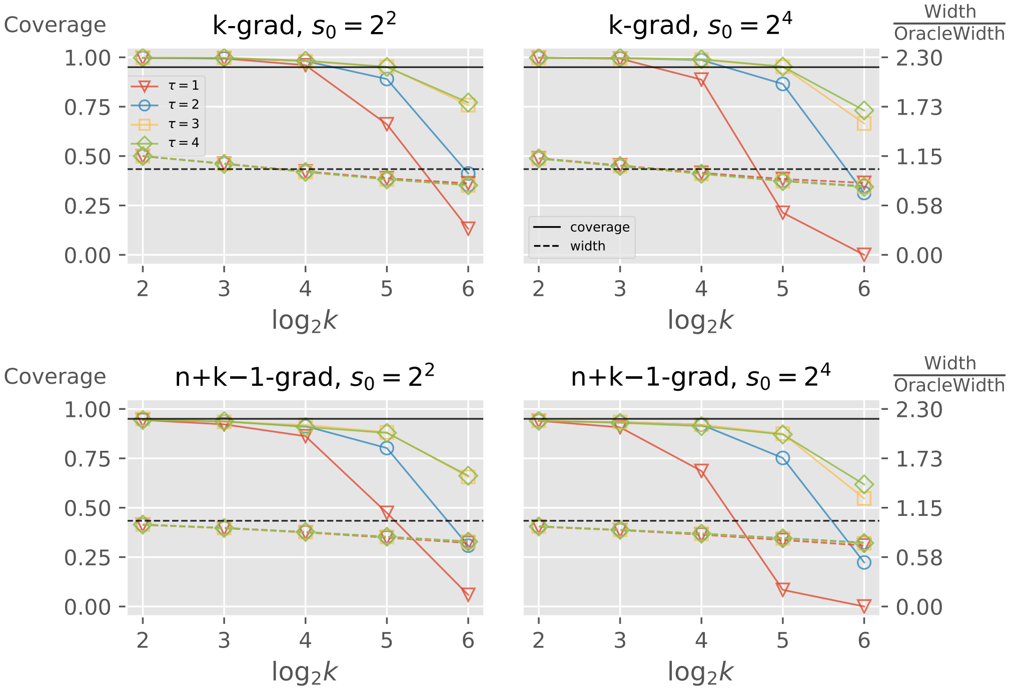

Figure 3: Empirical coverage probability (**left axis, solid lines**) and average width (**right axis, dashed lines**) of simultaneous confidence intervals by `k-grad` and `n+k-1-grad` in sparse linear regression with Toeplitz design and varying sparsity. Black solid line represents the 95% nominal level and black dashed line represents 1 on the right y-axis.

dataset. The code to retrieve data and reproduce the analyses, results, and plots is available in Supplementary Material.

5.1 Data

The US Airline On-Time Performance dataset (DVN, 2008), available at `http://stat-computing.org/dataexpo/2009`, consists of flight arrival and departure details for all commercial flights within the US from 1987 to 2008. Given the high dimensionality after dummy transformation and the huge sample size of the entire dataset, the most efficient way to process the data is using a distributed computational system, with sample size on each worker node likely to be smaller than the dimension. Our goal here is to uncover statistically significant independent variables associated with flight delay. We use variables Year, Month, DayOfWeek, CRSDepTime, CRSArrTime, UniqueCarrier, Origin, Dest, and ArrDelay in our model; descriptions are deferred to Appendix (Section D).

The response variable is labeled by 1 to denote a delay if ArrDelay is greater than zero, and by 0 otherwise. The rest of the variables are treated as categorical explanatory variables and are converted into dummy variables; refer to Appendix (Section E) for the details of the dummy variable creation. This results in a total of 203 predictors. The total sample size is 113.9 million observations. We randomly sample a dataset $\mathcal{D}_1$ of $N = 500{,}000$ observations, and conceptually distribute them across $k = 1{,}000$ nodes such that each node

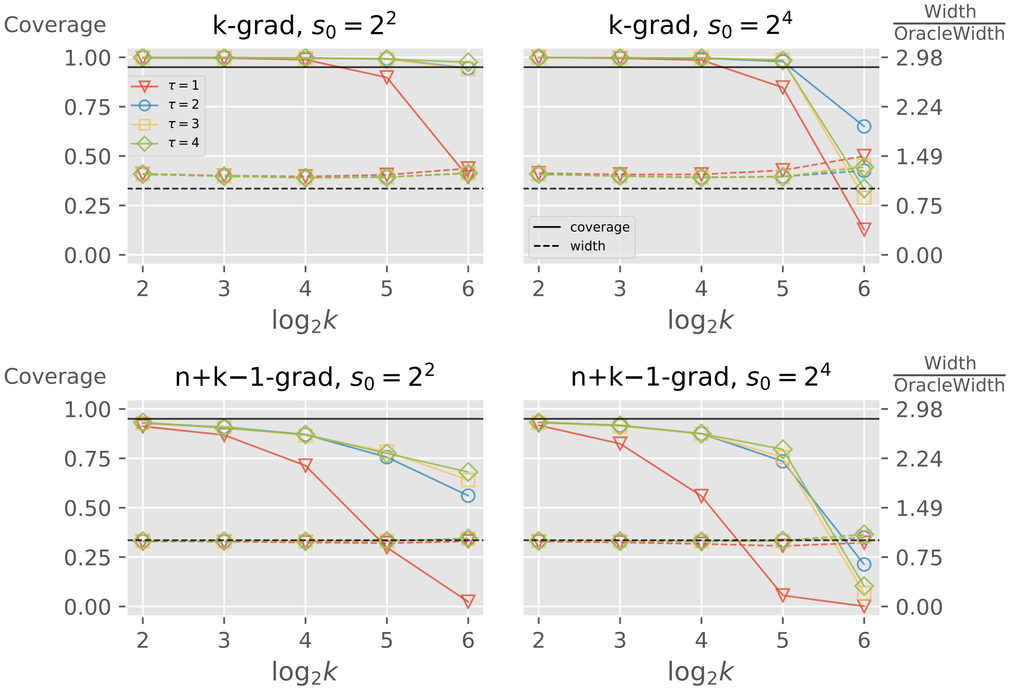

Figure 4: Empirical coverage probability (**left axis, solid lines**) and average width (**right axis, dashed lines**) of simultaneous confidence intervals by `k-grad` and `n+k-1-grad` in sparse linear regression with equi-correlation design and varying sparsity. Black solid line represents the 95% nominal level and black dashed line represents 1 on the right y-axis.

receives $n = 500$ observations. We randomly sample another dataset $\mathcal{D}_2$ of $N = 500{,}000$ observations for a pilot study to select relevant variables, where $\mathcal{D}_1 \cap \mathcal{D}_2 = \emptyset$.

5.2 An Artificial Design Matrix and Variable Screening

In the first stage, we perform a preliminary study that informs us some seemingly relevant variables to include in an artificial design matrix, which will be used to demonstrate variable screening performance of our method in the second stage. Note that the purpose of this stage is only to preliminarily discover possibly relevant variables, rather than to select variables in a fully rigorous manner. We perform a logistic regression in a centralized manner with intercept and without regularization using the N observations in $\mathcal{D}_2$. Standard Wald tests reveal that 144 out of 203 slopes are significantly non-zero (p-values less than 0.05).

The four predictors with the least p-values correspond to the dummy variables of years 2001–2004, and the coefficients are all negative, which suggests less likelihood of flight delay in these years. This interesting finding matches the results of previous study that the September 11 terrorist attacks have negatively impacted the US airline demand (Ito and Lee, 2005), which led to less flights and congestion. In addition, the *Notice of Market-based Actions to Relieve Airport Congestion and Delay*, (Docket No. OST-2001-9849) issued by Department of Transportation on August 21, 2001, might also alleviate the US airline delay.

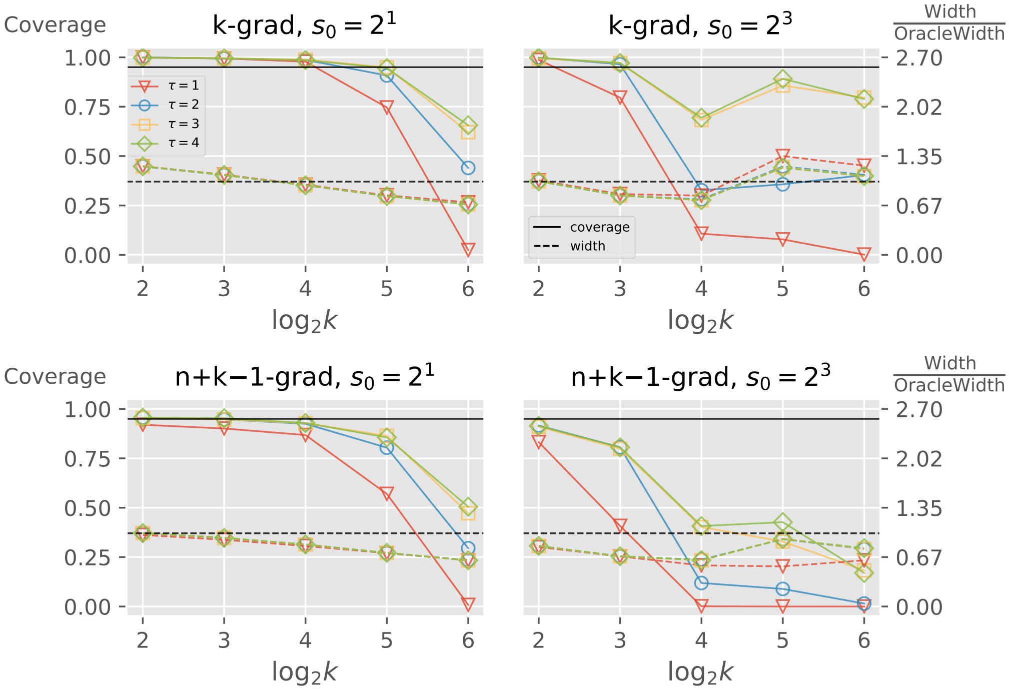

Figure 5: Empirical coverage probability (**left axis, solid lines**) and average width (**right axis, dashed lines**) of simultaneous confidence intervals by `k-grad` and `n+k-1-grad` in sparse logistic regression with Toeplitz design and varying sparsity. Black solid line represents the 95% nominal level and black dashed line represents 1 on the right y-axis.

To construct the artificial design matrix, we group the 4 predictors with the least p-values mentioned above and the intercept, so the number of the relevant columns is 5. Given d, we artificially create $d-5$ columns of binary and real valued variables by first sampling rows from $\mathcal{N}(0, \mathcal{C}_{d-5})$, where $\mathcal{C}_{d-5}$ is a Toeplitz matrix $((\mathcal{C}_{d-5})_{l,l'} = 0.5^{|l-l'|})$, and then converting half of the columns to either 0 or 1 by their signs. Then, we combine these $d-5$ spurious columns with a column of intercept and the 4 columns in $\mathcal{D}_1$ that are associated with the selected relevant variables to obtain an artificial design matrix.

In the second stage, using the artificial design matrix with the binary response vector from the ArrDelay in $\mathcal{D}_1$, we test if our distributed bootstrap `n+k-1-grad` (Algorithm 1) can screen the artificially created spurious variables. Note that $\mathcal{D}_1$ and $\mathcal{D}_2$ are disjoint, where $\mathcal{D}_2$ is used in the first stage for the preliminary study. For model tuning, we select $\lambda^{(0)}$ by a local 10-fold cross-validation; for each $t \geq 1$, $\lambda^{(t)}$ is chosen by running a distributed 10-fold cross-validation in Algorithm 2. We select each λ_l by performing a 10-fold cross-validation for the nodewise Lasso of each variable. The same entire procedure is repeated under each dimensionality $d \in \{200, 500, 1{,}000\}$.

The left panel of Figure 7 plots the number of significant variables against the number of iterations τ, which was broken down into the number intersecting with the relevant variables (solid lines) and the number intersecting with the spurious variables (dashed lines). First, all of the 4 relevant variables are tested to be significant at all iterations. For the spurious

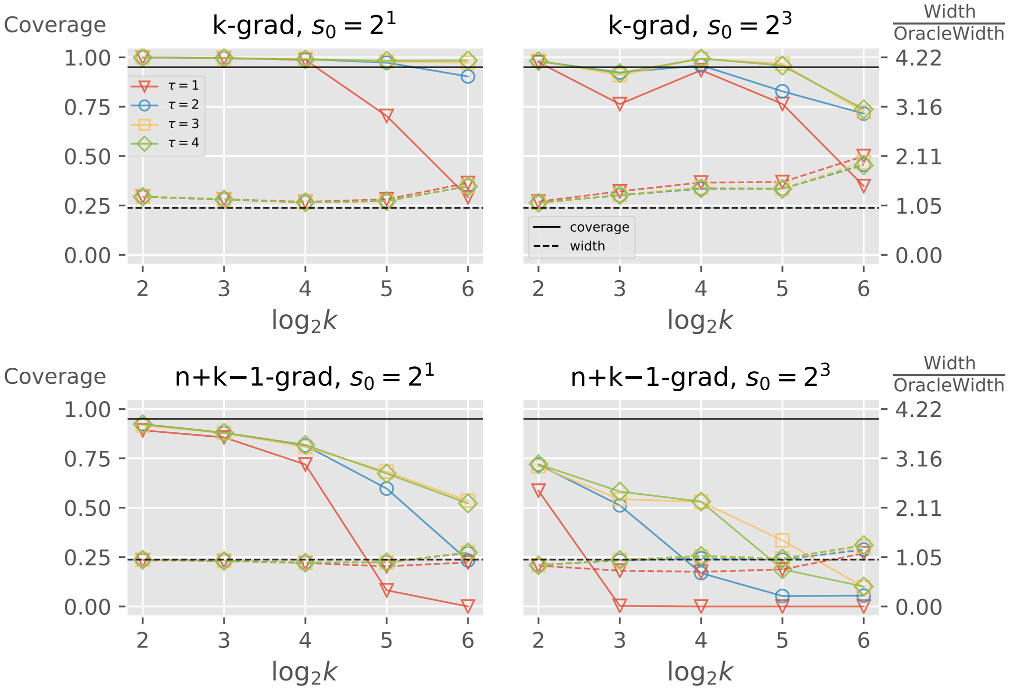

Figure 6: Empirical coverage probability (**left axis, solid lines**) and average width (**right axis, dashed lines**) of simultaneous confidence intervals by `k-grad` and `n+k-1-grad` in sparse logistic regression with equi-correlation design and varying sparsity. Black solid line represents the 95% nominal level and black dashed line represents 1 on the right y-axis.

variables, we see that with $\tau = 1$, the distributed bootstrap falsely detects one of them. However, as the number of iterations increases, less spurious variables are detected until none of them is detected. We also see that 2 iterations ($\tau = 2$) for $d = 500, 1{,}000$ and 3 iterations ($\tau = 3$) for $d = 200$ are sufficient, which empirically verifies that our method is not very sensitive to the nominal dimension d.

As an illustration that is potentially useful in practice, the confidence intervals computed with the simultaneous quantile for the 4 important slopes under $d = 1{,}000$ and $\tau = 2$ are plotted in the right panel of Figure 7. It can be seen that the flights in years 2002 and 2003 are relatively less likely to delay, which match the decreased air traffic in the aftermath of the September 11 terrorist attacks.

6. Conclusion

We propose a distributed bootstrap method for high-dimensional simultaneous inference based on the de-biased ℓ_1-CSL estimator as well as a distributed cross-validation method for hyperparameter tuning. The bootstrap validity and oracle efficiency are rigorously studied, and the merits are further shown via simulation study on coverage probability and efficiency, and a practical example on variable screening.

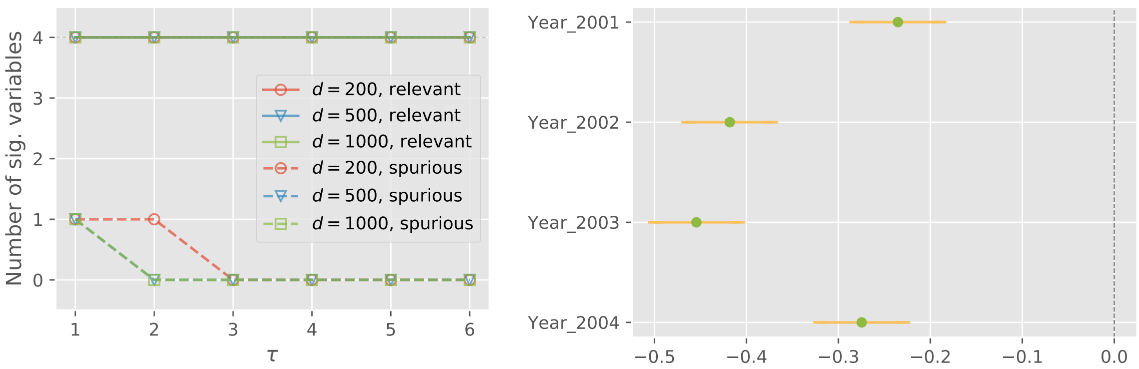

Figure 7: The left panel shows the number of significant variables uncovered by the simultaneous confidence intervals among the 4 relevant variables and among the $d - 5$ spurious variables for $d = 200, 500, 1{,}000$. The right panel shows the simultaneous confidence intervals of the 4 relevant variables for $d = 1{,}000$ and $\tau = 2$.

Acknowledgments

Shih-Kang Chao would like to acknowledge the financial support from the Research Council of the University of Missouri. Guang Cheng would like to acknowledge support from the National Science Foundation (NSF – SCALE MoDL (2134209)).

Appendix A. Pseudocode for `k-grad` and `n+k-1-grad`

Algorithm 3 `DistBoots`(`method`, $\widetilde{\theta}$, $\{\mathbf{g}_j\}_{j=1}^k$, $\widetilde{\Theta}$): only need the master node $\mathcal{M}_1$

```
Require: local gradient $\mathbf{g}_j$ and estimate $\widetilde{\Theta}$ of inverse Hessian obtained at $\mathcal{M}_1$
$\bar{\mathbf{g}} \leftarrow k^{-1} \sum_{j=1}^k \mathbf{g}_j$
for $b = 1, \dots, B$ do
    if method=‘k-grad’ then
        Draw $\epsilon_1^{(b)}, \dots, \epsilon_k^{(b)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and compute $W^{(b)}$ by (5)
    else if method=‘n+k-1-grad’ then
        Draw $\epsilon_{11}^{(b)}, \dots, \epsilon_{n1}^{(b)}, \epsilon_2^{(b)}, \dots, \epsilon_k^{(b)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and compute $W^{(b)}$ by (6)
    end if
end for
Compute the quantile $c_W(\alpha)$ of $\{W^{(1)}, \dots, W^{(B)}\}$ for $\alpha \in (0,1)$
Return $\widetilde{\theta}_l \pm N^{-1/2} c_W(\alpha)$, $l = 1, \dots, d$
```

Remark 13 *Although in Algorithm 3 the same $\widetilde{\theta}$ is used for the center of the confidence interval and for evaluating the gradients $\mathbf{g}_{ij}$, allowing them to be different (such as in Algorithm 1) can save one round of communication. For example, we can use $\widetilde{\theta}^{(\tau)}$ for the center of the confidence interval, while the gradients are evaluated with $\widetilde{\theta}^{(\tau-1)}$.*

Appendix B. Nodewise Lasso

In Algorithm 4, we state the nodewise Lasso method for constructing approximate inverse Hessian matrix used in Section 3.1.1 of van de Geer et al. (2014), which we apply in Algorithm 1. We define the components of $\widehat{\gamma}_l$ as $\widehat{\gamma}_l = \{\widehat{\gamma}_{l,l'}; l' = 1, \dots, d, l' \neq l\}$. We denote by $\widehat{M}_{l,-l}$ the l-th row of $\widehat{M}$ without the diagonal element (l,l), and by $\widehat{M}_{-l,-l}$ the submatrix without the l-th row and l-th column.

Algorithm 4 `Node`($\widehat{M}$)

```
Require: sample Hessian matrix $\widehat{M} \in \mathbb{R}^{d \times d}$, hyperparameters $\{\lambda_l\}_{l=1}^d$
for $l = 1, \dots, d$ do
    Compute $\widehat{\gamma}_l = \arg\min_{\gamma \in \mathbb{R}^{d-1}} \widehat{M}_{l,l} - 2\widehat{M}_{l,-l}\gamma + \gamma^\top \widehat{M}_{-l,-l}\gamma + 2\lambda_l \|\gamma\|_1$
    Compute $\widehat{\tau}_l^2 = \widehat{M}_{l,l} - \widehat{M}_{l,-l}\widehat{\gamma}_l$
end for
Construct $\widehat{M^{-1}}$ as
```

$$\widehat{M^{-1}} = \begin{pmatrix} \widehat{\tau}_1^{-2} & 0 & \dots & 0 \\ 0 & \widehat{\tau}_2^{-2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \widehat{\tau}_d^{-2} \end{pmatrix} \begin{pmatrix} 1 & -\widehat{\gamma}_{1,2} & \dots & -\widehat{\gamma}_{1,d} \\ -\widehat{\gamma}_{2,1} & 1 & \dots & -\widehat{\gamma}_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\gamma}_{d,1} & -\widehat{\gamma}_{d,2} & \dots & 1 \end{pmatrix}.$$

Remark 14 *Throughout this paper, we fix the choice of nodewise Lasso in Algorithm 1 for computing an approximate inverse Hessian matrix. In practice, various approaches (e.g., Zhang and Zhang (2014); Javanmard and Montanari (2014a)) can be chosen from in consideration of estimation accuracy and computational efficiency.*

Appendix C. CSL Estimator for GLMs

For the ℓ_1-penalized CSL estimator of generalized linear models, Theorem 3.3 of Wang et al. (2017) states that

$$\left\|\widetilde{\theta}^{(t+1)} - \theta^*\right\|_1 \lesssim s_0\sqrt{\frac{\log d}{N}} + s_0\sqrt{\frac{\log d}{n}}\left\|\widetilde{\theta}^{(t)} - \theta^*\right\|_1 + Ms_0\left\|\widetilde{\theta}^{(t)} - \theta^*\right\|_1^2, \tag{16}$$

where $M \geq 0$ is a Lipschitz constant of the g'', which exists due to Assumptions (B1). In linear models, $g(a, b) = (a - b)^2/2$, g'' is a constant, so $M = 0$ and CSL estimator has linear convergence to θ^* with rate $s_0(\log d)^{1/2}n^{-1/2}$ until it reaches the upper bound given by the first term, which is also the rate of the centralized (oracle) estimator. For GLMs, however, $M > 0$ and the third term can be dominant when t is small. For example, when $t = 0$, given that $\|\widetilde{\theta}^{(0)} - \theta^*\|_1 \lesssim s_0(\log d)^{1/2}n^{-1/2}$, it is easy to see that the third term is always s_0 times larger than the second term (up to a constant), and a larger n is required to ensure third term is less than $\left\|\widetilde{\theta}^{(t)} - \theta^*\right\|_1$ and the error is shrinking. However, when t is sufficiently large, this dominance reverses. The threshold is given by the τ_0 in (14), and this implies the three phases of convergence: When $t \leq \tau_0$, the third term dominates and the convergence is quadratic; when $t > \tau_0$, the second term dominates the third and the linear convergence kicks in. Finally, when t is sufficiently large, the first term dominates. Our analysis complements that of Wang et al. (2017), while in their Corollary 3.7 it is simply assumed that the second term dominates the third.

Appendix D. Variable Descriptions

We use the following variables in our model for the semi-synthetic study in Section 5:

- Year: from 1987 to 2008,
- Month: from 1 to 12,
- DayOfWeek: from 1 (Monday) to 7 (Sunday),
- CRSDepTime: scheduled departure time (in four digits, first two representing hour, last two representing minute),
- CRSArrTime: scheduled arrival time (in the same format as above),
- UniqueCarrier: unique carrier code,
- Origin: origin (in IATA airport code),
- Dest: destination (in IATA airport code),
- ArrDelay: arrival delay (in minutes). Positive value means there is a delay.

The complete variable information can be found at `http://stat-computing.org/dataexpo/2009/the-data.html`.

Appendix E. Creation of Dummy Variables

We categorize CRSDepTime and CRSArrTime into 24 one-hour time intervals (e.g., 1420 is converted to 14 to represent the interval [14:00,15:00]), and then treat Year, Month, DayOfWeek, CRSDepTime, CRSArrTime, UniqueCarrier, Origin, and Dest as nominal predictors. The nominal predictors are encoded by dummies with appropriate dimensions and merging all categories of lower counts into "others", and either "others" or the smallest ordinal value is treated as the baseline.

To ensure that none of the columns of the design matrix on the master node is completely zero so that the nodewise Lasso can be computed, we create the dummy variables using only the observations in the master node on the dataset $\mathcal{D}_1$. Specifically, for variables UniqueCarrier, Origin, and Dest, we keep the top categories that make up 90% of the data in the master node on $\mathcal{D}_1$; the rest categories are merged into "others" and are treated as baseline. For CRSDepTime and CRSArrTime, we merge the time intervals 23:00-6:00 and 1:00-7:00 respectively (due to their low counts) and use them as baseline. For Year, Month, and DayOfWeek, we treat year 1987, January, and Monday as baseline respectively.

Appendix F. Extension to Heteroscedastic Error Across Machines

As suggested by the associated editor, here we consider an extension to a more challenging scenario for linear models where the data across machines have heteroscedastic errors. In this scenario, Algorithm 2 can no longer apply as it relies on the homogeneity in data across machines. We provide a new Algorithm 5 by exploiting the multiplier bootstrap idea underlying the "High-Dimensional Metrics" (HDM, Chernozhukov et al. (2016)).

In Algorithm 5, we select the regularization parameters $\{\lambda^{(t)}\}_{t=1}^{\tau-1}$ in lines 12-16 by integrating the idea of Spindler et al. (2016). In addition, we handle heteroscedasticity by data-driven regularization loadings $\Psi^{(t)}$ in lines 17-21.

Under heteroscedasticity, we expect the `k-grad` in Algorithm 3 to continue being valid because it treats each machine equally as an independent data point. However, `n+k-1-grad` may no longer provide an accurate coverage because each single data point in the first machine is treated as equally important as the average of entire data in j machine for $j = 2, \cdots, k$, so the variance in the first machine could dominate so that the `n+k-1-grad` bootstrap could fail to precisely approximate the variance of the target empirical distribution. A careful theoretical study deserves future research.

The empirical performance of Algorithm 5 is verified by a simulation study based on a heteroscedastic Gaussian linear model. We fix total sample size $N = 2^{14}$ and the dimension $d = 2^{10}$, and choose the number of machines k from $\{2^2, 2^3, \ldots, 2^6\}$. The true coefficient θ^* is a d-dimensional vector in which the first s_0 coordinates are 1 and the rest is 0, where $s_0 \in \{2^2, 2^4\}$. We generate covariate vector x independently from $\mathcal{N}(0, \Sigma)$, where Σ is a Toeplitz matrix with $\Sigma_{l,l'} = 0.9^{|l-l'|}$. We introduce heteroscedasticity across machines by first independently generating the model noise from $\mathcal{N}(0, 1)$ for all data in the master node $\mathcal{M}_1$. Next, we generate model noise for each data point i in worker node $\mathcal{M}_j$ ($j = 2, \ldots, k$) independently from $\mathcal{N}(0, \sigma_j^2 + \omega_{ij})$, where the node level variance σ_j^2 is generated independently from Unif$(2, 3)$ and the idiosyncratic variance ω_{ij} is generated independently from Unif$(-0.2, 0.2)$. For each choice of s_0 and k, we run Algorithm 5 with `k-grad` and

Algorithm 5 Simultaneous inference for distributed data with heteroscedasticity

```
Require: $\tau \geq 1$ rounds of communication; nodewise Lasso procedure Node$(\cdot,\cdot)$ with hyperparameters $\{\lambda_l\}_{l=1}^d$, theoretical constant $c$
$\widetilde{\theta}^{(0)} \leftarrow \arg\min_\theta \mathcal{L}_1(\theta) + \lambda^{(0)}\|\theta\|_1$ at $\mathcal{M}_1$, where $\lambda^{(0)}$ is chosen by cross-validation using the data at $\mathcal{M}_1$
Compute $\widetilde{\Theta}$ by running Node$(\nabla^2\mathcal{L}_1(\widetilde{\theta}^{(0)}), \{\lambda_l\}_{l=1}^d)$ at $\mathcal{M}_1$
for $t = 1, \ldots, \tau$ do
    Transmit $\widetilde{\theta}^{(t-1)}$ to $\{\mathcal{M}_j\}_{j=2}^k$
    Compute $\nabla\mathcal{L}_1(\widetilde{\theta}^{(t-1)})$ and $\psi_1^{(t-1)} = n^{-1}\sum_{i=1}^n \nabla\mathcal{L}((x_{i1}, y_{i1}), \widetilde{\theta}^{(t-1)})^2$ at $\mathcal{M}_1$
    for $j = 2, \ldots, k$ do
        Compute $\nabla\mathcal{L}_j(\widetilde{\theta}^{(t-1)})$ and $\psi_j^{(t-1)} = n^{-1}\sum_{i=1}^n \nabla\mathcal{L}((x_{ij}, y_{ij}), \widetilde{\theta}^{(t-1)})^2$ at $\mathcal{M}_j$
        Transmit $\nabla\mathcal{L}_j(\widetilde{\theta}^{(t-1)})$ and $\psi_j^{(t-1)}$ to $\mathcal{M}_1$
    end for
    $\nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) \leftarrow k^{-1}\sum_{j=1}^k \nabla\mathcal{L}_j(\widetilde{\theta}^{(t-1)})$ at $\mathcal{M}_1$
    if $t < \tau$ then
        for $b = 1, \ldots, B$ do
            Draw $\epsilon_1^{(b)}, \ldots, \epsilon_k^{(b)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
            $\Lambda_b^{(t)} \leftarrow ck^{-1}\|\sum_{j=1}^k \epsilon_j^{(b)}\nabla\mathcal{L}_j(\widetilde{\theta}^{(t-1)})\|_\infty$
        end for
        $\lambda^{(t)} \leftarrow 90\%$ quantile of $\{\Lambda_1^{(t)}, \ldots, \Lambda_B^{(t)}\}$
        for $l = 1, \ldots, d$ do
            $\Psi_l^{(t)} \leftarrow \sqrt{k^{-1}\sum_{j=1}^k (\psi_j^{(t-1)})_l}$
        end for
        $\Psi^{(t)} \leftarrow \mathrm{diag}(\Psi_1^{(t)}, \ldots, \Psi_d^{(t)})$
        $\widetilde{\theta}^{(t)} \leftarrow \arg\min_\theta \mathcal{L}_1(\theta) - \theta^\top\left(\nabla\mathcal{L}_1(\widetilde{\theta}^{(t-1)}) - \nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)})\right) + \lambda^{(t)}\|\Psi^{(t)}\theta\|_1$ at $\mathcal{M}_1$
    else
        $\widetilde{\theta}^{(\tau)} \leftarrow \widetilde{\theta}^{(\tau-1)} - \widetilde{\Theta}\nabla\mathcal{L}_N(\widetilde{\theta}^{(\tau-1)})$ at $\mathcal{M}_1$
    end if
end for
Run DistBoots('k-grad' or 'n+k-1-grad', $\widetilde{\theta} = \widetilde{\theta}^{(\tau)}$, $\{\mathbf{g}_j = \nabla\mathcal{L}_j(\widetilde{\theta}^{(\tau-1)})\}_{j=1}^k$,
                    $\widetilde{\Theta} = \widetilde{\Theta}$) at $\mathcal{M}_1$
```

n+k-1-grad on 1,000 independently generated datasets, and compute the empirical coverage probability and the average width based on the results from these 1,000 replications. At each replication, we draw $B = 500$ bootstrap samples, from which we calculate the 95% empirical quantile to further obtain the 95% simultaneous confidence interval. For tuning the nodewise Lasso, we use the same approach as in the main text. The computation of the oracle width starts with fixing (N, d, s_0, k) and generating 500 independent datasets. For each dataset, we compute the centralized de-biased Lasso estimator $\widehat{\theta}$. The oracle width is defined as two times the 95% empirical quantile of $\|\widehat{\theta} - \theta^*\|_\infty$ of the 500 samples.

Figure 8 shows the coverage probability and efficiency in the form of relative widths of Algorithm 5. As expected, the coverage of the simultaneous confidence intervals is improved as the iteration goes using the new data-driven parameter tuning and heteroscedasticity-adapted regularization. The `k-grad` performs much better than the `n+k-1-grad`, which basically fails as the coverage probability of `n+k-1-grad` is nearly zero in all cases. The failure of the `n+k-1-grad` is due to the fact that it over-weigh the data in the master node $\mathcal{M}_1$ which leads to an under-estimation of the variance in other nodes, whereas in `k-grad` each node is weighed equally.

By comparing Figure 8 and Figure 9, we observe that our algorithm is generally robust to the selection of c as it performs similarly for $c = 0.5$ and $c = 1$. However, we note that $c = 0.5$ could be too small to stabilize the algorithm as the optimization solver in 21 fails to converge in about 2% of the replications, and the divergent results are not included in Figure 8. This suggests that the penalty at $c = 0.5$ may be so small that leads to an ill-conditioned objective function. After increasing c from 0.5 to 1, optimization solvers of all the replications stably converge.

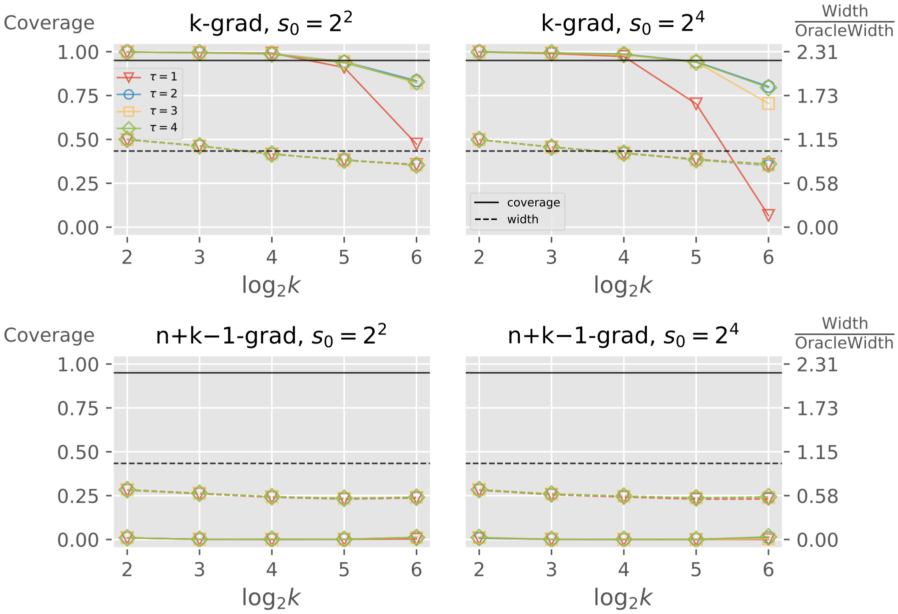

Figure 8: Under $c = 0.5$, empirical coverage probability (**left axis, solid lines**) and average relative width (**right axis, dashed lines**) of simultaneous confidence intervals by `k-grad` and `n+k-1-grad` in sparse linear regression with Toeplitz design and varying sparsity. Black solid line represents the 95% nominal level and black dashed line represents 1 on the right y-axis.

References

Data Expo 2009: Airline on time data, 2008. URL https://doi.org/10.7910/DVN/HG7NV7.

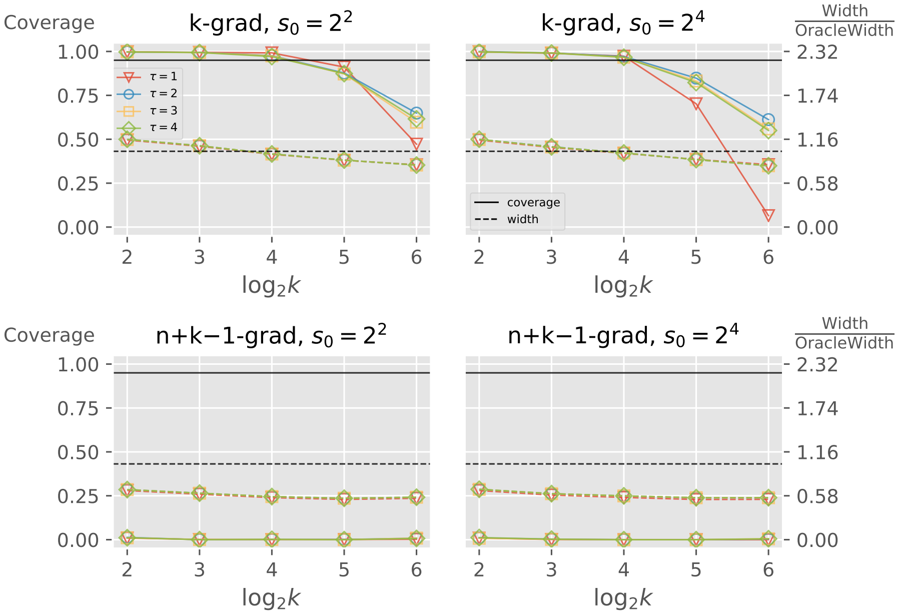

Figure 9: Under $c = 1$, empirical coverage probability (**left axis, solid lines**) and average relative width (**right axis, dashed lines**) of simultaneous confidence intervals by `k-grad` and `n+k-1-grad` in sparse linear regression with Toeplitz design and varying sparsity. Black solid line represents the 95% nominal level and black dashed line represents 1 on the right y-axis.

Moulinath Banerjee, Cecile Durot, Bodhisattva Sen, et al. Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *The Annals of Statistics*, 47(2): 720–757, 2019.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed estimation and inference with statistical guarantees. *Annals of Statistics*, 46(3):1352–1382, 2018.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Ying Wei. Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *Ann. Statist.*, 46(6B):3643–3675, 12 2018. doi: 10.1214/17-AOS1671.

Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association*, 114(526):749–758, 2019. doi: 10.1080/01621459.2018.1442339.

T Tony Cai and Wenguang Sun. Large-scale global and simultaneous inference: Estimation and testing in very high dimensions. *Annual Review of Economics*, 9:411–439, 2017.

Xi Chen, Weidong Liu, and Yichen Zhang. First-order newton-type estimator for distributed estimation and inference. *arXiv preprint arXiv:1811.11368*, 2018.

Xi Chen, Weidong Liu, and Yichen Zhang. Quantile regression under memory constraint. *Ann. Statist.*, 47(6):3244–3273, 12 2019. doi: 10.1214/18-AOS1777.

Xi Chen, Jason D Lee, He Li, and Yun Yang. Distributed estimation for principal component analysis: a gap-free approach. *arXiv preprint arXiv:2004.02336*, 2020.

Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, pages 1655–1684, 2014.

Victor Chernozhukov, Denis Chetverikov, Kengo Kato, et al. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

Victor Chernozhukov, Chris Hansen, and Martin Spindler. hdm: High-dimensional metrics. *The R Journal*, 8(2):185–199, 2016.

Ruben Dezeure, Peter Bühlmann, and Cun-Hui Zhang. High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719, 2017.

Jianqing Fan, Yongyi Guo, and Kaizheng Wang. Communication-efficient accurate statistical estimation. *arXiv preprint arXiv:1906.04870*, 2019a.

Jianqing Fan, Dong Wang, Kaizheng Wang, and Ziwei Zhu. Distributed estimation of principal eigenspaces. *Annals of statistics*, 47(6):3009, 2019b.

Cheng Huang and Xiaoming Huo. A distributed one-step estimator. *Mathematical Programming*, 174(1-2):41–76, 2019.

Harumi Ito and Darin Lee. Assessing the impact of the september 11 terrorist attacks on us airline demand. *Journal of Economics and Business*, 57(1):75–95, 2005.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014a.

Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014b.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, pages 1–48, 2018.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144, 2017.

Runze Li, Dennis KJ Lin, and Bing Li. Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29(5):399–409, 2013.

M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013. doi: 10.1109/TIT.2013.2243201.

Srijan Sengupta, Stanislav Volgushev, and Xiaofeng Shao. A subsampled double bootstrap for massive data. *Journal of the American Statistical Association*, 111(515):1222–1232, 2016.

Chengchun Shi, Wenbin Lu, and Rui Song. A massive data framework for m-estimators with cubic-rate. *Journal of the American Statistical Association*, 113(524):1698–1709, 2018.

Kamalpreet Singh and Ravinder Kaur. Hadoop: addressing challenges of big data. In *2014 IEEE International Advance Computing Conference (IACC)*, pages 686–689. IEEE, 2014.

Martin Spindler, Victor Chernozhukov, and Christian Hansen. hdm: High-dimensional metrics. *R Package Version 0.1. 0. Available at http://CRAN. R-project. org/package= hdm.[233]*, 2016.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Stanislav Volgushev, Shih-Kang Chao, Guang Cheng, et al. Distributed inference for quantile regression processes. *The Annals of Statistics*, 47(3):1634–1662, 2019.

Jialei Wang and Tong Zhang. Improved optimization of finite sums with minibatch stochastic variance reduced proximal iterations. *arXiv preprint arXiv:1706.07001*, 2017.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3636–3645. JMLR. org, 2017.

Ming Yu, Varun Gupta, and Mladen Kolar. Simultaneous inference for pairwise graphical models with generalized score matching. *Journal of Machine Learning Research*, 21(91):1–51, 2020a.

Yang Yu, Shih-Kang Chao, and Guang Cheng. Simultaneous inference for massive data: Distributed bootstrap. In *International Conference on Machine Learning*, pages 10892–10901. PMLR, 2020b.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Xianyang Zhang and Guang Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768, 2017.

Yuchen Zhang, Martin J Wainwright, and John C Duchi. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.

Tianqi Zhao, Guang Cheng, and Han Liu. A partially linear framework for massive heterogeneous data. *Annals of Statistics*, 44(4):1400, 2016.

Xuening Zhu, Feng Li, and Hansheng Wang. Least squares approximation for a distributed system. *ArXiv Preprint arXiv:1908.04904*, 2020.

SUPPLEMENTARY MATERIAL

1. Proofs of Main Results

To simplify the notation, in the proof we denote $\bar{\theta} = \widetilde{\theta}^{(\tau-1)}$, where $\widetilde{\theta}^{(\tau-1)}$ is the ℓ_1-penalized estimator at $\tau - 1$ iterator output by Algorithm 1. Denote $\widetilde{\theta} = \widetilde{\theta}^{(\tau)}$ output by Algorithm 1.
Proof of Theorem 3. We apply Theorem 3 of Wang et al. (2017), where their Assumption 2 is inherited from Assumption (A1), and obtain that if $n \gg s_0^2 \log d$,

$$\left\|\bar{\theta} - \theta^*\right\|_1 = \left\|\widetilde{\theta}^{(\tau-1)} - \theta^*\right\|_1 = O_P\left(s_0\sqrt{\frac{\log d}{N}} + \left(s_0\sqrt{\frac{\log d}{n}}\right)^{\tau}\right).$$

Then, by Lemma 15, we have that $\sup_{\alpha\in(0,1)} \left|P(T \le c_{\widetilde{W}}(\alpha)) - \alpha\right| = o(1)$, as long as $n \gg s^{*2} \log^{3+\kappa} d + s^* \log^{5+\kappa} d + s_0^2 \log d$, $k \gg s^{*2} \log^{5+\kappa} d$, and

$$s_0\sqrt{\frac{\log d}{N}} + \left(s_0\sqrt{\frac{\log d}{n}}\right)^{\tau} \ll \min\left\{\frac{1}{\sqrt{ks^*}\log^{1+\kappa} d}, \frac{1}{\sqrt{ns^*}\log^{1+\kappa} d}\right\}.$$

These conditions hold if $n \gg (s^{*2} + s^* s_0^2)\log^{3+\kappa} d + s^* \log^{5+\kappa} d$, $k \gg s^* s_0^2 \log^{3+\kappa} d + s^{*2} \log^{5+\kappa} d$, and

$$\tau > \max\left\{\frac{\log k + \log s^* + \log(C\log^{2+\kappa} d)}{\log n - \log(s_0^2) - \log\log d}, 1 + \frac{\log s^* + \log(s_0^2) + \log(C\log^{3+\kappa} d)}{\log n - \log(s_0^2) - \log\log d}\right\}.$$

If $n = d^{\gamma_n}$, $k = d^{\gamma_k}$, $\overline{s} = s_0 \vee s^* = d^{\gamma_s}$ for some constants γ_n, γ_k, and γ_s, then a sufficient condition is $\gamma_n > 3\gamma_s$, $\gamma_k > 3\gamma_s$, and

$$\tau \ge 1 + \left\lfloor \max\left\{\frac{\gamma_k + \gamma_s}{\gamma_n - 2\gamma_s}, 1 + \frac{3\gamma_s}{\gamma_n - 2\gamma_s}\right\}\right\rfloor.$$

■

Proof of Theorem 4. Similarly to the proof of Theorem 3, applying Theorem 3 of Wang et al. (2017) and Lemma 16, we have that $\sup_{\alpha\in(0,1)} \left|P(T \le c_{\widetilde{W}}(\alpha)) - \alpha\right| = o(1)$, as long as $n \gg s^{*2} \log^{3+\kappa} d + s^* \log^{5+\kappa} d + s_0^2 \log d$, $n + k \gg s^{*2} \log^{5+\kappa} d$, and

$$s_0\sqrt{\frac{\log d}{N}} + \left(s_0\sqrt{\frac{\log d}{n}}\right)^{\tau} \ll \min\left\{\frac{1}{\sqrt{ks^*}\log^{1+\kappa} d}, \frac{1}{s^*\sqrt{\log((n+k)d)}\log^{2+\kappa} d}\right\}.$$

These conditions hold if $n \gg (s^{*2} + s^* s_0^2)\log^{3+\kappa} d + s^* \log^{5+\kappa} d$, $n + k \gg s^{*2} \log^{5+\kappa} d$, $nk \gg s^{*2} s_0^2 \log^{5+\kappa} d$, and

$$\tau > \max\left\{\frac{\log k + \log s^* + \log(C\log^{2+\kappa} d)}{\log n - \log(s_0^2) - \log\log d}, \frac{\log s^{*2} + \log\log((n+k)d) + \log(C\log^{4+\kappa} d)}{\log n - \log(s_0^2) - \log\log d}\right\}.$$

If $n = d^{\gamma_n}$, $k = d^{\gamma_k}$, $\overline{s} = s_0 \vee s^* = d^{\gamma_s}$ for some constants γ_n, γ_k, and γ_s, then a sufficient condition is $\gamma_n > 3\gamma_s$, $\gamma_n + \gamma_k > 4\gamma_s$, and

$$\tau \ge 1 + \left\lfloor \frac{(\gamma_k \vee \gamma_s) + \gamma_s}{\gamma_n - 2\gamma_s}\right\rfloor.$$

■

Proof of Theorem 10. We apply Theorem 6 of Wang et al. (2017), where their Assumption 2 is inherited from Assumption (B3), and obtain that if $n \gg s_0^4 \log d$,

$$\left\|\bar{\theta}-\theta^*\right\|_1=\left\|\widetilde{\theta}^{(\tau-1)}-\theta^*\right\|_1=\begin{cases} O_P\left(s_0\sqrt{\frac{\log d}{N}}+\frac{1}{s_0}\left(s_0^2\sqrt{\frac{\log d}{n}}\right)^{2^{\tau-1}}\right), & \tau \leq \tau_0+1, \\ O_P\left(s_0\sqrt{\frac{\log d}{N}}+\frac{1}{s_0}\left(s_0^2\sqrt{\frac{\log d}{n}}\right)^{2^{\tau_0}}\left(s_0\sqrt{\frac{\log d}{n}}\right)^{\tau-\tau_0-1}\right), & \tau > \tau_0+1, \end{cases}$$

where τ_0 is the smallest integer t such that

$$\left(s_0^2\sqrt{\frac{\log d}{n}}\right)^{2^t} \lesssim s_0\sqrt{\frac{\log d}{n}},$$

that is,

$$\tau_0=\left\lceil \log_2\left(\frac{\log n-\log(s_0^2)-\log(C\log d)}{\log n-\log(s_0^4)-\log\log d}\right)\right\rceil.$$

Then, by Lemma 17, we have that $\sup_{\alpha\in(0,1)}\left|P(T\leq c_{\overline{W}}(\alpha))-\alpha\right|=o(1)$, as long as $n \gg (s_0^2+s^{*2})\log^{3+\kappa} d+(s_0+s^*)\log^{5+\kappa} d+s_0^4\log d$, $k \gg s^{*2}\log^{5+\kappa} d$, and

$$s_0\sqrt{\frac{\log d}{N}}+\frac{1}{s_0}\left(s_0^2\sqrt{\frac{\log d}{n}}\right)^{2^{\tau-1}} \ll \min\left\{\frac{1}{\sqrt{ks^*}s_0\log^{1+\kappa} d},\frac{1}{\sqrt{ns^*}\log^{1+\kappa} d}\right\},$$

if $\tau \leq \tau_0+1$, and

$$s_0\sqrt{\frac{\log d}{N}}+\frac{1}{s_0}\left(s_0^2\sqrt{\frac{\log d}{n}}\right)^{2^{\tau_0}}\left(s_0\sqrt{\frac{\log d}{n}}\right)^{\tau-\tau_0-1} \ll \min\left\{\frac{1}{\sqrt{ks^*}s_0\log^{1+\kappa} d},\frac{1}{\sqrt{ns^*}\log^{1+\kappa} d}\right\},$$

if $\tau > \tau_0+1$.

If $n=d^{\gamma_n}$, $k=d^{\gamma_k}$, $\overline{s}=s_0\vee s^*=d^{\gamma_s}$ for some constants γ_n, γ_k, and γ_s, then a sufficient condition is $\gamma_n>5\gamma_s$, $\gamma_k>3\gamma_s$, and

$$\begin{aligned}\tau &\geq 1+\left\lfloor \max\left\{1+\log_2\frac{\gamma_n-\gamma_s}{\gamma_n-4\gamma_s},\tau_0+1+\frac{\gamma_k+(4\cdot 2^{\tau_0}+1)\gamma_s-2^{\tau_0}\gamma_n}{\gamma_n-2\gamma_s}\right\}\right\rfloor \\ &= \left\lfloor \max\left\{2+\log_2\frac{\gamma_n-\gamma_s}{\gamma_n-4\gamma_s},\tau_0+2+\frac{\gamma_k+(4\cdot 2^{\tau_0}+1)\gamma_s-2^{\tau_0}\gamma_n}{\gamma_n-2}\right\}\right\rfloor \\ &= \left\lfloor \max\left\{2+\log_2\frac{\gamma_n-\gamma_s}{\gamma_n-4\gamma_s},\tau_0+\frac{\gamma_k+\gamma_s}{\gamma_n-2\gamma_s}+\nu_0\right\}\right\rfloor \\ &= \max\left\{\tau_0+\left\lfloor\frac{\gamma_k+\gamma_s}{\gamma_n-2\gamma_s}+\nu_0\right\rfloor,2+\left\lfloor\log_2\frac{\gamma_n-\gamma_s}{\gamma_n-4\gamma_s}\right\rfloor\right\},\end{aligned}$$

where

$$\tau_0=1+\left\lfloor\log_2\frac{\gamma_n-2\gamma_s}{\gamma_n-4\gamma_s}\right\rfloor,\quad \nu_0=2-\frac{2^{\tau_0}(\gamma_n-4\gamma_s)}{\gamma_n-2\gamma_s}\in(0,1].$$

■

Proof of Theorem 11. Similarly to the proof of Theorem 4, applying Theorem 3 of Wang et al. (2017) and Lemma 18, we have that $\sup_{\alpha\in(0,1)} \left|P(T \le c_{\overline{W}}(\alpha)) - \alpha\right| = o(1)$, as long as $n \gg (s_0 + s^*)\log^{5+\kappa} d + (s_0^2 + s^{*2})\log^{3+\kappa} d$, $n + k \gg s^{*2}\log^{5+\kappa} d$, and

$$\begin{aligned}
& s_0\sqrt{\frac{\log d}{N}} + \frac{1}{s_0}\left(s_0^2\sqrt{\frac{\log d}{n}}\right)^{2^{\tau-1}} \\
& \ll \min\left\{\frac{n+k}{s^*\left(n + k\sqrt{\log d} + k^{3/4}\log^{3/4} d\right)\log^{2+\kappa} d}, \frac{1}{\sqrt{ks^*}s_0\log^{1+\kappa} d}, \frac{1}{\left(nks^*\log^{1+\kappa} d\right)^{1/4}}\right\},
\end{aligned}$$

if $\tau \le \tau_0 + 1$, and

$$\begin{aligned}
& s_0\sqrt{\frac{\log d}{N}} + \frac{1}{s_0}\left(s_0^2\sqrt{\frac{\log d}{n}}\right)^{2^{\tau_0}}\left(s_0\sqrt{\frac{\log d}{n}}\right)^{\tau-\tau_0-1} \\
& \ll \min\left\{\frac{n+k}{s^*\left(n + k\sqrt{\log d} + k^{3/4}\log^{3/4} d\right)\log^{2+\kappa} d}, \frac{1}{\sqrt{ks^*}s_0\log^{1+\kappa} d}, \frac{1}{\left(nks^*\log^{1+\kappa} d\right)^{1/4}}\right\},
\end{aligned}$$

if $\tau > \tau_0 + 1$, where

$$\tau_0 = \left\lceil \log_2\left(\frac{\log n - \log(s_0^2) - \log(C\log d)}{\log n - \log(s_0^4) - \log\log d}\right)\right\rceil.$$

If $n = d^{\gamma_n}$, $k = d^{\gamma_k}$, $\overline{s} = s_0 \vee s^* = d^{\gamma_s}$ for some constants γ_n, γ_k, and γ_s, then a sufficient condition is $\gamma_n > 5\gamma_s$, and

Let $\overline{s} = s_0 \vee s^*$. If $n = \overline{s}^{\gamma_n}$, $k = \overline{s}^{\gamma_k}$, and $d = \overline{s}^{\gamma_d}$ for some constants γ_n, γ_k, and γ_d, then a sufficient condition is $\gamma_n > 5$, and if $\tau \le \tau_0 + 1$,

$$\tau \ge \max\left\{2 + \left\lfloor \log_2\frac{\gamma_k + 1}{\gamma_n - 4}\right\rfloor, 1\right\},$$

and if $\tau > \tau_0 + 1$

$$\begin{aligned}
\tau &\ge 1 + \left\lfloor \tau_0 + 1 + \frac{\gamma_k + 4\cdot 2^{\tau_0} + 1 - 2^{\tau_0}\gamma_n}{\gamma_n - 2}\right\rfloor \\
&= \left\lfloor \tau_0 + 2 + \frac{\gamma_k + 4\cdot 2^{\tau_0} + 1 - 2^{\tau_0}\gamma_n}{\gamma_n - 2}\right\rfloor \\
&= \left\lfloor \tau_0 + \frac{\gamma_k + 1}{\gamma_n - 2} + \nu_0\right\rfloor \\
&= \tau_0 + \left\lfloor \frac{\gamma_k + 1}{\gamma_n - 2} + \nu_0\right\rfloor,
\end{aligned}$$

where

$$\tau_0 = 1 + \left\lfloor \log_2\frac{\gamma_n - 2}{\gamma_n - 4}\right\rfloor, \quad \nu_0 = 2 - \frac{2^{\tau_0}(\gamma_n - 4)}{\gamma_n - 2} \in (0, 1].$$

■

2. Technical Lemmas

Lemma 15 (`k-grad`) *In sparse linear model, under Assumptions (A1) and (A2), if* $n \gg s^{*2} \log^{3+\kappa} d + s^* \log^{5+\kappa} d$, $k \gg s^{*2} \log^{5+\kappa} d$, *and*

$$\left\|\bar{\theta} - \theta^*\right\|_1 \ll \min\left\{\frac{1}{\sqrt{ks^*}\log^{1+\kappa} d}, \frac{1}{\sqrt{ns^*}\log^{1+\kappa} d}\right\},$$

for some $\kappa > 0$*, then we have that*

$$\sup_{\alpha\in(0,1)} \left|P(T \le c_{\overline{W}}(\alpha)) - \alpha\right| = o(1), \quad \textit{and} \tag{17}$$

$$\sup_{\alpha\in(0,1)} \left|P(\widehat{T} \le c_{\overline{W}}(\alpha)) - \alpha\right| = o(1). \tag{18}$$

Proof of Lemma 15. As noted by Zhang and Cheng (2017), since $\|\sqrt{N}(\widetilde{\theta} - \theta^*)\|_\infty = \max_l \sqrt{N}|\widetilde{\theta}_l - \theta^*_l| = \sqrt{N}\max_l \left((\widetilde{\theta}_l - \theta^*_l) \vee (\theta^*_l - \widetilde{\theta}_l)\right)$, the arguments for the bootstrap consistency result with

$$T = \max_l \sqrt{N}(\widetilde{\theta} - \theta^*)_l \quad \text{and} \tag{19}$$

$$\widehat{T} = \max_l \sqrt{N}(\widehat{\theta} - \theta^*)_l \tag{20}$$

imply the bootstrap consistency result for $T = \|\sqrt{N}(\widetilde{\theta} - \theta^*)\|_\infty$ and $\widehat{T} = \|\sqrt{N}(\widehat{\theta} - \theta^*)\|_\infty$. Hence, from now on, we redefine T and $\widehat{T}$ as (19) and (20). Define an oracle multiplier bootstrap statistic as

$$W^* := \max_{1\le l\le d} -\frac{1}{\sqrt{N}}\sum_{i=1}^n\sum_{j=1}^k \left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*; Z_{ij})\right)_l \epsilon^*_{ij}, \tag{21}$$

where $\{\epsilon^*_{ij}\}_{i=1,\dots,n;j=1,\dots,k}$ are N independent standard Gaussian variables, also independent of the entire dataset. The proof consists of two steps; the first step is to show that W^* achieves bootstrap consistency, i.e., $\sup_{\alpha\in(0,1)}|P(T \le c_{W^*}(\alpha)) - \alpha|$ converges to 0, where $c_{W^*}(\alpha) = \inf\{t \in \mathbb{R} : P_\epsilon(W^* \le t) \ge \alpha\}$, and the second step is to show the bootstrap consistency of our proposed bootstrap statistic by showing the quantiles of W and W^* are close.

Note that $\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*; Z) = \mathbb{E}[xx^\top]^{-1}x(x^\top\theta^* - y) = \Theta x e$ and

$$\mathbb{E}\left[\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*; Z)\right)\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*; Z)\right)^\top\right] = \Theta\mathbb{E}\left[xx^\top e^2\right]\Theta = \sigma^2\Theta\Sigma\Theta = \sigma^2\Theta.$$

Then, under Assumptions (A1) and (A2),

$$\min_l \mathbb{E}\left[\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*; Z)\right)_l^2\right] = \sigma^2\min_l \Theta_{l,l} \ge \sigma^2\lambda_{\min}(\Theta) = \frac{\sigma^2}{\lambda_{\max}(\Sigma)}, \tag{22}$$

is bounded away from zero. Under Assumption (A1), x is sub-Gaussian, that is, $w^\top x$ is sub-Gaussian with uniformly bounded ψ_2-norm for all $w \in S^{d-1}$. To show $w^\top\Theta x$ is also sub-Gaussian with uniformly bounded ψ_2-norm, we write it as

$$w^\top\Theta x = (\Theta w)^\top x = \|\Theta w\|_2\left(\frac{\Theta w}{\|\Theta w\|_2}\right)^\top x.$$

Since $\Theta w / \left\|\Theta w\right\|_2 \in S^{d-1}$, we have that $\left(\Theta w / \left\|\Theta w\right\|_2\right) x$ is sub-Gaussian with $O(1)$ ψ_2-norm, and hence, $w^\top \Theta x$ is sub-Gaussian with $O(\|\Theta w\|_2) = O(\lambda_{\max}(\Theta)) = O(\lambda_{\min}(\Sigma)^{-1}) = O(1)$ ψ_2-norm, under Assumption (A1). Since e is also sub-Gaussian under Assumption (A2) and is independent of $w^\top \Theta x$, we have that $w^\top \Theta x e$ is sub-exponential with uniformly bounded ψ_1-norm for all $w \in S^{d-1}$, and also, all $\left(\nabla^2 \mathcal{L}^*(\theta^*)^{-1} \nabla \mathcal{L}(\theta^*; Z)\right)_l$ are sub-exponential with uniformly bounded ψ_1-norm. Combining this with (22), we have verified Assumption (E.1) of Chernozhukov et al. (2013) for $\nabla^2 \mathcal{L}^*(\theta^*)^{-1} \nabla \mathcal{L}(\theta^*; Z)$.

Define

$$T_0 := \max_{1 \le l \le d} -\sqrt{N} \left(\nabla^2 \mathcal{L}^*(\theta^*)^{-1} \nabla \mathcal{L}_N(\theta^*)\right)_l, \tag{23}$$

which is a Bahadur representation of T. Under the condition $\log^7(dN)/N \lesssim N^{-c}$ for some constant $c > 0$, which holds if $N \gtrsim \log^{7+\kappa} d$ for some $\kappa > 0$, applying Theorem 3.2 and Corollary 2.1 of Chernozhukov et al. (2013), we obtain that for some constant $c > 0$ and for every $v, \zeta > 0$,

$$\begin{aligned} \sup_{\alpha \in (0,1)} \left|P(T \le c_{W^*}(\alpha)) - \alpha\right| &\lesssim N^{-c} + v^{1/3} \left(1 \vee \log \frac{d}{v}\right)^{2/3} + P\left(\left\|\widehat{\Omega} - \Omega_0\right\|_{\max} > v\right) \\ &\quad + \zeta \sqrt{1 \vee \log \frac{d}{\zeta}} + P\left(|T - T_0| > \zeta\right), \end{aligned} \tag{24}$$

where

$$\begin{aligned} \widehat{\Omega} &:= \operatorname{cov}_\epsilon \left(-\frac{1}{\sqrt{N}} \sum_{i=1}^n \sum_{j=1}^k \nabla^2 \mathcal{L}^*(\theta^*)^{-1} \nabla \mathcal{L}(\theta^*; Z_{ij}) \epsilon^*_{ij}\right) \\ &= \nabla^2 \mathcal{L}^*(\theta^*)^{-1} \left(\frac{1}{N} \sum_{i=1}^n \sum_{j=1}^k \nabla \mathcal{L}(\theta^*; Z_{ij}) \nabla \mathcal{L}(\theta^*; Z_{ij})^\top\right) \nabla^2 \mathcal{L}^*(\theta^*)^{-1}, \quad \text{and} \end{aligned} \tag{25}$$

$$\Omega_0 := \operatorname{cov}\left(-\nabla^2 \mathcal{L}^*(\theta^*)^{-1} \nabla \mathcal{L}(\theta^*; Z)\right) = \nabla^2 \mathcal{L}^*(\theta^*)^{-1} \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top\right] \nabla^2 \mathcal{L}^*(\theta^*)^{-1}. \tag{26}$$

To show the quantiles of $\overline{W}$ and W^* are close, we first have that for any ω such that $\alpha + \omega, \alpha - \omega \in (0,1)$,

$$\begin{aligned} &P(\{T \le c_{\overline{W}}(\alpha)\} \ominus \{T \le c_{W^*}(\alpha)\}) \\ &\le 2P(c_{W^*}(\alpha - \omega) < T \le c_{W^*}(\alpha + \omega)) + P(c_{W^*}(\alpha - \omega) > c_{\overline{W}}(\alpha)) + P(c_{\overline{W}}(\alpha) > c_{W^*}(\alpha + \omega)), \end{aligned}$$

where $\ominus$ denotes symmetric difference. Following the arguments in the proof of Lemma 3.2 of Chernozhukov et al. (2013), we have that

$$P(c_{\overline{W}}(\alpha) > c_{W^*}(\alpha + \pi(u))) \le P\left(\left\|\overline{\Omega} - \widehat{\Omega}\right\|_{\max} > u\right), \quad \text{and}$$

$$P(c_{W^*}(\alpha - \pi(u)) > c_{\overline{W}}(\alpha)) \le P\left(\left\|\overline{\Omega} - \widehat{\Omega}\right\|_{\max} > u\right),$$

where $\pi(u) := u^{1/3} \left(1 \vee \log(d/u)\right)^{2/3}$ and

$$\begin{aligned} \overline{\Omega} &:= \operatorname{cov}_\epsilon \left(-\frac{1}{\sqrt{k}} \sum_{j=1}^{k} \widetilde{\Theta} \sqrt{n} \left(\nabla \mathcal{L}_j(\bar{\theta}) - \nabla \mathcal{L}_N(\bar{\theta}) \right) \epsilon_j \right) \\ &= \widetilde{\Theta} \left(\frac{1}{k} \sum_{j=1}^{k} n \left(\nabla \mathcal{L}_j(\bar{\theta}) - \nabla \mathcal{L}_N(\bar{\theta}) \right) \left(\nabla \mathcal{L}_j(\bar{\theta}) - \nabla \mathcal{L}_N(\bar{\theta}) \right)^\top \right) \widetilde{\Theta}^\top. \end{aligned} \tag{27}$$

By letting $\omega = \pi(u)$, we have that

$$\begin{aligned} & P(\{T \le c_{\overline{W}}(\alpha)\} \ominus \{T \le c_{W^*}(\alpha)\}) \\ & \le 2P(c_{W^*}(\alpha - \pi(u)) < T \le c_{W^*}(\alpha + \pi(u))) + P(c_{W^*}(\alpha - \pi(u)) > c_{\overline{W}}(\alpha)) + P(c_{\overline{W}}(\alpha) > c_{W^*}(\alpha + \pi(u))) \\ & \le 2P(c_{W^*}(\alpha - \pi(u)) < T \le c_{W^*}(\alpha + \pi(u))) + 2P\left(\left\| \overline{\Omega} - \widehat{\Omega} \right\|_{\max} > u \right), \end{aligned}$$

where by (24),

$$\begin{aligned} P(c_{W^*}(\alpha - \pi(u)) < T \le c_{W^*}(\alpha + \pi(u))) &= P(T \le c_{W^*}(\alpha + \pi(u))) - P(T \le c_{W^*}(\alpha - \pi(u))) \\ &\lesssim \pi(u) + N^{-c} + \zeta \sqrt{1 \vee \log \frac{d}{\zeta}} + P\left(|T - T_0| > \zeta\right), \end{aligned}$$

and then,

$$\begin{aligned} \sup_{\alpha \in (0,1)} \left| P(T \le c_{\overline{W}}(\alpha)) - \alpha \right| &\lesssim N^{-c} + v^{1/3} \left(1 \vee \log \frac{d}{v} \right)^{2/3} + P\left(\left\| \widehat{\Omega} - \Omega_0 \right\|_{\max} > v \right) \\ &\quad + \zeta \sqrt{1 \vee \log \frac{d}{\zeta}} + P\left(|T - T_0| > \zeta\right) + u^{1/3} \left(1 \vee \log \frac{d}{u} \right)^{2/3} + P\left(\left\| \overline{\Omega} - \widehat{\Omega} \right\|_{\max} > u \right) \end{aligned} \tag{28}$$

Applying Lemmas 19, 24, and 23, we have that there exist some $\zeta, u, v > 0$ such that

$$\zeta \sqrt{1 \vee \log \frac{d}{\zeta}} + P\left(|T - T_0| > \zeta\right) = o(1), \quad \text{and} \tag{29}$$

$$u^{1/3} \left(1 \vee \log \frac{d}{u} \right)^{2/3} + P\left(\left\| \overline{\Omega} - \widehat{\Omega} \right\|_{\max} > u \right) = o(1), \quad \text{and} \tag{30}$$

$$v^{1/3} \left(1 \vee \log \frac{d}{v} \right)^{2/3} + P\left(\left\| \widehat{\Omega} - \Omega_0 \right\|_{\max} > v \right) = o(1), \tag{31}$$

and hence, after simplifying the conditions, obtain the first result in the lemma. To obtain the second result, we use Lemma 20, which yields

$$\xi \sqrt{1 \vee \log \frac{d}{\xi}} + P\left(|\widehat{T} - T_0| > \xi \right) = o(1). \tag{32}$$

■

Lemma 16 (`n+k-1-grad`) *In sparse linear model, under Assumptions (A1) and (A2), if* $n \gg s^{*2} \log^{3+\kappa} d + s^* \log^{5+\kappa} d$, $n + k \gg s^{*2} \log^{5+\kappa} d$, $nk \gtrsim \log^{7+\kappa} d$, *and*

$$\left\|\bar{\theta} - \theta^*\right\|_1 \ll \min\left\{ \frac{1}{\sqrt{ks^*} \log^{1+\kappa} d}, \frac{1}{s^* \sqrt{\log((n+k)d)} \log^{2+\kappa} d} \right\},$$

for some $\kappa > 0$*, then we have that*

$$\sup_{\alpha\in(0,1)} \left|P(T \leq c_{\widetilde{W}}(\alpha)) - \alpha\right| = o(1), \quad \textit{and} \tag{33}$$

$$\sup_{\alpha\in(0,1)} \left|P(\widehat{T} \leq c_{\widetilde{W}}(\alpha)) - \alpha\right| = o(1). \tag{34}$$

Proof of Lemma 16. By the argument in the proof of Lemma 15, we have that

$$\sup_{\alpha\in(0,1)} \left|P(T \leq c_{\widetilde{W}}(\alpha)) - \alpha\right| \lesssim N^{-c} + v^{1/3}\left(1 \vee \log\frac{d}{v}\right)^{2/3} + P\left(\left\|\widehat{\Omega} - \Omega_0\right\|_{\max} > v\right)$$
$$+ \zeta\sqrt{1 \vee \log\frac{d}{\zeta}} + P\left(|T - T_0| > \zeta\right) + u^{1/3}\left(1 \vee \log\frac{d}{u}\right)^{2/3} + P\left(\left\|\widetilde{\Omega} - \widehat{\Omega}\right\|_{\max} > u\right) \tag{35}$$

where

$$\widetilde{\Omega} := \mathrm{cov}_\epsilon\left(-\frac{1}{\sqrt{n+k-1}}\left(\sum_{i=1}^n \widetilde{\Theta}\left(\nabla\mathcal{L}(\bar{\theta}; Z_{i1}) - \nabla\mathcal{L}_N(\bar{\theta})\right)\epsilon_{i1} + \sum_{j=2}^k \widetilde{\Theta}\sqrt{n}\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta})\right)\epsilon_j\right)\right)$$
$$= \widetilde{\Theta}\frac{1}{n+k-1}\left(\sum_{i=1}^n \left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}_N(\theta)\right)^\top\right.$$
$$\left. + \sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)^\top\right)\widetilde{\Theta}^\top, \tag{36}$$

if $N \gtrsim \log^{7+\kappa} d$ for some $\kappa > 0$. Applying Lemmas 19, 24, and 25, we have that there exist some $\zeta, u, v > 0$ such that (29),

$$u^{1/3}\left(1 \vee \log\frac{d}{u}\right)^{2/3} + P\left(\left\|\widetilde{\Omega} - \widehat{\Omega}\right\|_{\max} > u\right) = o(1), \tag{37}$$

and (31) hold, and hence, after simplifying the conditions, obtain the first result in the lemma. To obtain the second result, we use Lemma 20, which yields (32). ■

Lemma 17 (`k-grad`) *In sparse GLM, under Assumptions (B1)–(B4), if* $n \gg (s_0^2 + s^{*2}) \log^{3+\kappa} d + (s_0 + s^*) \log^{5+\kappa} d$, $k \gg s^{*2} \log^{5+\kappa} d$, *and*

$$\left\|\bar{\theta} - \theta^*\right\|_1 \ll \min\left\{ \frac{1}{\sqrt{ks^*} s_0 \log^{1+\kappa} d}, \frac{1}{\sqrt{ns^*} \log^{1+\kappa} d} \right\},$$

for some $\kappa > 0$*, then we have that* (17) *and* (18) *hold.*

Proof of Lemma 17. We redefine T and $\widehat{T}$ as (19) and (20). We define an oracle multiplier bootstrap statistic as in (21). Under Assumption (B3),

$$\begin{aligned}\min_l \mathbb{E}\left[\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*;Z)\right)_l^2\right] &= \min_l \left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\nabla^2\mathcal{L}^*(\theta^*)^{-1}\right)_{l,l}\\ &\geq \lambda_{\min}\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\nabla^2\mathcal{L}^*(\theta^*)^{-1}\right)\\ &\geq \lambda_{\min}\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\right)^2\lambda_{\min}\left(\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right)\\ &= \frac{\lambda_{\min}\left(\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right)}{\lambda_{\max}\left(\nabla^2\mathcal{L}^*(\theta^*)\right)^2}\end{aligned}$$

is bounded away from zero. Combining this with Assumption (B4), we have verified Assumption (E.1) of Chernozhukov et al. (2013) for $\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*;Z)$. Then, we use the same argument as in the proof of Lemma 15, and obtain (28) with

$$\overline{\Omega} := \widetilde{\Theta}(\widetilde{\theta}^{(0)})\left(\frac{1}{k}\sum_{j=1}^k n\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top\right)\widetilde{\Theta}(\widetilde{\theta}^{(0)})^\top, \tag{38}$$

under the condition $\log^7(dN)/N \lesssim N^{-c}$ for some constant $c > 0$, which holds if $N \gtrsim \log^{7+\kappa} d$ for some $\kappa > 0$. Applying Lemmas 21, 27, and 26, we have that there exist some $\zeta, u, v > 0$ such that (29), (30), and (31) hold, and hence, after simplifying the conditions, obtain the first result in the lemma. To obtain the second result, we use Lemma 22, which yields (32). ■

Lemma 18 (`n+k-1-grad`) *In sparse GLM, under Assumptions (B1)–(B4), if* $n \gg (s_0 + s^*)\log^{5+\kappa} d + (s_0^2 + s^{*2})\log^{3+\kappa} d$, $n + k \gg s^{*2}\log^{5+\kappa} d$, $nk \gtrsim \log^{7+\kappa} d$, *and*

$$\left\|\bar{\theta}-\theta^*\right\|_1 \ll \min\left\{\frac{n+k}{s^*\left(n+k\sqrt{\log d}+k^{3/4}\log^{3/4} d\right)\log^{2+\kappa} d}, \frac{1}{\sqrt{ks^*}s_0\log^{1+\kappa} d}, \frac{1}{\left(nks^*\log^{1+\kappa} d\right)^{1/4}}\right\},$$

for some $\kappa > 0$, *then we have that* (33) *and* (34) *hold.*

Proof of Lemma 18. By the argument in the proof of Lemma 17, we obtain (35) with

$$\begin{aligned}\widetilde{\Omega} := \widetilde{\Theta}(\widetilde{\theta}^{(0)})\frac{1}{n+k-1}\Bigg(&\sum_{i=1}^n\left(\nabla\mathcal{L}(\theta;Z_{i1})-\nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}(\theta;Z_{i1})-\nabla\mathcal{L}_N(\theta)\right)^\top\\ &+\sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\theta)-\nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}_j(\theta)-\nabla\mathcal{L}_N(\theta)\right)^\top\Bigg)\widetilde{\Theta}(\widetilde{\theta}^{(0)})^\top,\end{aligned} \tag{39}$$

if $N \gtrsim \log^{7+\kappa} d$ for some $\kappa > 0$. Applying Lemmas 21, 27, and 28, we have that there exist some $\zeta, u, v > 0$ such that (29), (37), and (31) hold, and hence, after simplifying the conditions, obtain the first result in the lemma. To obtain the second result, we use Lemma 22, which yields (32). ■

Lemma 19 *T and T_0 are defined as in (7) and (23) respectively. In sparse linear model, under Assumptions (A1) and (A2), provided that $\left\|\bar{\theta}-\theta^*\right\|_1 = O_P(r_{\bar{\theta}})$ and $n \gg s^* \log d$, we have that*

$$|T - T_0| = O_P\left(r_{\bar{\theta}}\sqrt{s^* k \log d} + \frac{s^* \log d}{\sqrt{n}}\right).$$

Moreover, if $n \gg s^{*2} \log^{3+\kappa} d$ and

$$\left\|\bar{\theta}-\theta^*\right\|_1 \ll \frac{1}{\sqrt{ks^*}\log^{1+\kappa} d},$$

for some $\kappa > 0$, then there exists some $\zeta > 0$ such that (29) holds.

Proof of Lemma 19. First, we note that

$$\begin{aligned}|T - T_0| &\leq \max_{1\leq l\leq d}\left|\sqrt{N}(\widetilde{\theta}-\theta^*)_l + \sqrt{N}\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right)_l\right| \\ &= \sqrt{N}\left\|\widetilde{\theta}-\theta^* + \nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty,\end{aligned}$$

where we use the fact that $|\max_l a_l - \max_l b_l| \leq \max_l |a_l - b_l|$ for any two vectors a and b of the same dimension. Next, we bound $\left\|\widetilde{\theta}-\theta^* + \nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty$. In linear model, we have that

$$\widetilde{\theta}-\theta^* + \nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*) = \bar{\theta} + \widetilde{\Theta}\frac{X_N^\top(y_N - X_N\bar{\theta})}{N} - \theta^* - \Theta\frac{X_N^\top(y_N - X_N\theta^*)}{N},$$

and then,

$$\begin{aligned}&\left\|\widetilde{\theta}-\theta^* + \nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty \\ &= \left\|\bar{\theta} + \widetilde{\Theta}\frac{X_N^\top(y_N - X_N\bar{\theta})}{N} - \theta^* - \Theta\frac{X_N^\top(y_N - X_N\theta^*)}{N}\right\|_\infty \\ &= \left\|\bar{\theta} + \widetilde{\Theta}\frac{X_N^\top(y_N - X_N\bar{\theta})}{N} - \theta^* - \widetilde{\Theta}\frac{X_N^\top(y_N - X_N\theta^*)}{N} + \widetilde{\Theta}\frac{X_N^\top(y_N - X_N\theta^*)}{N} - \Theta\frac{X_N^\top(y_N - X_N\theta^*)}{N}\right\|_\infty \\ &\leq \left\|\left(\widetilde{\Theta}\frac{X_N^\top X_N}{N} - I_d\right)(\bar{\theta}-\theta^*)\right\|_\infty + \left\|\left(\widetilde{\Theta}-\Theta\right)\frac{X_N^\top e_N}{N}\right\|_\infty \\ &\leq \left\|\left|\widetilde{\Theta}\frac{X_N^\top X_N}{N} - I_d\right\|\right|_{\max}\|\bar{\theta}-\theta^*\|_1 + \left\|\left|\widetilde{\Theta}-\Theta\right\|\right|_\infty\left\|\frac{X_N^\top e_N}{N}\right\|_\infty,\end{aligned}$$

where we use the triangle inequality in the second to last inequality and the fact that for any matrix A and vector a with compatible dimensions, $\|Aa\|_\infty \leq |||A|||_{\max}\|a\|_1$ and $\|Aa\|_\infty \leq |||A|||_\infty\|a\|_\infty$, in the last inequality. Further applying the triangle inequality and the fact that for any two matrices A and B with compatible dimensions, $|||AB|||_{\max} \leq |||A|||_\infty |||B|||_{\max}$, we have that

$$\begin{aligned}\left\|\widetilde{\Theta}\frac{X_N^\top X_N}{N}-I_d\right\|_{\max} &= \left\|\widetilde{\Theta}\frac{X_N^\top X_N}{N}-\widetilde{\Theta}\frac{X_1^\top X_1}{n}+\widetilde{\Theta}\frac{X_1^\top X_1}{n}-I_d\right\|_{\max} \\ &\le \left\|\widetilde{\Theta}\left(\frac{X_N^\top X_N}{N}-\frac{X_1^\top X_1}{n}\right)\right\|_{\max}+\left\|\widetilde{\Theta}\frac{X_1^\top X_1}{n}-I_d\right\|_{\max} \\ &\le \left\|\widetilde{\Theta}\right\|_{\infty}\left\|\frac{X_N^\top X_N}{N}-\frac{X_1^\top X_1}{n}\right\|_{\max}+\left\|\widetilde{\Theta}\frac{X_1^\top X_1}{n}-I_d\right\|_{\max}.\end{aligned}$$

Under Assumption (A1), X_N has sub-Gaussian rows. Then, by Lemma 35, if $n \gg s^* \log d$, we have that

$$\left\|\widetilde{\Theta}\right\|_{\infty}=\max_l\left\|\widetilde{\Theta}_l\right\|_1=O_P\left(\sqrt{s^*}\right),$$

$$\left\|\widetilde{\Theta}\frac{X_1^\top X_1}{n}-I_d\right\|_{\max}=O_P\left(\sqrt{\frac{\log d}{n}}\right),$$

and

$$\left\|\widetilde{\Theta}-\Theta\right\|_{\infty}=\max_l\left\|\widetilde{\Theta}_l-\Theta_l\right\|_1=O_P\left(s^*\sqrt{\frac{\log d}{n}}\right).$$

It remains to bound $\left\|\frac{X_N^\top X_N}{N}-\frac{X_1^\top X_1}{n}\right\|_{\max}$ and $\left\|\frac{X_N^\top e_N}{N}\right\|_{\infty}$.

Under Assumptions (A1), each $x_{ij,l}$ is sub-Gaussian, and therefore, the product $x_{ij,l}x_{ij,l'}$ of any two is sub-exponential. By Bernstein's inequality, we have that for any $t>0$,

$$P\left(\left|\frac{(X_N^\top X_N)_{l,l'}}{N}-\Sigma_{l,l'}\right|>t\right)\le 2\exp\left(-cN\left(\frac{t^2}{\Sigma_{l,l'}^2}\wedge\frac{t}{|\Sigma_{l,l'}|}\right)\right),$$

or for any $\delta\in(0,1)$,

$$P\left(\left|\frac{(X_N^\top X_N)_{l,l'}}{N}-\Sigma_{l,l'}\right|>|\Sigma_{l,l'}|\left(\frac{\log\frac{2d^2}{\delta}}{cN}\vee\sqrt{\frac{\log\frac{2d^2}{\delta}}{cN}}\right)\right)\le\frac{\delta}{d^2},$$

for some constant $c>0$. Then, by the union bound, we have that

$$P\left(\left\|\frac{X_N^\top X_N}{N}-\Sigma\right\|_{\max}>\|\Sigma\|_{\max}\left(\frac{\log\frac{2d^2}{\delta}}{cN}\vee\sqrt{\frac{\log\frac{2d^2}{\delta}}{cN}}\right)\right)\le\delta. \tag{40}$$

Similarly, we have that

$$P\left(\left\|\frac{X_1^\top X_1}{n}-\Sigma\right\|_{\max}>\|\Sigma\|_{\max}\left(\frac{\log\frac{2d^2}{\delta}}{cn}\vee\sqrt{\frac{\log\frac{2d^2}{\delta}}{cn}}\right)\right)\le\delta. \tag{41}$$

Then, by the triangle inequality, we have that

$$\begin{aligned}\left\lVert\!\left\lvert \frac{X_N^\top X_N}{N} - \frac{X_1^\top X_1}{n} \right\rvert\!\right\rVert_{\max} &\leq \left\lVert\!\left\lvert \frac{X_1^\top X_1}{n} - \Sigma \right\rvert\!\right\rVert_{\max} + \left\lVert\!\left\lvert \frac{X_N^\top X_N}{N} - \Sigma \right\rvert\!\right\rVert_{\max} \\ &\lesssim \lVert\!\lvert \Sigma \rvert\!\rVert_{\max} \left(\frac{\log \frac{2d^2}{\delta}}{n} \vee \sqrt{\frac{\log \frac{2d^2}{\delta}}{n}} \right) \\ &\lesssim \left(\frac{\log \frac{2d^2}{\delta}}{n} \vee \sqrt{\frac{\log \frac{2d^2}{\delta}}{n}} \right),\end{aligned}$$

with probability at least $1-\delta$, where we use $\lVert\!\lvert \Sigma \rvert\!\rVert_{\max} \leq \lVert\!\lvert \Sigma \rvert\!\rVert_2 = \lambda_{\max}(\Sigma) = O(1)$ under Assumption (A1). This implies that

$$\left\lVert\!\left\lvert \frac{X_N^\top X_N}{N} - \frac{X_1^\top X_1}{n} \right\rvert\!\right\rVert_{\max} = O_P\left(\sqrt{\frac{\log d}{n}} \right).$$

Under Assumptions (A1) and (A2), each $x_{ij,l}$ and e_{ij} are sub-Gaussian, and therefore, their product $x_{ij,l} e_{ij}$ is sub-exponential. Applying Bernstein's inequality, we have that for any $\delta \in (0,1)$,

$$P\left(\left| \frac{(X_N^\top e_N)_l}{N} \right| > \sqrt{\Sigma_{l,l}}\sigma \left(\frac{\log \frac{2d}{\delta}}{cN} \vee \sqrt{\frac{\log \frac{2d}{\delta}}{cN}} \right) \right) \leq \frac{\delta}{d},$$

for some constant $c > 0$. Then, by the union bound, we have that

$$P\left(\left\lVert \frac{X_N^\top e_N}{N} \right\rVert_\infty > \max_l \sqrt{\Sigma_{l,l}}\sigma \left(\frac{\log \frac{2d}{\delta}}{cN} \vee \sqrt{\frac{\log \frac{2d}{\delta}}{cN}} \right) \right) \leq \delta, \tag{42}$$

and then,

$$\left\lVert \frac{X_N^\top e_N}{N} \right\rVert_\infty = O_P\left(\sqrt{\frac{\log d}{N}} \right).$$

Putting all the preceding bounds together, we obtain that

$$\begin{aligned}&\left\lVert \widetilde{\theta} - \theta^* + \nabla^2 \mathcal{L}^*(\theta^*)^{-1} \nabla \mathcal{L}_N(\theta^*) \right\rVert_\infty \\ &\leq \left(\lVert\!\lvert \widetilde{\Theta} \rvert\!\rVert_\infty \left\lVert\!\left\lvert \frac{X_N^\top X_N}{N} - \frac{X_1^\top X_1}{n} \right\rvert\!\right\rVert_{\max} + \left\lVert\!\left\lvert \widetilde{\Theta} \frac{X_1^\top X_1}{n} - I_d \right\rvert\!\right\rVert_{\max} \right) \lVert \bar{\theta} - \theta^* \rVert_1 + \lVert\!\lvert \widetilde{\Theta} - \Theta \rvert\!\rVert_\infty \left\lVert \frac{X_N^\top e_N}{N} \right\rVert_\infty \\ &= \left(O_P\left(\sqrt{s^*} \right) O_P\left(\sqrt{\frac{\log d}{n}} \right) + O_P\left(\sqrt{\frac{\log d}{n}} \right) \right) O_P(r_{\bar{\theta}}) + O_P\left(s^* \sqrt{\frac{\log d}{n}} \right) O_P\left(\sqrt{\frac{\log d}{N}} \right) \\ &= O_P\left(\sqrt{\frac{s^* \log d}{n}} r_{\bar{\theta}} + \frac{s^* \log d}{n\sqrt{k}} \right),\end{aligned}$$

where we assume that $\lVert \bar{\theta} - \theta^* \rVert_1 = O_P(r_{\bar{\theta}})$, and hence,

$$|T - T_0| = O_P\left(r_{\bar{\theta}} \sqrt{s^* k \log d} + \frac{s^* \log d}{\sqrt{n}} \right).$$

Choosing

$$\zeta = \left(r_{\bar{\theta}} \sqrt{s^* k \log d} + \frac{s^* \log d}{\sqrt{n}} \right)^{1-\kappa},$$

with any $\kappa > 0$, we deduce that

$$P\left(|T - T_0| > \zeta\right) = o(1).$$

We also have that

$$\zeta \sqrt{1 \vee \log \frac{d}{\zeta}} = o(1),$$

provided that

$$\left(r_{\bar{\theta}} \sqrt{s^* k \log d} + \frac{s^* \log d}{\sqrt{n}} \right) \log^{1/2+\kappa} d = o(1),$$

which holds if

$$n \gg s^{*2} \log^{3+\kappa} d,$$

and

$$r_{\bar{\theta}} \ll \frac{1}{\sqrt{ks^*} \log^{1+\kappa} d}.$$

■

Lemma 20 *$\widehat{T}$ and T_0 are defined as in* (20) *and* (23) *respectively. In sparse linear model, under Assumptions (A1) and (A2), provided that $n \gg s^* \log d$, we have that*

$$|\widehat{T} - T_0| = O_P\left(\frac{\left(s_0 \sqrt{s^*} + s^*\right) \log d}{\sqrt{n}} \right).$$

Moreover, if $n \gg \left(s_0^2 s^* + s^{*2}\right) \log^{3+\kappa} d$ and for some $\kappa > 0$, then there exists some $\xi > 0$ such that (32) *holds.*

Proof of Lemma 20. By the proof of Lemma 19, we obtain that

$$\begin{aligned}
|\widehat{T} - T_0| &\le \max_{1 \le l \le d} \left| \sqrt{N}(\widehat{\theta} - \theta^*)_l + \sqrt{N} \left(\nabla^2 \mathcal{L}^*(\theta^*)^{-1} \nabla \mathcal{L}_N(\theta^*) \right)_l \right| \\
&= \sqrt{N} \left\| \widehat{\theta} - \theta^* + \nabla^2 \mathcal{L}^*(\theta^*)^{-1} \nabla \mathcal{L}_N(\theta^*) \right\|_\infty \\
&= \sqrt{N} \left\| \widehat{\theta}_L + \widetilde{\Theta} \frac{X_N^\top (y_N - X_N \widehat{\theta}_L)}{N} - \theta^* - \Theta \frac{X_N^\top (y_N - X_N \theta^*)}{N} \right\|_\infty \\
&\le \left\| \widetilde{\Theta} \frac{X_N^\top X_N}{N} - I_d \right\|_{\max} \left\| \widehat{\theta}_L - \theta^* \right\|_1 + \left\| \widetilde{\Theta} - \Theta \right\|_\infty \left\| \frac{X_N^\top e_N}{N} \right\|_\infty \\
&= O_P\left(\sqrt{s^* k \log d} \right) \left\| \widehat{\theta}_L - \theta^* \right\|_1 + O_P\left(\frac{s^* \log d}{\sqrt{n}} \right).
\end{aligned}$$

Since

$$\left\| \widehat{\theta}_L - \theta^* \right\|_1 = O_P\left(s_0 \sqrt{\frac{\log d}{N}} \right),$$

we have that

$$|\widehat{T} - T_0| = O_P\left(\frac{\left(s_0\sqrt{s^*} + s^*\right)\log d}{\sqrt{n}}\right).$$

Choosing

$$\xi = \left(\frac{\left(s_0\sqrt{s^*} + s^*\right)\log d}{\sqrt{n}}\right)^{1-\kappa},$$

with any $\kappa > 0$, we deduce that

$$P\left(|\widehat{T} - T_0| > \xi\right) = o(1).$$

We also have that

$$\xi\sqrt{1 \vee \log\frac{d}{\xi}} = o(1),$$

provided that

$$\left(\frac{\left(s_0\sqrt{s^*} + s^*\right)\log d}{\sqrt{n}}\right)\log^{1/2+\kappa} d = o(1),$$

which holds if

$$n \gg \left(s_0^2 s^* + s^{*2}\right)\log^{3+\kappa} d.$$

■

Lemma 21 *T and T_0 are defined as in* (7) *and* (23) *respectively. In sparse GLM, under Assumptions (B1) and (B2), provided that $\left\|\bar{\theta} - \theta^*\right\|_1 = O_P(r_{\bar{\theta}})$ and $n \gg s_0^2\log^2 d + s^{*2}\log d$, we have that*

$$|T - T_0| = O_P\left(r_{\bar{\theta}}\sqrt{s^* k\log d} + \frac{s^*\log d}{\sqrt{n}}\right).$$

Moreover, if $n \gg (s^{*2} + s_0^2)\log^{3+\kappa} d$ and

$$\left\|\bar{\theta} - \theta^*\right\|_1 \ll \min\left\{\frac{1}{\sqrt{ks^*}s_0\log^{1+\kappa} d}, \frac{1}{\left(nks^*\log^{1+\kappa} d\right)^{1/4}}\right\},$$

for some $\kappa > 0$, then there exists some $\zeta > 0$ such that (29) *holds.*

Proof of Lemma 21. Following the argument in the proof of Lemma 19, we have that

$$\begin{aligned}|T - T_0| &\le \max_{1\le l\le d}\left|\sqrt{N}(\widetilde{\theta}_l - \theta_l^*) + \sqrt{N}\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right)_l\right| \\ &= \sqrt{N}\left\|\widetilde{\theta} - \theta^* + \nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty,\end{aligned}$$

and

$$\begin{aligned}
&\left\|\widetilde{\theta} - \theta^* + \nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty \\
&= \left\|\bar{\theta} - \widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla\mathcal{L}_N(\bar{\theta}) - \theta^* + \Theta\nabla\mathcal{L}_N(\theta^*)\right\|_\infty \\
&= \left\|\bar{\theta} - \widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla\mathcal{L}_N(\bar{\theta}) - \theta^* + \widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla\mathcal{L}_N(\theta^*) - \widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla\mathcal{L}_N(\theta^*) + \Theta\nabla\mathcal{L}_N(\theta^*)\right\|_\infty \\
&\le \left\|\bar{\theta} - \theta^* - \widetilde{\Theta}(\widetilde{\theta}^{(0)})\left(\nabla\mathcal{L}_N(\bar{\theta}) - \nabla\mathcal{L}_N(\theta^*)\right)\right\|_\infty + \left\|\left(\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta\right)\nabla\mathcal{L}_N(\theta^*)\right\|_\infty .
\end{aligned}$$

By Taylor's theorem, we have that

$$\nabla\mathcal{L}_N(\bar{\theta}) - \nabla\mathcal{L}_N(\theta^*) = \int_0^1 \nabla^2\mathcal{L}_N(\theta^* + t(\bar{\theta} - \theta^*))dt(\bar{\theta} - \theta^*), \tag{43}$$

and then,

$$\begin{aligned}
&\left\|\widetilde{\theta} - \theta^* + \nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty \\
&\le \left\|\bar{\theta} - \theta^* - \widetilde{\Theta}(\widetilde{\theta}^{(0)})\int_0^1 \nabla^2\mathcal{L}_N(\theta^* + t(\bar{\theta} - \theta^*))dt(\bar{\theta} - \theta^*)\right\|_\infty + \left\|\left(\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta\right)\nabla\mathcal{L}_N(\theta^*)\right\|_\infty \\
&= \left\|\int_0^1 \left(\widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_N(\theta^* + t(\bar{\theta} - \theta^*)) - I_d\right)dt(\bar{\theta} - \theta^*)\right\|_\infty + \left\|\left(\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta\right)\nabla\mathcal{L}_N(\theta^*)\right\|_\infty \\
&\le \int_0^1 \left\|\left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_N(\theta^* + t(\bar{\theta} - \theta^*)) - I_d\right\|\right\|_{\max} dt \left\|\bar{\theta} - \theta^*\right\|_1 + \left\|\left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta\right\|\right\|_\infty \left\|\nabla\mathcal{L}_N(\theta^*)\right\|_\infty .
\end{aligned}$$

By the triangle inequality, we have that

$$\begin{aligned}
&\left\|\left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_N(\theta^* + t(\bar{\theta} - \theta^*)) - I_d\right\|\right\|_{\max} \\
&= \Big|\Big|\Big|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_N(\theta^* + t(\bar{\theta} - \theta^*)) - \widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_N(\theta^*) + \widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_N(\theta^*) - \widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_1(\theta^*) \\
&\quad + \widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_1(\theta^*) - \widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_1(\widetilde{\theta}^{(0)}) + \widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_1(\widetilde{\theta}^{(0)}) - I_d\Big|\Big|\Big|_{\max} \\
&\le \left\|\left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\left(\nabla^2\mathcal{L}_N(\theta^* + t(\bar{\theta} - \theta^*)) - \nabla^2\mathcal{L}_N(\theta^*)\right)\right\|\right\|_{\max} + \left\|\left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\left(\nabla^2\mathcal{L}_N(\theta^*) - \nabla^2\mathcal{L}_1(\theta^*)\right)\right\|\right\|_{\max} \\
&\quad + \left\|\left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\left(\nabla^2\mathcal{L}_1(\theta^*) - \nabla^2\mathcal{L}_1(\widetilde{\theta}^{(0)})\right)\right\|\right\|_{\max} + \left\|\left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_1(\widetilde{\theta}^{(0)}) - I_d\right\|\right\|_{\max} \\
&\le \left\|\left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\right\|\right\|_\infty \Bigg(\left\|\left\|\nabla^2\mathcal{L}_N(\theta^* + t(\bar{\theta} - \theta^*)) - \nabla^2\mathcal{L}_N(\theta^*)\right\|\right\|_{\max} + \left\|\left\|\nabla^2\mathcal{L}_N(\theta^*) - \nabla^2\mathcal{L}_1(\theta^*)\right\|\right\|_{\max} \\
&\quad + \left\|\left\|\nabla^2\mathcal{L}_1(\theta^*) - \nabla^2\mathcal{L}_1(\widetilde{\theta}^{(0)})\right\|\right\|_{\max}\Bigg) + \left\|\left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_1(\widetilde{\theta}^{(0)}) - I_d\right\|\right\|_{\max} .
\end{aligned}$$

Under Assumption (B1), we have by Taylor's theorem that

$$\begin{aligned}
\left|g''(y_{ij}, x_{ij}^\top(\theta^* + t(\bar{\theta} - \theta^*))) - g''(y_{ij}, x_{ij}^\top\theta^*)\right| &= \left|\int_0^1 g'''(y_{ij}, x_{ij}^\top(\theta^* + st(\bar{\theta} - \theta^*)))ds \cdot t x_{ij}^\top(\bar{\theta} - \theta^*)\right| \\
&\lesssim \left|x_{ij}^\top(\bar{\theta} - \theta^*)\right|,
\end{aligned}$$

and then by the triangle inequality,

$$\begin{aligned}
\left\| \nabla^2 \mathcal{L}_N(\theta^* + t(\bar{\theta} - \theta^*)) - \nabla^2 \mathcal{L}_N(\theta^*) \right\|_{\max} &= \left\| \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} x_{ij}^\top \Big(g''(y_{ij}, x_{ij}^\top(\theta^* + t(\bar{\theta} - \theta^*))) - g''(y_{ij}, x_{ij}^\top \theta^*) \Big) \right\|_{\mathrm{ma}} \\
&\le \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{k} \left\| x_{ij} x_{ij}^\top \Big(g''(y_{ij}, x_{ij}^\top(\theta^* + t(\bar{\theta} - \theta^*))) - g''(y_{ij}, x_{ij}^\top \theta^*) \Big) \right\|_{\mathrm{ma}} \\
&= \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{k} \left\| x_{ij} x_{ij}^\top \right\|_{\max} \Big| g''(y_{ij}, x_{ij}^\top(\theta^* + t(\bar{\theta} - \theta^*))) - g''(y_{ij}, x_{ij}^\top \theta^*) \Big| \\
&\lesssim \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{k} \|x_{ij}\|_\infty^2 \left| x_{ij}^\top (\bar{\theta} - \theta^*) \right| \\
&\le \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{k} \|x_{ij}\|_\infty^3 \|\bar{\theta} - \theta^*\|_1 \\
&\lesssim \left\| \bar{\theta} - \theta^* \right\|_1, \qquad (44)
\end{aligned}$$

where we use that $\|x_{ij}\|_\infty = O(1)$ under Assumption (B2) in the last inequality. Similarly, we have that

$$\left\| \nabla^2 \mathcal{L}_1(\theta^*) - \nabla^2 \mathcal{L}_1(\widetilde{\theta}^{(0)}) \right\|_{\max} \lesssim \|\widetilde{\theta}^{(0)} - \theta^*\|_1 = O_P\left(s_0 \sqrt{\frac{\log d}{n}} \right),$$

by noticing that $\widetilde{\theta}^{(0)}$ is a local Lasso estimator computed using n observations. Note that

$$\left\| \nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}^*(\theta^*) \right\|_{\max} = \left\| \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{k} g''(y_{ij}, x_{ij}^\top \theta^*) x_{ij} x_{ij}^\top - \mathbb{E}[g''(y, x^\top \theta^*) x x^\top] \right\|_{\max},$$

and $g''(y_{ij}, x_{ij}^\top \theta^*) = O(1)$ under Assumption (B1). Then, we have that by Hoeffding's inequality,

$$P\left(\frac{\sum_{i=1}^{n} \sum_{j=1}^{k} g''(y_{ij}, x_{ij}^\top \theta^*) x_{ij,l} x_{ij,l'}}{N} - \mathbb{E}[g''(y, x^\top \theta^*) x_l x_{l'}] > \sqrt{\frac{2 \log(\frac{2d^2}{\delta})}{N}} \right) \le \frac{\delta}{d^2},$$

and by the union bound, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\left\| \nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}^*(\theta^*) \right\|_{\max} \le \sqrt{\frac{2 \log(\frac{2d^2}{\delta})}{N}},$$

which implies that

$$\left\| \nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}^*(\theta^*) \right\|_{\max} = O_P\left(\sqrt{\frac{\log d}{N}} \right). \qquad (45)$$

Similarly, we have that

$$\left\|\left|\nabla^2 \mathcal{L}_1(\theta^*) - \nabla^2 \mathcal{L}^*(\theta^*)\right|\right\|_{\max} = O_P\left(\sqrt{\frac{\log d}{n}}\right),$$

and then, by the triangle inequality,

$$\begin{aligned}\left\|\left|\nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}_1(\theta^*)\right|\right\|_{\max} &\leq \left\|\left|\nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}^*(\theta^*)\right|\right\|_{\max} + \left\|\left|\nabla^2 \mathcal{L}_1(\theta^*) - \nabla^2 \mathcal{L}^*(\theta^*)\right|\right\|_{\max} \\ &= O_P\left(\sqrt{\frac{\log d}{n}}\right).\end{aligned}$$

Note that $\nabla \mathcal{L}_N(\theta^*) = \sum_{i=1}^n \sum_{j=1}^k g'(y_{ij}, x_{ij}^\top \theta^*) x_{ij}/N$ and $g'(y_{ij}, x_{ij}^\top \theta^*) x_{ij,l} = O(1)$ for each $l = 1, \ldots, d$ under Assumptions (B1) and (B2). Then, by Hoeffding's inequality, we have that

$$P\left(|\nabla \mathcal{L}_N(\theta^*)_l| > t\right) \leq 2 \exp\left(-\frac{Nt^2}{c}\right), \tag{46}$$

for any $t > 0$, or

$$P\left(|\nabla \mathcal{L}_N(\theta^*)_l| > \sqrt{\frac{c \log \frac{2d}{\delta}}{N}}\right) \leq \frac{\delta}{d},$$

for any $\delta \in (0, 1)$. By the union bound, we have with probability at least $1 - \delta$ that

$$\|\nabla \mathcal{L}_N(\theta^*)\|_\infty \leq \sqrt{\frac{c \log \frac{2d}{\delta}}{N}},$$

which implies that

$$\|\nabla \mathcal{L}_N(\theta^*)\|_\infty = O_P\left(\sqrt{\frac{\log d}{N}}\right). \tag{47}$$

By Lemma 36, provided that $n \gg s_0^2 \log^2 d + s^{*2} \log d$, we have that

$$\left\|\left|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\right|\right\|_\infty = O_P\left(\sqrt{s^*}\right),$$

$$\left\|\left|\widetilde{\Theta}(\widetilde{\theta}^{(0)}) \nabla^2 \mathcal{L}_1(\widetilde{\theta}^{(0)}) - I_d\right|\right\|_{\max} = O_P\left(\sqrt{\frac{\log d}{n}}\right),$$

and

$$\left\|\left|\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta\right|\right\|_\infty = O_P\left((s_0 + s^*) \sqrt{\frac{\log d}{n}}\right).$$

Putting all the preceding bounds together, we obtain that

$$
\begin{aligned}
&\left|\!\left|\!\left| \widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_N(\theta^*+t(\bar{\theta}-\theta^*))-I_d \right|\!\right|\!\right|_{\max} \\
&= O_P\left(\sqrt{s^*}\right)\left(O_P(r_{\bar{\theta}})+O_P\left(\sqrt{\frac{\log d}{n}}\right)+O_P\left(s_0\sqrt{\frac{\log d}{n}}\right)\right)+O_P\left(\sqrt{\frac{\log d}{n}}\right) \\
&= O_P\left(\sqrt{s^*}\left(r_{\bar{\theta}}+s_0\sqrt{\frac{\log d}{n}}\right)\right),
\end{aligned}
$$

and then,

$$
\begin{aligned}
&\left\|\widetilde{\theta}-\theta^*+\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty \\
&= O_P\left(\sqrt{s^*}\left(r_{\bar{\theta}}+s_0\sqrt{\frac{\log d}{n}}\right)\right)O_P(r_{\bar{\theta}})+O_P\left((s_0+s^*)\sqrt{\frac{\log d}{n}}\right)O_P\left(\sqrt{\frac{\log d}{N}}\right) \\
&= O_P\left(\sqrt{s^*}\left(r_{\bar{\theta}}+s_0\sqrt{\frac{\log d}{n}}\right)r_{\bar{\theta}}+(s_0+s^*)\frac{\log d}{n\sqrt{k}}\right),
\end{aligned}
$$

where we assume that $\left\|\bar{\theta}-\theta^*\right\|_1=O_P(r_{\bar{\theta}})$, and hence,

$$
|T-T_0| = O_P\left(\sqrt{s^*}\left(\sqrt{n}r_{\bar{\theta}}+s_0\sqrt{\log d}\right)\sqrt{k}r_{\bar{\theta}}+(s_0+s^*)\frac{\log d}{\sqrt{n}}\right).
$$

Choosing

$$
\zeta = \left(\sqrt{s^*}\left(\sqrt{n}r_{\bar{\theta}}+s_0\sqrt{\log d}\right)\sqrt{k}r_{\bar{\theta}}+(s_0+s^*)\frac{\log d}{\sqrt{n}}\right)^{1-\kappa},
$$

with any $\kappa>0$, we deduce that

$$
P\left(|T-T_0|>\zeta\right)=o(1).
$$

We also have that

$$
\zeta\sqrt{1\vee\log\frac{d}{\zeta}}=o(1),
$$

provided that

$$
\left(\sqrt{s^*}\left(\sqrt{n}r_{\bar{\theta}}+s_0\sqrt{\log d}\right)\sqrt{k}r_{\bar{\theta}}+(s_0+s^*)\frac{\log d}{\sqrt{n}}\right)\log^{1/2+\kappa} d=o(1),
$$

which holds if

$$
n \gg \left(s^{*2}+s_0^2\right)\log^{3+\kappa} d,
$$

and

$$
r_{\bar{\theta}} \ll \min\left\{\frac{1}{\sqrt{ks^*}s_0\log^{1+\kappa} d},\frac{1}{\left(nks^*\log^{1+\kappa} d\right)^{1/4}}\right\}.
$$

■

Lemma 22 *$\widehat{T}$ and T_0 are defined as in (20) and (23) respectively. In sparse GLM, under Assumptions (B1) and (B2), provided that $n \gg s_0^2 \log^2 d + s^{*2} \log d$, we have that*

$$|\widehat{T} - T_0| = O_P\left(\frac{\left(s_0^2\sqrt{s^*} + s^*\right)\log d}{\sqrt{n}}\right).$$

Moreover, if $n \gg \left(s^4 s^* + s^{*2}\right)\log^{3+\kappa} d$ for some $\kappa > 0$, then there exists some $\xi > 0$ such that (32) holds.

Proof of Lemma 22. By the proof of Lemma 21, we obtain that

$$\begin{aligned}
|\widehat{T} - T_0| &\le \max_{1\le l\le d}\left|\sqrt{N}(\widehat{\theta} - \theta^*)_l + \sqrt{N}\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right)_l\right| \\
&= \sqrt{N}\left\|\widehat{\theta} - \theta^* + \nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty \\
&= \sqrt{N}\left\|\widehat{\theta}_L - \widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla\mathcal{L}_N(\widehat{\theta}_L) - \theta^* + \Theta\nabla\mathcal{L}_N(\theta^*)\right\|_\infty \\
&\le \int_0^1 \left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2\mathcal{L}_N(\theta^* + t(\widehat{\theta}_L - \theta^*)) - I_d\right\|_{\max} dt \left\|\widehat{\theta}_L - \theta^*\right\|_1 + \left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta\right\|_\infty \|\nabla\mathcal{L}_N(\theta^*)\|_\infty \\
&= O_P\left(\sqrt{nks^*}\right)\left\|\widehat{\theta}_L - \theta^*\right\|_1^2 + O_P\left(s_0\sqrt{ks^*\log d}\right)\left\|\widehat{\theta}_L - \theta^*\right\|_1 + O_P\left((s_0 + s^*)\frac{\log d}{\sqrt{n}}\right).
\end{aligned}$$

Since

$$\left\|\widehat{\theta}_L - \theta^*\right\|_1 = O_P\left(s_0\sqrt{\frac{\log d}{N}}\right),$$

we have that

$$|\widehat{T} - T_0| = O_P\left(\frac{\left(s_0^2\sqrt{s^*} + s^*\right)\log d}{\sqrt{n}}\right).$$

Choosing

$$\xi = \left(\frac{\left(s_0^2\sqrt{s^*} + s^*\right)\log d}{\sqrt{n}}\right)^{1-\kappa},$$

with any $\kappa > 0$, we deduce that

$$P\left(|\widehat{T} - T_0| > \xi\right) = o(1).$$

We also have that

$$\xi\sqrt{1 \vee \log\frac{d}{\xi}} = o(1),$$

provided that

$$\left(\frac{\left(s_0^2\sqrt{s^*} + s^*\right)\log d}{\sqrt{n}}\right)\log^{1/2+\kappa} d = o(1),$$

which holds if

$$n \gg \left(s^4 s^* + s^{*2}\right)\log^{3+\kappa} d.$$

■

Lemma 23 *$\overline{\Omega}$ and $\widehat{\Omega}$ are defined as in* (27) *and* (25) *respectively. In sparse linear model, under Assumptions (A1) and (A2), provided that $\left\|\bar{\theta}-\theta^*\right\|_1 = O_P(r_{\bar{\theta}})$, $r_{\bar{\theta}}\sqrt{\log(kd)} \lesssim 1$, $n \gg s^*\log d$, and $k \gtrsim \log^2(dk)\log d$, we have that*

$$\left\|\!\left|\overline{\Omega}-\widehat{\Omega}\right|\!\right\|_{\max} = O_P\left(s^*\left(\sqrt{\frac{\log d}{k}}+\frac{\log^2(dk)\log d}{k}+\sqrt{\log(kd)}r_{\bar{\theta}}+nr_{\bar{\theta}}^2\right)+\sqrt{\frac{s^*\log d}{n}}\right).$$

Moreover, if $n \gg s^*\log^{5+\kappa} d$, $k \gg s^{*2}\log^{5+\kappa} d$, and

$$\left\|\bar{\theta}-\theta^*\right\|_1 \ll \min\left\{\frac{1}{s^*\sqrt{\log(kd)}\log^{2+\kappa} d}, \frac{1}{\sqrt{ns^*}\log^{1+\kappa} d}\right\},$$

for some $\kappa > 0$, then there exists some $u > 0$ such that (30) *holds.*

Proof of Lemma 23. Note by the triangle inequality that

$$\left\|\!\left|\overline{\Omega}-\widehat{\Omega}\right|\!\right\|_{\max} \leq \left\|\!\left|\overline{\Omega}-\Omega_0\right|\!\right\|_{\max} + \left\|\!\left|\widehat{\Omega}-\Omega_0\right|\!\right\|_{\max},$$

where Ω_0 is defined as in (26). First, we bound $\left\|\!\left|\widehat{\Omega}-\Omega_0\right|\!\right\|_{\max}$. With Assumption (E.1) of Chernozhukov et al. (2013) verified for $\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*;Z)$ in the proof of Lemma 15, by the proof of Corollary 3.1 of Chernozhukov et al. (2013), we have that

$$\mathbb{E}\left[\left\|\!\left|\widehat{\Omega}-\Omega_0\right|\!\right\|_{\max}\right] \lesssim \sqrt{\frac{\log d}{N}}+\frac{\log^2(dN)\log d}{N},$$

and then, by Markov's inequality, with probability at least $1-\delta$,

$$\left\|\!\left|\widehat{\Omega}-\Omega_0\right|\!\right\|_{\max} \lesssim \frac{1}{\delta}\left(\sqrt{\frac{\log d}{N}}+\frac{\log^2(dN)\log d}{N}\right),$$

for any $\delta \in (0,1)$, which implies that

$$\left\|\!\left|\widehat{\Omega}-\Omega_0\right|\!\right\|_{\max} = O_P\left(\sqrt{\frac{\log d}{N}}+\frac{\log^2(dN)\log d}{N}\right).$$

Next, we bound $\left\|\!\left|\overline{\Omega}-\Omega_0\right|\!\right\|_{\max}$. By the triangle inequality, we have that

$$\begin{aligned}
&\left\|\!\left|\overline{\Omega}-\Omega_0\right|\!\right\|_{\max} \\
&= \left\|\!\left|\widetilde{\Theta}\left(\frac{1}{k}\sum_{j=1}^k n\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top\right)\widetilde{\Theta}^\top - \Theta\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\Theta\right|\!\right\|_{\max} \\
&\leq \left\|\!\left|\widetilde{\Theta}\left(\frac{1}{k}\sum_{j=1}^k n\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right)\widetilde{\Theta}\right|\!\right\|_{\max} \\
&\quad + \left\|\!\left|\widetilde{\Theta}\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\widetilde{\Theta}^\top - \Theta\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\Theta\right|\!\right\|_{\max} \\
&:= I_1(\bar{\theta}) + I_2.
\end{aligned}$$

To bound $I_1(\bar{\theta})$, we use the fact that for any two matrices A and B with compatible dimensions, $\left\|\!\left\|AB\right\|\!\right\|_{\max} \leq \left\|\!\left\|A\right\|\!\right\|_\infty \left\|\!\left\|B\right\|\!\right\|_{\max}$ and $\left\|\!\left\|AB\right\|\!\right\|_{\max} \leq \left\|\!\left\|A\right\|\!\right\|_{\max} \left\|\!\left\|B\right\|\!\right\|_1$, and obtain that

$$\begin{aligned} I_1(\bar{\theta}) &\leq \left\|\!\left\|\widetilde{\Theta}\right\|\!\right\|_\infty \left\|\!\left\|\frac{1}{k}\sum_{j=1}^k n\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta})\right)^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|\!\right\|_{\max} \left\|\!\left\|\widetilde{\Theta}^\top\right\|\!\right\|_1 \\ &= \left\|\!\left\|\widetilde{\Theta}\right\|\!\right\|_\infty^2 \left\|\!\left\|\frac{1}{k}\sum_{j=1}^k n\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta})\right)^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|\!\right\|_{\max}. \end{aligned}$$

Under Assumption (A1), by Lemma 35, if $n \gg s^* \log d$, we have that

$$\left\|\!\left\|\widetilde{\Theta}\right\|\!\right\|_\infty = \max_l \left\|\widetilde{\Theta}_l\right\|_1 = O_P\left(\sqrt{s^*}\right).$$

Then, applying Lemma 30, we have that

$$\begin{aligned} I_1(\bar{\theta}) &= O_P\left(s^*\right) O_P\left(\sqrt{\frac{\log d}{k}} + \frac{\log^2(dk)\log d}{k} + \sqrt{\log(kd)}r_{\bar{\theta}} + nr_{\bar{\theta}}^2\right) \\ &= O_P\left(s^*\left(\sqrt{\frac{\log d}{k}} + \frac{\log^2(dk)\log d}{k} + \sqrt{\log(kd)}r_{\bar{\theta}} + nr_{\bar{\theta}}^2\right)\right), \end{aligned}$$

under Assumptions (A1) and (A2), provided that $\left\|\bar{\theta} - \theta^*\right\|_1 = O_P(r_{\bar{\theta}})$, $r_{\bar{\theta}}\sqrt{\log(kd)} \lesssim 1$, and $k \gtrsim \log^2(dk)\log d$.

It remains to bound I_2. In linear model, we have that

$$I_2 = \left\|\!\left\|\widetilde{\Theta}\left(\sigma^2\Sigma\right)\widetilde{\Theta}^\top - \Theta\left(\sigma^2\Sigma\right)\Theta\right\|\!\right\|_{\max} = \sigma^2\left\|\!\left\|\widetilde{\Theta}\Sigma\widetilde{\Theta}^\top - \Theta\right\|\!\right\|_{\max},$$

and by the triangle inequality,

$$\begin{aligned} I_2 &= \sigma^2\left\|\!\left\|(\widetilde{\Theta} - \Theta + \Theta)\Sigma(\widetilde{\Theta} - \Theta + \Theta)^\top - \Theta\right\|\!\right\|_{\max} \\ &= \sigma^2\left\|\!\left\|(\widetilde{\Theta} - \Theta)\Sigma(\widetilde{\Theta} - \Theta)^\top + \Theta\Sigma(\widetilde{\Theta} - \Theta)^\top + (\widetilde{\Theta} - \Theta)\Sigma\Theta + \Theta\Sigma\Theta - \Theta\right\|\!\right\|_{\max} \\ &\leq \sigma^2\left\|\!\left\|(\widetilde{\Theta} - \Theta)\Sigma(\widetilde{\Theta} - \Theta)^\top\right\|\!\right\|_{\max} + 2\sigma^2\left\|\!\left\|\widetilde{\Theta} - \Theta\right\|\!\right\|_{\max}. \end{aligned}$$

By Lemma 35, we have that

$$\left\|\!\left\|\widetilde{\Theta} - \Theta\right\|\!\right\|_{\max} \leq \max_l \left\|\widetilde{\Theta}_l - \Theta_l\right\|_2 = O_P\left(\sqrt{\frac{s^*\log d}{n}}\right),$$

and

$$\left\|\!\left\|(\widetilde{\Theta} - \Theta)\Sigma(\widetilde{\Theta} - \Theta)^\top\right\|\!\right\|_{\max} \leq \left\|\!\left\|\Sigma\right\|\!\right\|_2 \max_l \left\|\widetilde{\Theta}_l - \Theta_l\right\|_2^2 = O_P\left(\frac{s^*\log d}{n}\right),$$

where we use that $\left\|\!\left\|\Sigma\right\|\!\right\|_{\max} \leq \left\|\!\left\|\Sigma\right\|\!\right\|_2 = O(1)$ under Assumption (A1). Then, we obtain that

$$I_2 = O_P\left(\frac{s^*\log d}{n}\right) + O_P\left(\sqrt{\frac{s^*\log d}{n}}\right) = O_P\left(\sqrt{\frac{s^*\log d}{n}}\right).$$

Putting all the preceding bounds together, we obtain that

$$\left\|\!\left|\overline{\Omega} - \Omega_0\right|\!\right\|_{\max} = O_P\left(s^*\left(\sqrt{\frac{\log d}{k}} + \frac{\log^2(dk)\log d}{k} + \sqrt{\log(kd)}r_{\bar\theta} + nr_{\bar\theta}^2\right) + \sqrt{\frac{s^*\log d}{n}}\right),$$

and

$$\left\|\!\left|\overline{\Omega} - \widehat{\Omega}\right|\!\right\|_{\max} = O_P\left(s^*\left(\sqrt{\frac{\log d}{k}} + \frac{\log^2(dk)\log d}{k} + \sqrt{\log(kd)}r_{\bar\theta} + nr_{\bar\theta}^2\right) + \sqrt{\frac{s^*\log d}{n}}\right).$$

Choosing

$$u = \left(s^*\sqrt{\frac{\log d}{k}} + \frac{s^*\log^2(dk)\log d}{k} + s^*\sqrt{\log(kd)}r_{\bar\theta} + ns^*r_{\bar\theta}^2 + \sqrt{\frac{s^*\log d}{n}}\right)^{1-\kappa},$$

with any $\kappa > 0$, we deduce that

$$P\left(\left\|\!\left|\overline{\Omega} - \widehat{\Omega}\right|\!\right\|_{\max} > u\right) = o(1).$$

We also have that

$$u^{1/3}\left(1 \vee \log\frac{d}{u}\right)^{2/3} = o(1),$$

provided that

$$\left(s^*\sqrt{\frac{\log d}{k}} + \frac{s^*\log^2(dk)\log d}{k} + s^*\sqrt{\log(kd)}r_{\bar\theta} + ns^*r_{\bar\theta}^2 + \sqrt{\frac{s^*\log d}{n}}\right)\log^{2+\kappa} d = o(1),$$

which holds if

$$n \gg s^*\log^{5+\kappa} d,$$

$$k \gg s^{*2}\log^{5+\kappa} d,$$

and

$$r_{\bar\theta} \ll \min\left\{\frac{1}{s^*\sqrt{\log(kd)}\log^{2+\kappa} d}, \frac{1}{\sqrt{ns^*}\log^{1+\kappa} d}\right\}.$$

■

Lemma 24 $\widehat{\Omega}$ *and* Ω_0 *is defined as in* (25) *and* (26) *respectively. In sparse linear model, under Assumptions (A1) and (A2), we have that*

$$\left\|\!\left|\widehat{\Omega} - \Omega_0\right|\!\right\|_{\max} = O_P\left(\sqrt{\frac{\log d}{N}} + \frac{\log^2(dN)\log d}{N}\right).$$

Moreover, if $N \gg \log^{5+\kappa} d$ *for some* $\kappa > 0$*, then there exists some* $v > 0$ *such that* (31) *holds.*

Proof of Lemma 24. In the proof of Lemma 23, we have shown that

$$\left\|\widehat{\Omega}-\Omega_0\right\|_{\max}=O_P\left(\sqrt{\frac{\log d}{N}}+\frac{\log^2(dN)\log d}{N}\right).$$

Choosing

$$v=\left(\sqrt{\frac{\log d}{N}}+\frac{\log^2(dN)\log d}{N}\right)^{1-\kappa},$$

with any $\kappa>0$, we deduce that

$$P\left(\left\|\widehat{\Omega}-\Omega_0\right\|_{\max}>v\right)=o(1).$$

We also have that

$$v^{1/3}\left(1\vee\log\frac{d}{v}\right)^{2/3}=o(1),$$

provided that

$$\left(\sqrt{\frac{\log d}{N}}+\frac{\log^2(dN)\log d}{N}\right)\log^{2+\kappa}d=o(1),$$

which holds if

$$N\gg\log^{5+\kappa}d.$$

The same result applies to the low-dimensional case as well. ∎

Lemma 25 *$\widetilde{\Omega}$ and $\widehat{\Omega}$ are defined as in* (36) *and* (25) *respectively. In sparse linear model, under Assumptions (A1) and (A2), provided that $\left\|\bar{\theta}-\theta^*\right\|_1=O_P(r_{\bar{\theta}})$, $r_{\bar{\theta}}\sqrt{\log((n+k)d)}\lesssim 1$, $n\gg s^*\log d$, and $n+k\gtrsim\log^3 d$, we have that*

$$\left\|\widetilde{\Omega}-\widehat{\Omega}\right\|_{\max}=O_P\left(s^*\left(\sqrt{\frac{\log d}{n+k}}+\frac{\log^2(d(n+k))\log d}{n+k}+\sqrt{\log((n+k)d)}r_{\bar{\theta}}+\frac{nk}{n+k}r_{\bar{\theta}}^2\right)+\sqrt{\frac{s^*\log d}{n}}\right).$$

Moreover, if $n\gg s^*\log^{5+\kappa}d$, $n+k\gg s^{*2}\log^{5+\kappa}d$, and

$$\left\|\bar{\theta}-\theta^*\right\|_1\ll\min\left\{\frac{1}{s^*\sqrt{\log((n+k)d)}\log^{2+\kappa}d},\frac{1}{\sqrt{s^*}\log^{1+\kappa}d}\sqrt{\frac{1}{n}+\frac{1}{k}}\right\},$$

for some $\kappa>0$, then there exists some $u>0$ such that (37) *holds.*

Proof of Lemma 25. Note by the triangle inequality that

$$\left\|\widetilde{\Omega}-\widehat{\Omega}\right\|_{\max}\le\left\|\widetilde{\Omega}-\Omega_0\right\|_{\max}+\left\|\widehat{\Omega}-\Omega_0\right\|_{\max},$$

where Ω_0 is defined as in (26). By the proof of Lemma 23, we have that

$$\left\|\widehat{\Omega}-\Omega_0\right\|_{\max}=O_P\left(\sqrt{\frac{\log d}{N}}+\frac{\log^2(dN)\log d}{N}\right).$$

Next, we bound $\left\|\widetilde{\Omega}-\Omega_0\right\|_{\max}$ using the same argument as in the proof of Lemma 23. By the triangle inequality, we have that

$$\begin{aligned}
&\left\|\widetilde{\Omega}-\Omega_0\right\|_{\max}\\
&=\Bigg\|\widetilde{\Theta}\frac{1}{n+k-1}\Bigg(\sum_{i=1}^n\left(\nabla\mathcal{L}(\theta;Z_{i1})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}(\theta;Z_{i1})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top\\
&\quad+\sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top\Bigg)\widetilde{\Theta}^\top-\Theta\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\Theta\Bigg\|_{\max}\\
&\le\Bigg\|\widetilde{\Theta}\Bigg(\frac{1}{n+k-1}\Bigg(\sum_{i=1}^n\left(\nabla\mathcal{L}(\theta;Z_{i1})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}(\theta;Z_{i1})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top\\
&\quad+\sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top\Bigg)-\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\Bigg)\widetilde{\Theta}^\top\Bigg\|_{\max}\\
&\quad+\left\|\widetilde{\Theta}\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\widetilde{\Theta}^\top-\Theta\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\Theta\right\|_{\max}\\
&:=I_1'(\bar{\theta})+I_2.
\end{aligned}$$

We have shown in the proof of Lemma 23 that

$$I_2=O_P\left(\sqrt{\frac{s^*\log d}{n}}\right).$$

To bound $I_1'(\bar{\theta})$, we note that

$$\begin{aligned}
I_1'(\bar{\theta})\le\left\|\widetilde{\Theta}\right\|_\infty^2&\Bigg\|\frac{1}{n+k-1}\Bigg(\sum_{i=1}^n\left(\nabla\mathcal{L}(\bar{\theta};Z_{i1})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}(\bar{\theta};Z_{i1})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top\\
&\quad+\sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top\Bigg)-\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\Bigg\|_{\max}.
\end{aligned}$$

Under Assumption (A1), by Lemma 35, if $n\gg s^*\log d$, we have that

$$\left\|\widetilde{\Theta}\right\|_\infty=\max_l\left\|\widetilde{\Theta}_l\right\|_1=O_P\left(\sqrt{s^*}\right).$$

Then, applying Lemma 32, we have that

$$I_1'(\bar{\theta})=O_P\left(s^*\left(\sqrt{\frac{\log d}{n+k}}+\frac{\log^2(d(n+k))\log d}{n+k}+\sqrt{\log((n+k)d)}r_{\bar{\theta}}+\frac{nk}{n+k}r_{\bar{\theta}}^2\right)\right),$$

under Assumptions (A1) and (A2), provided that $\left\|\bar{\theta}-\theta^*\right\|_1=O_P(r_{\bar{\theta}})$, $r_{\bar{\theta}}\sqrt{\log((n+k)d)}\lesssim 1$, and $n+k\gtrsim\log^2(d(n+k))\log d$. Putting all the preceding bounds together, we obtain that

$$\left\|\widetilde{\Omega}-\Omega_0\right\|_{\max}=O_P\left(s^*\left(\sqrt{\frac{\log d}{n+k}}+\frac{\log^2(d(n+k))\log d}{n+k}+\sqrt{\log((n+k)d)}r_{\bar{\theta}}+\frac{nk}{n+k}r_{\bar{\theta}}^2\right)+\sqrt{\frac{s^*\log d}{n}}\right),$$

and

$$\left\|\widetilde{\Omega}-\widehat{\Omega}\right\|_{\max}=O_P\left(s^*\left(\sqrt{\frac{\log d}{n+k}}+\frac{\log^2(d(n+k))\log d}{n+k}+\sqrt{\log((n+k)d)}r_{\bar{\theta}}+\frac{nk}{n+k}r_{\bar{\theta}}^2\right)+\sqrt{\frac{s^*\log d}{n}}\right).$$

Choosing

$$u=\left(s^*\sqrt{\frac{\log d}{n+k}}+\frac{s^*\log^2(d(n+k))\log d}{n+k}+s^*\sqrt{\log((n+k)d)}r_{\bar{\theta}}+\frac{nks^*}{n+k}r_{\bar{\theta}}^2+\sqrt{\frac{s^*\log d}{n}}\right)^{1-\kappa},$$

with any $\kappa>0$, we deduce that

$$P\left(\left\|\widetilde{\Omega}-\widehat{\Omega}\right\|_{\max}>u\right)=o(1).$$

We also have that

$$u^{1/3}\left(1\vee\log\frac{d}{u}\right)^{2/3}=o(1),$$

provided that

$$\left(s^*\sqrt{\frac{\log d}{n+k}}+\frac{s^*\log^2(d(n+k))\log d}{n+k}+s^*\sqrt{\log((n+k)d)}r_{\bar{\theta}}+\frac{nks^*}{n+k}r_{\bar{\theta}}^2+\sqrt{\frac{s^*\log d}{n}}\right)\log^{2+\kappa}d=o(1),$$

which holds if

$$n\gg s^*\log^{5+\kappa}d,$$

$$n+k\gg s^{*2}\log^{5+\kappa}d,$$

and

$$r_{\bar{\theta}}\ll\min\left\{\frac{1}{s^*\sqrt{\log((n+k)d)}\log^{2+\kappa}d},\frac{1}{\sqrt{s^*}\log^{1+\kappa}d}\sqrt{\frac{1}{n}+\frac{1}{k}}\right\}.$$

■

Lemma 26 *$\overline{\Omega}$ and $\widehat{\Omega}$ are defined as in* (38) *and* (25) *respectively. In sparse GLM, under Assumptions (B1)–(B4), provided that $\left\|\bar{\theta}-\theta^*\right\|_1=O_P(r_{\bar{\theta}})$, $r_{\bar{\theta}}\lesssim 1$, $n\gg s_0^2\log^2 d+s^{*2}\log d$, and $k\gtrsim\log d$, we have that*

$$\left\|\overline{\Omega}-\widehat{\Omega}\right\|_{\max}=O_P\left(s^*\left(\sqrt{\frac{\log d}{k}}+\sqrt{\log d}r_{\bar{\theta}}+nr_{\bar{\theta}}^2\right)+\sqrt{\frac{(s_0+s^*)\log d}{n}}+\frac{\log^2(dN)\log d}{N}\right).$$

Moreover, if $n\gg(s_0+s^*)\log^{5+\kappa}d$, $k\gg s^{*2}\log^{5+\kappa}d$, and

$$\left\|\bar{\theta}-\theta^*\right\|_1\ll\min\left\{\frac{1}{s^*\log^{5/2+\kappa}d},\frac{1}{\sqrt{ns^*}\log^{1+\kappa}d}\right\},$$

for some $\kappa>0$, then there exists some $u>0$ such that (30) *holds.*

Proof of Lemma 26. We use the same argument as in the proof of Lemma 23. Note by the triangle inequality that

$$\left\| \overline{\Omega} - \widehat{\Omega} \right\|_{\max} \leq \left\| \overline{\Omega} - \Omega_0 \right\|_{\max} + \left\| \widehat{\Omega} - \Omega_0 \right\|_{\max},$$

where Ω_0 is defined as in (26). First, we bound $\left\| \widehat{\Omega} - \Omega_0 \right\|_{\max}$. With Assumption (E.1) of Chernozhukov et al. (2013) verified for $\nabla^2 \mathcal{L}^*(\theta^*)^{-1} \nabla \mathcal{L}(\theta^*; Z)$ in the proof of Lemma 17, by the proof of Corollary 3.1 of Chernozhukov et al. (2013), we have that

$$\left\| \widehat{\Omega} - \Omega_0 \right\|_{\max} = O_P\left(\sqrt{\frac{\log d}{N}} + \frac{\log^2(dN) \log d}{N} \right).$$

Next, we bound $\left\| \overline{\Omega} - \Omega_0 \right\|_{\max}$. By the triangle inequality, we have that

$$\begin{aligned}
&\left\| \overline{\Omega} - \Omega_0 \right\|_{\max} \\
&= \Bigg\| \widetilde{\Theta}(\widetilde{\theta}^{(0)}) \left(\frac{1}{k} \sum_{j=1}^{k} n \left(\nabla \mathcal{L}_j(\bar{\theta}) - \nabla \mathcal{L}_N(\bar{\theta}) \right) \left(\nabla \mathcal{L}_j(\bar{\theta}) - \nabla \mathcal{L}_N(\bar{\theta}) \right)^\top \right) \widetilde{\Theta}(\widetilde{\theta}^{(0)})^\top \\
&\quad - \Theta \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \Theta \Bigg\|_{\max} \\
&\leq \left\| \widetilde{\Theta}(\widetilde{\theta}^{(0)}) \left(\frac{1}{k} \sum_{j=1}^{k} n \left(\nabla \mathcal{L}_j(\bar{\theta}) - \nabla \mathcal{L}_N(\bar{\theta}) \right) \left(\nabla \mathcal{L}_j(\bar{\theta}) - \nabla \mathcal{L}_N(\bar{\theta}) \right)^\top - \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \right) \widetilde{\Theta}(\widetilde{\theta}^{(0)})^\top \right\|_{\max} \\
&\quad + \left\| \widetilde{\Theta}(\widetilde{\theta}^{(0)}) \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \widetilde{\Theta}(\widetilde{\theta}^{(0)})^\top - \Theta \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \Theta \right\|_{\max} \\
&:= I_1(\bar{\theta}) + I_2.
\end{aligned}$$

Note that

$$\begin{aligned}
&\widetilde{\Theta}(\widetilde{\theta}^{(0)}) \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \widetilde{\Theta}(\widetilde{\theta}^{(0)})^\top \\
&= \left(\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta \right) \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \left(\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta \right)^\top + \Theta \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \left(\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta \right)^\top \\
&\quad + \left(\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta \right) \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \Theta + \Theta \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \Theta.
\end{aligned}$$

By the triangle inequality, we have that

$$\begin{aligned}
I_2 &\leq \left\| \left(\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta \right) \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \left(\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta \right)^\top \right\|_{\max} \\
&\quad + 2 \left\| \Theta \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \left(\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta \right)^\top \right\|_{\max} \\
&\leq \left\| \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \right\|_2 \max_l \left\| \widetilde{\Theta}(\widetilde{\theta}^{(0)})_l - \Theta_l \right\|_2^2 \\
&\quad + 2 \left\| \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top \right] \right\|_2 \max_l \|\Theta_l\|_2 \max_l \left\| \widetilde{\Theta}(\widetilde{\theta}^{(0)})_l - \Theta_l \right\|_2.
\end{aligned}$$

Note that $\max_l \left\|\Theta_l\right\|_2 \leq |\!|\!|\Theta|\!|\!|_2 = O(1)$ under Assumption (B3). By Lemma 36, provided that $n \gg s_0^2 \log^2 d + s^{*2} \log d$, we have that

$$I_2 = O_P\left((s_0 + s^*)\frac{\log d}{n} + \sqrt{s_0 + s^*}\sqrt{\frac{\log d}{n}}\right) = O_P\left(\sqrt{\frac{(s_0 + s^*)\log d}{n}}\right).$$

To bound $I_1(\bar{\theta})$, we note that

$$I_1(\bar{\theta}) \leq \left|\!\left|\!\left|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\right|\!\right|\!\right|_\infty^2 \left\|\frac{1}{k}\sum_{j=1}^{k} n\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta})\right)^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*; Z)\nabla\mathcal{L}(\theta^*; Z)^\top\right]\right\|_{\max}.$$

By Lemma 36, we have that

$$\left|\!\left|\!\left|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\right|\!\right|\!\right|_\infty = O_P\left(\sqrt{s^*}\right).$$

Then, applying Lemma 33, we obtain that

$$I_1(\bar{\theta}) = O_P\left(s^*\left(\sqrt{\frac{\log d}{k}} + \sqrt{\log d}\, r_{\bar{\theta}} + n r_{\bar{\theta}}^2\right)\right),$$

provided that $\left\|\bar{\theta} - \theta^*\right\|_1 = O_P(r_{\bar{\theta}})$, $r_{\bar{\theta}} \lesssim 1$, $n \gtrsim \log d$, and $k \gtrsim \log d$. Putting all the preceding bounds together, we obtain that

$$|\!|\!|\overline{\Omega} - \Omega_0|\!|\!|_{\max} = O_P\left(s^*\left(\sqrt{\frac{\log d}{k}} + \sqrt{\log d}\, r_{\bar{\theta}} + n r_{\bar{\theta}}^2\right) + \sqrt{\frac{(s_0 + s^*)\log d}{n}}\right),$$

and

$$\left|\!\left|\!\left|\overline{\Omega} - \widehat{\Omega}\right|\!\right|\!\right|_{\max} = O_P\left(s^*\left(\sqrt{\frac{\log d}{k}} + \sqrt{\log d}\, r_{\bar{\theta}} + n r_{\bar{\theta}}^2\right) + \sqrt{\frac{(s_0 + s^*)\log d}{n}} + \frac{\log^2(dN)\log d}{N}\right).$$

Choosing

$$u = \left(s^*\sqrt{\frac{\log d}{k}} + s^*\sqrt{\log d}\, r_{\bar{\theta}} + n s^* r_{\bar{\theta}}^2 + \sqrt{\frac{(s_0 + s^*)\log d}{n}} + \frac{\log^2(dN)\log d}{N}\right)^{1-\kappa},$$

with any $\kappa > 0$, we deduce that

$$P\left(\left|\!\left|\!\left|\overline{\Omega} - \widehat{\Omega}\right|\!\right|\!\right|_{\max} > u\right) = o(1).$$

We also have that

$$u^{1/3}\left(1 \vee \log\frac{d}{u}\right)^{2/3} = o(1),$$

provided that

$$\left(s^*\sqrt{\frac{\log d}{k}} + s^*\sqrt{\log d}\, r_{\bar{\theta}} + n s^* r_{\bar{\theta}}^2 + \sqrt{\frac{(s_0 + s^*)\log d}{n}} + \frac{\log^2(dN)\log d}{N}\right)\log^{2+\kappa} d = o(1),$$

which holds if

$$n \gg (s_0 + s^*) \log^{5+\kappa} d,$$

$$k \gg s^{*2} \log^{5+\kappa} d,$$

and

$$r_{\bar{\theta}} \ll \min\left\{\frac{1}{s^* \log^{5/2+\kappa} d}, \frac{1}{\sqrt{ns^*} \log^{1+\kappa} d}\right\}.$$

■

Lemma 27 $\widehat{\Omega}$ *and* Ω_0 *is defined as in* (25) *and* (26) *respectively. In sparse GLM, under Assumptions (B3)–(B4), we have that*

$$\left\|\left\|\widehat{\Omega} - \Omega_0\right\|\right\|_{\max} = O_P\left(\sqrt{\frac{\log d}{N}} + \frac{\log^2(dN)\log d}{N}\right).$$

Moreover, if $N \gg \log^{5+\kappa} d$ *for some* $\kappa > 0$*, then there exists some* $v > 0$ *such that* (31) *holds.*

Proof of Lemma 27. In the proof of Lemma 26, we have shown that

$$\left\|\left\|\widehat{\Omega} - \Omega_0\right\|\right\|_{\max} = O_P\left(\sqrt{\frac{\log d}{N}} + \frac{\log^2(dN)\log d}{N}\right).$$

Choosing

$$v = \left(\sqrt{\frac{\log d}{N}} + \frac{\log^2(dN)\log d}{N}\right)^{1-\kappa},$$

with any $\kappa > 0$, we deduce that

$$P\left(\left\|\left\|\widehat{\Omega} - \Omega_0\right\|\right\|_{\max} > v\right) = o(1).$$

We also have that

$$v^{1/3}\left(1 \vee \log\frac{d}{v}\right)^{2/3} = o(1),$$

provided that

$$\left(\sqrt{\frac{\log d}{N}} + \frac{\log^2(dN)\log d}{N}\right)\log^{2+\kappa} d = o(1),$$

which holds if

$$N \gg \log^{5+\kappa} d.$$

The same result applies to the low-dimensional case as well. ■

Lemma 28 *$\widetilde{\Omega}$ and $\widehat{\Omega}$ are defined as in (39) and (25) respectively. In sparse GLM, under Assumptions (B1)–(B4), provided that $\left\|\bar{\theta}-\theta^*\right\|_1 = O_P(r_{\bar{\theta}})$, $r_{\bar{\theta}} \lesssim 1$, and $n \gg s_0^2 \log^2 d + s^{*2} \log d$, we have that*

$$\left\|\!\left|\widetilde{\Omega}-\widehat{\Omega}\right|\!\right\|_{\max} = O_P\Bigg(s^*\left(\sqrt{\frac{\log d}{n+k}} + \frac{n+k\sqrt{\log d}+k^{3/4}\log^{3/4} d}{n+k} r_{\bar{\theta}} + \frac{nk}{n+k} r_{\bar{\theta}}^2\right) + \sqrt{\frac{(s_0+s^*)\log d}{n}} \\ + \frac{\log^2(dN)\log d}{N}\Bigg).$$

Moreover, if $n \gg (s_0+s^*)\log^{5+\kappa} d + s_0^2\log^2 d + s^{*2}\log d$, $n+k \gg s^{*2}\log^{5+\kappa} d$, and

$$\left\|\bar{\theta}-\theta^*\right\|_1 \ll \min\left\{\frac{n+k}{s^*\left(n+k\sqrt{\log d}+k^{3/4}\log^{3/4} d\right)\log^{2+\kappa} d}, \frac{1}{\sqrt{s^*}\log^{1+\kappa} d}\sqrt{\frac{1}{n}+\frac{1}{k}}\right\},$$

for some $\kappa > 0$, then there exists some $u > 0$ such that (37) holds.

Proof of Lemma 28. Note by the triangle inequality that

$$\left\|\!\left|\widetilde{\Omega}-\widehat{\Omega}\right|\!\right\|_{\max} \le \left\|\!\left|\widetilde{\Omega}-\Omega_0\right|\!\right\|_{\max} + \left\|\!\left|\widehat{\Omega}-\Omega_0\right|\!\right\|_{\max},$$

where Ω_0 is defined as in (26). By the proof of Lemma 26, we have that

$$\left\|\!\left|\widehat{\Omega}-\Omega_0\right|\!\right\|_{\max} = O_P\left(\sqrt{\frac{\log d}{N}} + \frac{\log^2(dN)\log d}{N}\right).$$

Next, we bound $\left\|\!\left|\widetilde{\Omega}-\Omega_0\right|\!\right\|_{\max}$ using the same argument as in the proof of Lemma 26. By the triangle inequality, we have that

$$\begin{aligned}
&\left\|\!\left|\widetilde{\Omega}-\Omega_0\right|\!\right\|_{\max} \\
&= \Bigg\|\!\Bigg|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\frac{1}{n+k-1}\Bigg(\sum_{i=1}^n \left(\nabla\mathcal{L}(\theta;Z_{i1})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}(\theta;Z_{i1})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top \\
&\quad + \sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top\Bigg)\widetilde{\Theta}(\widetilde{\theta}^{(0)})^\top - \Theta\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\Theta\Bigg|\!\Bigg\|_{\max} \\
&\le \Bigg\|\!\Bigg|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\Bigg(\frac{1}{n+k-1}\Bigg(\sum_{i=1}^n \left(\nabla\mathcal{L}(\theta;Z_{i1})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}(\theta;Z_{i1})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top \\
&\quad + \sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top\Bigg) - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\Bigg)\widetilde{\Theta}(\widetilde{\theta}^{(0)})^\top\Bigg|\!\Bigg\|_{\max} \\
&\quad + \left\|\!\left|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\widetilde{\Theta}(\widetilde{\theta}^{(0)})^\top - \Theta\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\Theta\right|\!\right\|_{\max} \\
&:= I_1'(\bar{\theta}) + I_2.
\end{aligned}$$

We have shown in the proof of Lemma 26 that

$$I_2 = O_P\left(\sqrt{\frac{(s_0+s^*)\log d}{n}}\right).$$

To bound $I_1'(\bar\theta)$, we note that

$$\begin{aligned}I_1'(\bar\theta) \le \left|\!\left|\!\left|\widetilde\Theta(\widetilde\theta^{(0)})\right|\!\right|\!\right|_\infty^2 \Bigg\|\frac{1}{n+k-1}\Bigg(&\sum_{i=1}^n \left(\nabla\mathcal{L}(\bar\theta;Z_{i1})-\nabla\mathcal{L}_N(\bar\theta)\right)\left(\nabla\mathcal{L}(\bar\theta;Z_{i1})-\nabla\mathcal{L}_N(\bar\theta)\right)^\top \\ &+\sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\bar\theta)-\nabla\mathcal{L}_N(\bar\theta)\right)\left(\nabla\mathcal{L}_j(\bar\theta)-\nabla\mathcal{L}_N(\bar\theta)\right)^\top\Bigg) - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\Bigg\|_{\max}.\end{aligned}$$

By Lemma 36, provided that $n \gg s_0^2\log^2 d + s^{*2}\log d$, we have that

$$\left|\!\left|\!\left|\widetilde\Theta(\widetilde\theta^{(0)})\right|\!\right|\!\right|_\infty = O_P\left(\sqrt{s^*}\right).$$

Then, applying Lemma 34, we have that

$$I_1'(\bar\theta) = O_P\left(s^*\left(\sqrt{\frac{\log d}{n+k}} + \frac{n+k\sqrt{\log d}+k^{3/4}\log^{3/4} d}{n+k} r_{\bar\theta} + \frac{nk}{n+k} r_{\bar\theta}^2\right)\right),$$

under Assumptions (B1)–(B3), provided that $\left\|\bar\theta-\theta^*\right\|_1 = O_P(r_{\bar\theta})$, $r_{\bar\theta} \lesssim 1$, and $n+k \gtrsim \log d$.

Putting all the preceding bounds together, we obtain that

$$\left|\!\left|\!\left|\widetilde\Omega-\Omega_0\right|\!\right|\!\right|_{\max} = O_P\left(s^*\left(\sqrt{\frac{\log d}{n+k}} + \frac{n+k\sqrt{\log d}+k^{3/4}\log^{3/4} d}{n+k} r_{\bar\theta} + \frac{nk}{n+k} r_{\bar\theta}^2\right) + \sqrt{\frac{(s_0+s^*)\log d}{n}}\right),$$

and

$$\begin{aligned}\left|\!\left|\!\left|\widetilde\Omega-\widehat\Omega\right|\!\right|\!\right|_{\max} = O_P\Bigg(&s^*\left(\sqrt{\frac{\log d}{n+k}} + \frac{n+k\sqrt{\log d}+k^{3/4}\log^{3/4} d}{n+k} r_{\bar\theta} + \frac{nk}{n+k} r_{\bar\theta}^2\right) + \sqrt{\frac{(s_0+s^*)\log d}{n}} \\ &+ \frac{\log^2(dN)\log d}{N}\Bigg).\end{aligned}$$

Choosing

$$u = \left(s^*\sqrt{\frac{\log d}{n+k}} + \frac{n+k\sqrt{\log d}+k^{3/4}\log^{3/4} d}{n+k} s^* r_{\bar\theta} + \frac{nks^*}{n+k} r_{\bar\theta}^2 + \sqrt{\frac{(s_0+s^*)\log d}{n}} + \frac{\log^2(dN)\log d}{N}\right)^{1-\kappa},$$

with any $\kappa > 0$, we deduce that

$$P\left(\left|\!\left|\!\left|\widetilde\Omega-\widehat\Omega\right|\!\right|\!\right|_{\max} > u\right) = o(1).$$

We also have that

$$u^{1/3}\left(1 \vee \log\frac{d}{u}\right)^{2/3} = o(1),$$

provided that

$$\left(s^*\sqrt{\frac{\log d}{n+k}}+\frac{n+k\sqrt{\log d}+k^{3/4}\log^{3/4} d}{n+k}s^* r_{\bar\theta}+\frac{nks^*}{n+k}r_{\bar\theta}^2+\sqrt{\frac{(s_0+s^*)\log d}{n}}+\frac{\log^2(dN)\log d}{N}\right)\log^{2+\kappa} d$$
$$=o(1),$$

which holds if

$$n \gg (s_0+s^*)\log^{5+\kappa} d + s_0^2\log^2 d + s^{*2}\log d,$$
$$n+k \gg s^{*2}\log^{5+\kappa} d,$$

and

$$r_{\bar\theta} \ll \min\left\{\frac{n+k}{s^*\left(n+k\sqrt{\log d}+k^{3/4}\log^{3/4} d\right)\log^{2+\kappa} d}, \frac{1}{\sqrt{s^*}\log^{1+\kappa} d}\sqrt{\frac{1}{n}+\frac{1}{k}}\right\}.$$

■

Lemma 29 *For any θ, we have that*

$$\left\|\left\|\frac{1}{k}\sum_{j=1}^k n\left(\nabla\mathcal{L}_j(\theta)-\nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}_j(\theta)-\nabla\mathcal{L}_N(\theta)\right)^\top-\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|\right\|_{\max} \le U_1(\theta)+U_2+U_3(\theta),$$

where

$$U_1(\theta) := \left\|\left\|\frac{1}{k}\sum_{j=1}^k n\left(\nabla\mathcal{L}_j(\theta)-\nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_j(\theta)-\nabla\mathcal{L}^*(\theta)\right)^\top - n\nabla\mathcal{L}_j(\theta^*)\nabla\mathcal{L}_j(\theta^*)^\top\right\|\right\|_{\max},$$

$$U_2 := \left\|\left\|\frac{1}{k}\sum_{j=1}^k n\nabla\mathcal{L}_j(\theta^*)\nabla\mathcal{L}_j(\theta^*)^\top-\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|\right\|_{\max},$$

and

$$U_3(\theta) := n\left\|\nabla\mathcal{L}_N(\theta)-\nabla\mathcal{L}^*(\theta)\right\|_\infty^2.$$

Lemma 29 is the same as Lemma F.1 of Yu et al. (2020b). We omit the proof.

Lemma 30 *In sparse linear model, under Assumptions (A1) and (A2), provided that $\left\|\bar\theta-\theta^*\right\|_1 = O_P(r_{\bar\theta})$, we have that*

$$\left\|\left\|\frac{1}{k}\sum_{j=1}^k n\left(\nabla\mathcal{L}_j(\bar\theta)-\nabla\mathcal{L}_N(\bar\theta)\right)\left(\nabla\mathcal{L}_j(\bar\theta)-\nabla\mathcal{L}_N(\bar\theta)\right)^\top-\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|\right\|_{\max}$$
$$=O_P\left(\sqrt{\frac{\log d}{k}}+\frac{\log^2(dk)\log d}{k}+\left(1+\left(\frac{\log d}{k}\right)^{1/4}+\sqrt{\frac{\log^2(dk)\log d}{k}}\right)\sqrt{\log(kd)}r_{\bar\theta}\right.$$
$$\left.+\left(n+\sqrt{\frac{n\log d}{k}}+\log(kd)\right)r_{\bar\theta}^2\right).$$

Proof of Lemma 30. By Lemma 29, it suffices to bound $U_1(\bar{\theta})$, U_2, and $U_3(\bar{\theta})$. We begin by bounding U_2. In linear model, we have that

$$U_2 = \left\| \frac{1}{k} \sum_{j=1}^{k} n \left(\frac{X_j^\top e_j}{n} \right) \left(\frac{X_j^\top e_j}{n} \right)^\top - \sigma^2 \Sigma \right\|_{\max} .$$

Note that

$$\mathbb{E}\left[\left(\frac{(X_j^\top e_j)_l}{\sqrt{n}} \right)^2 \right] = \mathbb{E}\left[\frac{\sum_{i=1}^n X_{ij,l}^2 e_{ij}^2}{n} \right] = \sigma^2 \Sigma_{l,l}$$

is bounded away from zero, under Assumptions (A1) and (A2). Also, using same argument for obtaining (42), we have that for any $t > 0$,

$$P\left(\left| \frac{(X_j^\top e_j)_l}{n} \right| > t \right) \le 2 \exp\left(-cn \left(\frac{t^2}{\Sigma_{l,l}\sigma^2} \wedge \frac{t}{\sqrt{\Sigma_{l,l}}\sigma} \right) \right),$$

and then,

$$P\left(\left| \frac{(X_j^\top e_j)_l}{\sqrt{n}} \right| > t \right) \le 2 \exp\left(-c \left(\frac{t^2}{\Sigma_{l,l}\sigma^2} \wedge \frac{t\sqrt{n}}{\sqrt{\Sigma_{l,l}}\sigma} \right) \right) \le C \exp\left(-c' t \right),$$

for some positive constants c, c', and C, that is, $(X_j^\top e_j)_l/\sqrt{n}$ is sub-exponential with O(1) ψ_1-norm for each (j, l). Then, by the proof of Corollary 3.1 of Chernozhukov et al. (2013), we have that

$$\mathbb{E}[U_2] = \mathbb{E}\left[\left\| \frac{1}{k} \sum_{j=1}^{k} \left(\frac{X_j^\top e_j}{\sqrt{n}} \right) \left(\frac{X_j^\top e_j}{\sqrt{n}} \right)^\top - \sigma^2 \Sigma \right\|_{\max} \right] \lesssim \sqrt{\frac{\log d}{k}} + \frac{\log^2(dk) \log d}{k},$$

and then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$U_2 \lesssim \frac{1}{\delta} \left(\sqrt{\frac{\log d}{k}} + \frac{\log^2(dk) \log d}{k} \right),$$

by Markov's inequality, which implies that

$$U_2 = O_P\left(\sqrt{\frac{\log d}{k}} + \frac{\log^2(dk) \log d}{k} \right).$$

Next, we bound $U_3(\bar{\theta})$. By the triangle inequality and the fact that for any matrix A and vector a with compatible dimensions, $\|Aa\|_\infty \leq \|\!|A|\!\|_{\max} \|a\|_1$, we have that

$$\begin{aligned}
\left\|\nabla\mathcal{L}_N(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta})\right\|_\infty &\leq \left\|\nabla\mathcal{L}_N(\bar{\theta}) - \nabla\mathcal{L}_N(\theta^*)\right\|_\infty + \|\nabla\mathcal{L}_N(\theta^*)\|_\infty + \left\|\nabla\mathcal{L}^*(\bar{\theta})\right\|_\infty \\
&= \left\|\frac{X_N^\top(X_N\bar{\theta} - y_N)}{N} - \frac{X_N^\top(X_N\theta^* - y_N)}{N}\right\|_\infty + \left\|\frac{X_N^\top(X_N\theta^* - y_N)}{N}\right\|_\infty + \left\|\Sigma(\bar{\theta} - \theta^*)\right\|_\infty \\
&= \left\|\frac{X_N^\top X_N}{N}(\bar{\theta} - \theta^*)\right\|_\infty + \left\|\frac{X_N^\top e_N}{N}\right\|_\infty + \left\|\Sigma(\bar{\theta} - \theta^*)\right\|_\infty \\
&\leq \left\|\!\left|\frac{X_N^\top X_N}{N}\right|\!\right\|_{\max} \left\|\bar{\theta} - \theta^*\right\|_1 + \left\|\frac{X_N^\top e_N}{N}\right\|_\infty + \|\!|\Sigma|\!\|_{\max} \left\|\bar{\theta} - \theta^*\right\|_1 \\
&\lesssim \left\|\!\left|\frac{X_N^\top X_N}{N} - \Sigma\right|\!\right\|_{\max} \left\|\bar{\theta} - \theta^*\right\|_1 + \left\|\frac{X_N^\top e_N}{N}\right\|_\infty + \|\!|\Sigma|\!\|_{\max} \left\|\bar{\theta} - \theta^*\right\|_1 .
\end{aligned}$$

By (40) and (42), we have that

$$\left\|\!\left|\frac{X_N^\top X_N}{N} - \Sigma\right|\!\right\|_{\max} \leq \|\!|\Sigma|\!\|_{\max} \left(\frac{\log\frac{2d^2}{\delta}}{cN} \vee \sqrt{\frac{\log\frac{2d^2}{\delta}}{cN}}\right) = O_P\left(\sqrt{\frac{\log d}{N}}\right),$$

and

$$\left\|\frac{X_N^\top e_N}{N}\right\|_\infty \leq \max_l \sqrt{\Sigma_{l,l}}\sigma \left(\frac{\log\frac{2d}{\delta}}{cN} \vee \sqrt{\frac{\log\frac{2d}{\delta}}{cN}}\right) = O_P\left(\sqrt{\frac{\log d}{N}}\right),$$

where $\max_l \sqrt{\Sigma_{l,l}} \leq \|\!|\Sigma|\!\|_{\max} = O(1)$ under Assumption (A1). Then, assuming that $\left\|\bar{\theta} - \theta^*\right\|_1 = O_P(r_{\bar{\theta}})$, we have that

$$\begin{aligned}
\left\|\nabla\mathcal{L}_N(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta})\right\|_\infty &= \left(O(1) + O_P\left(\sqrt{\frac{\log d}{N}}\right)\right) O_P(r_{\bar{\theta}}) + O_P\left(\sqrt{\frac{\log d}{N}}\right) \\
&= O_P\left(\left(1 + \sqrt{\frac{\log d}{N}}\right) r_{\bar{\theta}} + \sqrt{\frac{\log d}{N}}\right),
\end{aligned}$$

and then,

$$U_3(\bar{\theta}) = O_P\left(\left(1 + \sqrt{\frac{\log d}{N}}\right) n r_{\bar{\theta}}^2 + \frac{\log d}{k}\right).$$

Lastly, we bound $U_1(\bar{\theta})$. We write $\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta})$ as $\big(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_j(\theta^*)\big) + \nabla\mathcal{L}_j(\theta^*)$, and obtain by the triangle inequality that

$$\begin{aligned}
U_1(\bar{\theta}) \leq & \left|\!\left|\!\left| \frac{1}{k}\sum_{j=1}^{k} n \left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_j(\theta^*)\right)\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_j(\theta^*)\right)^\top \right|\!\right|\!\right|_{\max} \\
& + \left|\!\left|\!\left| \frac{1}{k}\sum_{j=1}^{k} n \nabla\mathcal{L}_j(\theta^*)\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_j(\theta^*)\right)^\top \right|\!\right|\!\right|_{\max} \\
& + \left|\!\left|\!\left| \frac{1}{k}\sum_{j=1}^{k} n \left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_j(\theta^*)\right)\nabla\mathcal{L}_j(\theta^*)^\top \right|\!\right|\!\right|_{\max} \\
= & \left|\!\left|\!\left| \frac{1}{k}\sum_{j=1}^{k} n \left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_j(\theta^*)\right)\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_j(\theta^*)\right)^\top \right|\!\right|\!\right|_{\max} \\
& + 2\left|\!\left|\!\left| \frac{1}{k}\sum_{j=1}^{k} n \nabla\mathcal{L}_j(\theta^*)\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_j(\theta^*)\right)^\top \right|\!\right|\!\right|_{\max} \\
:= & \, U_{11}(\bar{\theta}) + 2U_{12}(\bar{\theta}).
\end{aligned}$$

To bound $U_{12}(\bar{\theta})$, we first define an inner product $\langle A, B\rangle = |\!|\!| AB^\top |\!|\!|_{\max}$ for any $A, B \in \mathbb{R}^{d\times k}$, the validity of which is easy to check. We then apply Cauchy-Schwarz inequality on $\langle A, B\rangle$ with

$$A = \sqrt{\frac{n}{k}}\begin{bmatrix}\nabla\mathcal{L}_1(\theta^*) & \dots & \nabla\mathcal{L}_k(\theta^*)\end{bmatrix}$$

and

$$B = \sqrt{\frac{n}{k}}\begin{bmatrix}\nabla\mathcal{L}_1(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_1(\theta^*) & \dots & \nabla\mathcal{L}_k(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_k(\theta^*))\end{bmatrix}$$

and obtain that

$$\begin{aligned}
U_{12}(\bar{\theta}) \leq & \left|\!\left|\!\left| \frac{1}{k}\sum_{j=1}^{k} n \nabla\mathcal{L}_j(\theta^*)\nabla\mathcal{L}_j(\theta^*)^\top \right|\!\right|\!\right|_{\max}^{1/2} \\
& \cdot \left|\!\left|\!\left| \frac{1}{k}\sum_{j=1}^{k} n \left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_j(\theta^*)\right)\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_j(\theta^*)\right)^\top \right|\!\right|\!\right|_{\max}^{1/2} \\
= & \left|\!\left|\!\left| \frac{1}{k}\sum_{j=1}^{k} n \nabla\mathcal{L}_j(\theta^*)\nabla\mathcal{L}_j(\theta^*)^\top \right|\!\right|\!\right|_{\max}^{1/2} U_{11}(\bar{\theta})^{1/2}.
\end{aligned}$$

By the triangle inequality, we have that

$$\begin{aligned}
&\left\|\!\left\| \frac{1}{k}\sum_{j=1}^{k} n\nabla\mathcal{L}_j(\theta^*)\nabla\mathcal{L}_j(\theta^*)^\top \right\|\!\right\|_{\max} \\
&\le \left\|\!\left\| \frac{1}{k}\sum_{j=1}^{k} n\nabla\mathcal{L}_j(\theta^*)\nabla\mathcal{L}_j(\theta^*)^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right] \right\|\!\right\|_{\max} + \left\|\!\left\| \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right] \right\|\!\right\|_{\max} \\
&= U_2 + \sigma^2 \left\|\!\left|\Sigma\right|\!\right\|_{\max} \\
&= O_P\left(1 + \sqrt{\frac{\log d}{k}} + \frac{\log^2(dk)\log d}{k}\right).
\end{aligned}$$

It remains to bound $U_{11}(\bar{\theta})$. Note that

$$\begin{aligned}
\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_j(\theta^*) &= \frac{X_j^\top(X_j\bar{\theta} - y_j)}{n} - \Sigma(\bar{\theta}-\theta^*) + \frac{X_j^\top(X_j\theta^* - y_j)}{n} \\
&= \left(\frac{X_j^\top X_j}{n} - \Sigma\right)(\bar{\theta}-\theta^*).
\end{aligned}$$

Then, we have that

$$\begin{aligned}
U_{11}(\bar{\theta}) &= \left\|\!\left\| \frac{1}{k}\sum_{j=1}^{k} n\left(\frac{X_j^\top X_j}{n} - \Sigma\right)(\bar{\theta}-\theta^*)(\bar{\theta}-\theta^*)^\top\left(\frac{X_j^\top X_j}{n} - \Sigma\right) \right\|\!\right\|_{\max} \\
&\le \frac{1}{k}\sum_{j=1}^{k} n\left\|\!\left\| \left(\frac{X_j^\top X_j}{n} - \Sigma\right)(\bar{\theta}-\theta^*)(\bar{\theta}-\theta^*)^\top\left(\frac{X_j^\top X_j}{n} - \Sigma\right) \right\|\!\right\|_{\max} \\
&= \frac{1}{k}\sum_{j=1}^{k} n\left\|\!\left\| \left(\frac{X_j^\top X_j}{n} - \Sigma\right)(\bar{\theta}-\theta^*) \right\|\!\right\|_{\infty}^2 \\
&\le \frac{1}{k}\sum_{j=1}^{k} n\left\|\!\left\| \frac{X_j^\top X_j}{n} - \Sigma \right\|\!\right\|_{\max}^2 \left\|\bar{\theta}-\theta^*\right\|_1^2,
\end{aligned}$$

where we use the triangle inequality and the fact that $\left\|\!\left| aa^\top \right|\!\right\|_{\max} = \|a\|_\infty^2$ for any vector a, and $\|Aa\|_\infty \le \left\|\!\left|A\right|\!\right\|_{\max}\|a\|_1$ for any matrix A and vector a with compatible dimensions. By (41), we have that

$$P\left(\left\|\!\left\| \frac{X_j^\top X_j}{n} - \Sigma \right\|\!\right\|_{\max} > \left\|\!\left|\Sigma\right|\!\right\|_{\max}\left(\frac{\log\frac{2kd^2}{\delta}}{cn} \vee \sqrt{\frac{\log\frac{2kd^2}{\delta}}{cn}}\right)\right) \le \frac{\delta}{k},$$

and then, by the union bound,

$$P\left(\max_j \left\|\!\left\| \frac{X_j^\top X_j}{n} - \Sigma \right\|\!\right\|_{\max} > \left\|\!\left|\Sigma\right|\!\right\|_{\max}\left(\frac{\log\frac{2kd^2}{\delta}}{cn} \vee \sqrt{\frac{\log\frac{2kd^2}{\delta}}{cn}}\right)\right) \le \delta,$$

which implies that

$$\max_j \left\| \frac{X_j^\top X_j}{n} - \Sigma \right\|_{\max} = O_P\left(\sqrt{\frac{\log(kd)}{n}}\right).$$

Putting all the preceding bounds together, we obtain that

$$U_{11}(\bar{\theta}) = O_P\left(\log(kd) r_{\bar{\theta}}^2\right),$$

$$U_{12}(\bar{\theta}) = O_P\left(\left(1 + \left(\frac{\log d}{k}\right)^{1/4} + \sqrt{\frac{\log^2(dk)\log d}{k}}\right)\sqrt{\log(kd)} r_{\bar{\theta}}\right),$$

$$U_1(\bar{\theta}) = O_P\left(\left(1 + \left(\frac{\log d}{k}\right)^{1/4} + \sqrt{\frac{\log^2(dk)\log d}{k}}\right)\sqrt{\log(kd)} r_{\bar{\theta}} + \log(kd) r_{\bar{\theta}}^2\right),$$

and finally,

$$\begin{aligned}
&\left\| \frac{1}{k}\sum_{j=1}^k n\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta})\right)^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*; Z)\nabla\mathcal{L}(\theta^*; Z)^\top\right] \right\|_2 \\
&= O_P\left(\sqrt{\frac{\log d}{k}} + \frac{\log^2(dk)\log d}{k} + \left(1 + \left(\frac{\log d}{k}\right)^{1/4} + \sqrt{\frac{\log^2(dk)\log d}{k}}\right)\sqrt{\log(kd)} r_{\bar{\theta}}\right. \\
&\quad \left. + \left(n + \sqrt{\frac{n\log d}{k}} + \log(kd)\right) r_{\bar{\theta}}^2\right).
\end{aligned}$$

■

Lemma 31 *For any θ, we have that*

$$\begin{aligned}
&\Bigg\| \frac{1}{n+k-1}\Bigg(\sum_{i=1}^n \left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}_N(\theta)\right)^\top \\
&\quad + \sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)^\top\Bigg) - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*; Z)\nabla\mathcal{L}(\theta^*; Z)^\top\right] \Bigg\|_{\max} \\
&\le V_1(\theta) + V_1'(\theta) + V_2 + V_2' + V_3(\theta),
\end{aligned}$$

where

$$V_1(\theta) := \frac{k-1}{n+k-1}\left\| \frac{1}{k-1}\sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top - n\nabla\mathcal{L}_j(\theta^*)\nabla\mathcal{L}_j(\theta^*)^\top \right\|_{\max},$$

$$V_1'(\theta) := \frac{n}{n+k-1}\left\| \frac{1}{n}\sum_{i=1}^n \left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}^*(\theta)\right)^\top - \nabla\mathcal{L}(\theta^*; Z_{i1})\nabla\mathcal{L}(\theta^*; Z_{i1})^\top \right\|_{\max},$$

$$V_2 := \frac{k-1}{n+k-1} \left\|\left\| \frac{1}{k-1} \sum_{j=2}^{k} n \nabla \mathcal{L}_j(\theta^*) \nabla \mathcal{L}_j(\theta^*)^\top - \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top\right] \right\|\right\|_{\max},$$

$$V_2' := \frac{n}{n+k-1} \left\|\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \mathcal{L}(\theta^*; Z_{i1}) \nabla \mathcal{L}(\theta^*; Z_{i1})^\top - \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top\right] \right\|\right\|_{\max},$$

and

$$V_3(\theta) := \frac{nk}{n+k-1} \left\| \nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right\|_\infty^2 .$$

Lemma 31 is the same as Lemma F.3 of Yu et al. (2020b). We omit the proof.

Lemma 32 *In sparse linear model, under Assumptions (A1) and (A2), provided that $\left\|\bar{\theta} - \theta^*\right\|_1 = O_P(r_{\bar{\theta}})$, we have that*

$$\begin{aligned} &\left\|\left\| \frac{1}{n+k-1} \left(\sum_{i=1}^{n} \left(\nabla \mathcal{L}(\bar{\theta}; Z_{i1}) - \nabla \mathcal{L}_N(\bar{\theta})\right) \left(\nabla \mathcal{L}(\bar{\theta}; Z_{i1}) - \nabla \mathcal{L}_N(\bar{\theta})\right)^\top \right.\right.\right. \\ &\quad \left.\left.\left. + \sum_{j=2}^{k} n \left(\nabla \mathcal{L}_j(\bar{\theta}) - \nabla \mathcal{L}_N(\bar{\theta})\right) \left(\nabla \mathcal{L}_j(\bar{\theta}) - \nabla \mathcal{L}_N(\bar{\theta})\right)^\top \right) - \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top\right] \right\|\right\|_{\max} \\ &= O_P\Bigg(\sqrt{\frac{\log d}{n+k}} + \frac{\log^2(d(n+k)) \log d}{n+k} + \left(\left(1 + \sqrt{\frac{\log d}{N}}\right) \frac{nk}{n+k} + \log((n+k)d) \right) r_{\bar{\theta}}^2 \\ &\quad + \left(\sqrt{\log((n+k)d)} + \frac{\log^{1/4} d \sqrt{\log((n+k)d)}}{(n+k)^{1/4}} + \sqrt{\frac{\log^3(d(n+k)) \log d}{n+k}} \right) r_{\bar{\theta}} \Bigg). \end{aligned}$$

Proof of Lemma 32. By Lemma 31, it suffices to bound $V_1(\bar{\theta})$, $V_1'(\bar{\theta})$, V_2, V_2', and $V_3(\bar{\theta})$. By the proof of Lemma 30, we have that under Assumptions (A1) and (A2), assuming that $\left\|\bar{\theta} - \theta^*\right\|_1 = O_P(r_{\bar{\theta}})$,

$$\begin{aligned} V_1(\bar{\theta}) &= \frac{k-1}{n+k-1} O_P\left(\left(1 + \left(\frac{\log d}{k}\right)^{1/4} + \sqrt{\frac{\log^2(dk) \log d}{k}} \right) \sqrt{\log(kd)} r_{\bar{\theta}} + \log(kd) r_{\bar{\theta}}^2 \right) \\ &= O_P\left(\left(1 + \left(\frac{\log d}{k}\right)^{1/4} + \sqrt{\frac{\log^2(dk) \log d}{k}} \right) \frac{k\sqrt{\log(kd)}}{n+k} r_{\bar{\theta}} + \frac{k \log(kd)}{n+k} r_{\bar{\theta}}^2 \right), \end{aligned}$$

$$V_2 = \frac{k-1}{n+k-1} O_P\left(\sqrt{\frac{\log d}{k}} + \frac{\log^2(dk) \log d}{k} \right) = O_P\left(\frac{\sqrt{k \log d}}{n+k} + \frac{\log^2(dk) \log d}{n+k} \right),$$

and

$$V_3(\bar{\theta}) = \frac{nk}{n+k-1} O_P\left(\left(1 + \sqrt{\frac{\log d}{N}}\right) r_{\bar{\theta}}^2 + \frac{\log d}{N} \right) = O_P\left(\left(1 + \sqrt{\frac{\log d}{N}}\right) \frac{nk}{n+k} r_{\bar{\theta}}^2 + \frac{\log d}{n+k} \right).$$

It remains to bound $V_1'(\bar{\theta})$ and V_2'.

To bound V_2', we have that in linear model, under Assumptions (A1) and (A2),

$$V_2' = \frac{n}{n+k-1} \left\| \frac{1}{n} \sum_{i=1}^{n} (x_{i1} e_{i1}) (x_{i1} e_{i1})^\top - \sigma^2 \Sigma \right\|_{\max}.$$

Note that

$$\mathbb{E}\left[(x_{i1} e_{i1})_l^2 \right] = \sigma^2 \Sigma_{l,l}$$

is bounded away from zero, and also, $(x_{i1} e_{i1})_l$ is sub-exponential with O(1) ψ_1-norm for each (i,l). Then, by the proof of Corollary 3.1 of Chernozhukov et al. (2013), we have that

$$\mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} (x_{i1} e_{i1}) (x_{i1} e_{i1})^\top - \sigma^2 \Sigma \right\|_{\max} \right] \lesssim \sqrt{\frac{\log d}{n}} + \frac{\log^2(dn) \log d}{n},$$

and then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (x_{i1} e_{i1}) (x_{i1} e_{i1})^\top - \sigma^2 \Sigma \right\|_{\max} \lesssim \frac{1}{\delta} \left(\sqrt{\frac{\log d}{n}} + \frac{\log^2(dn) \log d}{n} \right),$$

by Markov's inequality, which implies that

$$V_2' = \frac{n}{n+k-1} O_P\left(\sqrt{\frac{\log d}{n}} + \frac{\log^2(dn) \log d}{n} \right) = O_P\left(\frac{\sqrt{n \log d}}{n+k} + \frac{\log^2(dn) \log d}{n+k} \right).$$

Lastly, we bound $V_1'(\bar{\theta})$ using the same argument as in bounding $U_1(\bar{\theta})$ in the proof of Lemma 30. We write $\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta)$ as $(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) - \nabla \mathcal{L}(\theta^*; Z_{i1})) + \nabla \mathcal{L}(\theta^*; Z_{i1})$, and obtain by the triangle inequality that

$$\begin{aligned}
\frac{n+k-1}{n} V_1'(\bar{\theta}) \leq & \left\| \frac{1}{n} \sum_{i=1}^{n} (\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) - \nabla \mathcal{L}(\theta^*; Z_{i1})) (\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) - \nabla \mathcal{L}(\theta^*; Z_{i1}))^\top \right\|_{\max} \\
& + \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \mathcal{L}(\theta^*; Z_{i1}) (\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) - \nabla \mathcal{L}(\theta^*; Z_{i1}))^\top \right\|_{\max} \\
& + \left\| \frac{1}{n} \sum_{i=1}^{n} (\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) - \nabla \mathcal{L}(\theta^*; Z_{i1})) \nabla \mathcal{L}(\theta^*; Z_{i1})^\top \right\|_{\max} \\
= & \left\| \frac{1}{n} \sum_{i=1}^{n} (\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) - \nabla \mathcal{L}(\theta^*; Z_{i1})) (\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) - \nabla \mathcal{L}(\theta^*; Z_{i1}))^\top \right\|_{\max} \\
& + 2 \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \mathcal{L}(\theta^*; Z_{i1}) (\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) - \nabla \mathcal{L}(\theta^*; Z_{i1}))^\top \right\|_{\max} \\
:= & V_{11}'(\bar{\theta}) + 2 V_{12}'(\bar{\theta}).
\end{aligned}$$

Applying Cauchy-Schwarz inequality, we obtain that

$$\begin{aligned}
V'_{12}(\bar\theta) \le{}& \left\|\!\left\| \frac{1}{n}\sum_{i=1}^n \nabla\mathcal{L}(\theta^*;Z_{i1})\nabla\mathcal{L}(\theta^*;Z_{i1})^\top \right\|\!\right\|_{\max}^{1/2} \\
&\cdot \left\|\!\left\| \frac{1}{n}\sum_{i=1}^n \left(\nabla\mathcal{L}(\theta;Z_{i1}) - \nabla\mathcal{L}^*(\theta) - \nabla\mathcal{L}(\theta^*;Z_{i1})\right)\left(\nabla\mathcal{L}(\theta;Z_{i1}) - \nabla\mathcal{L}^*(\theta) - \nabla\mathcal{L}(\theta^*;Z_{i1})\right)^\top \right\|\!\right\|_{\max}^{1/2} \\
={}& \left\|\!\left\| \frac{1}{n}\sum_{i=1}^n \nabla\mathcal{L}(\theta^*;Z_{i1})\nabla\mathcal{L}(\theta^*;Z_{i1})^\top \right\|\!\right\|_{\max}^{1/2} V'_{11}(\bar\theta)^{1/2}.
\end{aligned}$$

By the triangle inequality, we have that

$$\begin{aligned}
&\left\|\!\left\| \frac{1}{n}\sum_{i=1}^n \nabla\mathcal{L}(\theta^*;Z_{i1})\nabla\mathcal{L}(\theta^*;Z_{i1})^\top \right\|\!\right\|_{\max} \\
&\le \left\|\!\left\| \frac{1}{n}\sum_{i=1}^n \nabla\mathcal{L}(\theta^*;Z_{i1})\nabla\mathcal{L}(\theta^*;Z_{i1})^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right] \right\|\!\right\|_{\max} + \left\|\!\left| \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right] \right\|\!\right\|_{\max} \\
&= \frac{n+k-1}{n} V'_2 + \sigma^2 \left\|\!\left|\Sigma\right\|\!\right\|_{\max} \\
&= O_P\left(1 + \sqrt{\frac{\log d}{n}} + \frac{\log^2(dn)\log d}{n}\right).
\end{aligned}$$

It remains to bound $V'_{11}(\bar\theta)$. Note that

$$\begin{aligned}
\nabla\mathcal{L}(\theta;Z_{i1}) - \nabla\mathcal{L}^*(\theta) - \nabla\mathcal{L}(\theta^*;Z_{i1}) &= x_{ij}(x_{ij}^\top\bar\theta - y_{ij}) - \Sigma(\bar\theta - \theta^*) + x_{ij}(x_{ij}^\top\theta^* - y_{ij}) \\
&= \left(x_{ij}x_{ij}^\top - \Sigma\right)(\bar\theta - \theta^*).
\end{aligned}$$

Then, we have by the triangle inequality that

$$\begin{aligned}
V'_{11}(\bar\theta) &= \left\|\!\left\| \frac{1}{n}\sum_{i=1}^n \left(x_{i1}x_{i1}^\top - \Sigma\right)(\bar\theta - \theta^*)(\bar\theta - \theta^*)^\top\left(x_{i1}x_{i1}^\top - \Sigma\right) \right\|\!\right\|_{\max} \\
&\le \frac{1}{n}\sum_{i=1}^n \left\|\!\left\| \left(x_{i1}x_{i1}^\top - \Sigma\right)(\bar\theta - \theta^*)(\bar\theta - \theta^*)^\top\left(x_{i1}x_{i1}^\top - \Sigma\right) \right\|\!\right\|_{\max} \\
&= \frac{1}{n}\sum_{i=1}^n \left\| \left(x_{i1}x_{i1}^\top - \Sigma\right)(\bar\theta - \theta^*) \right\|_\infty^2 \\
&\le \frac{1}{n}\sum_{i=1}^n \left\|\!\left\| x_{i1}x_{i1}^\top - \Sigma \right\|\!\right\|_{\max}^2 \left\|\bar\theta - \theta^*\right\|_1^2.
\end{aligned}$$

Similarly to obtaining (41), we have that

$$P\left(\left\|\!\left\| x_{i1}x_{i1}^\top - \Sigma \right\|\!\right\|_{\max} > \left\|\!\left|\Sigma\right\|\!\right\|_{\max}\left(\frac{\log\frac{2nd^2}{\delta}}{c} \vee \sqrt{\frac{\log\frac{2nd^2}{\delta}}{c}}\right)\right) \le \frac{\delta}{n},$$

and then, by the union bound,

$$P\left(\max_i \left\|\!\left| x_{i1}x_{i1}^\top - \Sigma \right|\!\right\|_{\max} > \left\|\!\left|\Sigma\right|\!\right\|_{\max}\left(\frac{\log\frac{2nd^2}{\delta}}{c} \vee \sqrt{\frac{\log\frac{2nd^2}{\delta}}{c}}\right)\right) \le \delta,$$

which implies that

$$\max_i \left\|\!\left| x_{i1}x_{i1}^\top - \Sigma \right|\!\right\|_{\max} = O_P\left(\sqrt{\log(nd)}\right).$$

Putting all the preceding bounds together, we obtain that

$$V_{11}'(\bar\theta) = O_P\left(\log(nd) r_{\bar\theta}^2\right),$$

$$V_{12}'(\bar\theta) = O_P\left(\left(1 + \left(\frac{\log d}{n}\right)^{1/4} + \sqrt{\frac{\log^2(dn)\log d}{n}}\right)\sqrt{\log(nd)}r_{\bar\theta}\right),$$

$$\begin{aligned}V_1'(\bar\theta) &= \frac{n}{n+k-1}O_P\left(\left(1 + \left(\frac{\log d}{n}\right)^{1/4} + \sqrt{\frac{\log^2(dn)\log d}{n}}\right)\sqrt{\log(nd)}r_{\bar\theta} + \log(nd)r_{\bar\theta}^2\right)\\ &= O_P\left(\left(1 + \left(\frac{\log d}{n}\right)^{1/4} + \sqrt{\frac{\log^2(dn)\log d}{n}}\right)\frac{n\sqrt{\log(nd)}}{n+k}r_{\bar\theta} + \frac{n\log(nd)}{n+k}r_{\bar\theta}^2\right),\end{aligned}$$

and finally,

$$\begin{aligned}&\left\|\!\left|\frac{1}{n+k-1}\left(\sum_{i=1}^n \left(\nabla\mathcal{L}(\bar\theta;Z_{i1}) - \nabla\mathcal{L}_N(\bar\theta)\right)\left(\nabla\mathcal{L}(\bar\theta;Z_{i1}) - \nabla\mathcal{L}_N(\bar\theta)\right)^\top\right.\right.\right.\\ &\quad\left.\left.\left. + \sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\bar\theta) - \nabla\mathcal{L}_N(\bar\theta)\right)\left(\nabla\mathcal{L}_j(\bar\theta) - \nabla\mathcal{L}_N(\bar\theta)\right)^\top\right) - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right|\!\right\|_{\max}\\ &= O_P\left(\sqrt{\frac{\log d}{n+k}} + \frac{\log^2(d(n+k))\log d}{n+k} + \left(\left(1 + \sqrt{\frac{\log d}{N}}\right)\frac{nk}{n+k} + \log((n+k)d)\right)r_{\bar\theta}^2\right.\\ &\quad\left. + \left(\sqrt{\log((n+k)d)} + \frac{\log^{1/4}d\sqrt{\log((n+k)d)}}{(n+k)^{1/4}} + \sqrt{\frac{\log^3(d(n+k))\log d}{n+k}}\right)r_{\bar\theta}\right).\end{aligned}$$

$\blacksquare$

Lemma 33 *In sparse GLM, under Assumptions (B1)–(B3), provided that* $\left\|\bar\theta - \theta^*\right\|_1 = O_P(r_{\bar\theta})$*, we have that*

$$\begin{aligned}&\left\|\!\left|\frac{1}{k}\sum_{j=1}^k n\left(\nabla\mathcal{L}_j(\bar\theta) - \nabla\mathcal{L}_N(\bar\theta)\right)\left(\nabla\mathcal{L}_j(\bar\theta) - \nabla\mathcal{L}_N(\bar\theta)\right)^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right|\!\right\|_{\max}\\ &= O_P\left(\sqrt{\frac{\log d}{k}} + \frac{\log d}{k} + \left(1 + \left(\frac{\log d}{k}\right)^{1/4}\right)\left(\sqrt{\log d} + \sqrt{n}r_{\bar\theta}\right)r_{\bar\theta} + \left(n + \log d + nr_{\bar\theta}^2\right)r_{\bar\theta}^2\right).\end{aligned}$$

Proof of Lemma 33. By Lemma 29, it suffices to bound $U_1(\bar\theta)$, U_2, and $U_3(\bar\theta)$. We begin by bounding U_2. Using the argument for obtaining (46), we have that for any $t>0$,

$$P\left(|\nabla\mathcal{L}_j(\theta^*)_l|>t\right)\le 2\exp\left(-\frac{nt^2}{c}\right),$$

and then,

$$P\left(\sqrt{n}\,|\nabla\mathcal{L}_j(\theta^*)_l|>t\right)\le 2\exp\left(-\frac{t^2}{c}\right),$$

that is, $\sqrt{n}\nabla\mathcal{L}_j(\theta^*)_l$ is sub-Gaussian with $O(1)$ ψ_2-norm. Therefore, $n\nabla\mathcal{L}_j(\theta^*)_l\nabla\mathcal{L}_j(\theta^*)_{l'}$ is sub-exponential with $O(1)$ ψ_1-norm. Note that $\mathbb{E}[n\nabla\mathcal{L}_j(\theta^*)_l\nabla\mathcal{L}_j(\theta^*)_{l'}]=\mathbb{E}[\nabla\mathcal{L}(\theta^*;Z)_l\nabla\mathcal{L}(\theta^*;Z)_{l'}]$. Then, we apply Bernstein's inequality and obtain that for any $t>0$,

$$P\left(\left|\frac{1}{k}\sum_{j=1}^k n\nabla\mathcal{L}_j(\theta^*)_l\nabla\mathcal{L}_j(\theta^*)_{l'}-\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)_l\nabla\mathcal{L}(\theta^*;Z)_{l'}\right]\right|>t\right)\le 2\exp\left(-ck\left(t^2\wedge t\right)\right),$$

or, for any $\delta\in(0,1)$,

$$P\left(\left|\frac{1}{k}\sum_{j=1}^k n\nabla\mathcal{L}_j(\theta^*)_l\nabla\mathcal{L}_j(\theta^*)_{l'}-\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)_l\nabla\mathcal{L}(\theta^*;Z)_{l'}\right]\right|>\sqrt{\frac{\log\frac{2d^2}{\delta}}{ck}}\vee\frac{\log\frac{2d^2}{\delta}}{ck}\right)\le\frac{\delta}{d^2},$$

and by the union bound, with probability at least $1-\delta$,

$$U_2\le\sqrt{\frac{\log\frac{2d^2}{\delta}}{ck}}\vee\frac{\log\frac{2d^2}{\delta}}{ck},$$

which implies that

$$U_2=O_P\left(\sqrt{\frac{\log d}{k}}\right).$$

Next, we bound $U_3(\bar\theta)$. By the triangle inequality, we have that

$$\left\|\nabla\mathcal{L}_N(\bar\theta)-\nabla\mathcal{L}^*(\bar\theta)\right\|_\infty\le\left\|\nabla\mathcal{L}_N(\bar\theta)-\nabla\mathcal{L}_N(\theta^*)\right\|_\infty+\left\|\nabla\mathcal{L}_N(\theta^*)\right\|_\infty+\left\|\nabla\mathcal{L}^*(\bar\theta)\right\|_\infty.$$

By (43), we have that

$$\begin{aligned}\nabla\mathcal{L}_N(\bar\theta)-\nabla\mathcal{L}_N(\theta^*)&=\int_0^1\nabla^2\mathcal{L}_N(\theta^*+t(\bar\theta-\theta^*))dt(\bar\theta-\theta^*)\\&=\int_0^1\frac{1}{N}\sum_{i=1}^n\sum_{j=1}^k g''(y_{ij},x_{ij}^\top(\theta^*+t(\bar\theta-\theta^*)))x_{ij}x_{ij}^\top dt(\bar\theta-\theta^*),\end{aligned}$$

and then, under Assumptions (B1) and (B2),

$$\begin{aligned}\left\|\nabla\mathcal{L}_N(\bar\theta)-\nabla\mathcal{L}_N(\theta^*)\right\|_\infty&=\int_0^1\frac{1}{N}\sum_{i=1}^n\sum_{j=1}^k\left|g''(y_{ij},x_{ij}^\top(\theta^*+t(\bar\theta-\theta^*)))\right|\|x_{ij}\|_\infty^2\,dt\left\|\bar\theta-\theta^*\right\|_\infty\\&\lesssim\left\|\bar\theta-\theta^*\right\|_\infty.\end{aligned}$$

Note that for any θ,

$$\begin{aligned}
\|\nabla\mathcal{L}^*(\theta)\|_\infty &= \|\nabla\mathcal{L}^*(\theta) - \nabla\mathcal{L}^*(\theta^*)\|_\infty \\
&= \left\|\mathbb{E}\left[\left(g'(y, x^\top\theta) - g'(y, x^\top\theta^*))\right)x\right]\right\|_\infty \\
&= \left\|\mathbb{E}\left[\int_0^1 g''(y, x^\top(\theta^* + t(\theta - \theta^*)))dtxx^\top(\theta - \theta^*)\right]\right\|_\infty \\
&\le \mathbb{E}\left[\int_0^1 \left|g''(y, x^\top(\theta^* + t(\theta - \theta^*)))\right| dt \, \|x\|_\infty^2 \, \|\theta - \theta^*\|_\infty\right] \\
&\lesssim \|\theta - \theta^*\|_\infty .
\end{aligned}$$

Therefore,

$$\left\|\nabla\mathcal{L}^*(\bar\theta)\right\|_\infty \lesssim \left\|\bar\theta - \theta^*\right\|_\infty .$$

By (47), we have that

$$\|\nabla\mathcal{L}_N(\theta^*)\|_\infty = O_P\left(\sqrt{\frac{\log d}{N}}\right).$$

Then, assuming that $\left\|\bar\theta - \theta^*\right\|_1 = O_P(r_{\bar\theta})$, we have that

$$\left\|\nabla\mathcal{L}_N(\bar\theta) - \nabla\mathcal{L}^*(\bar\theta)\right\|_\infty = O_P\left(r_{\bar\theta} + \sqrt{\frac{\log d}{N}}\right),$$

and then,

$$U_3(\bar\theta) = O_P\left(nr_{\bar\theta}^2 + \frac{\log d}{k}\right).$$

Lastly, we bound $U_1(\bar\theta)$. As in the proof of Lemma 30, we have that

$$\begin{aligned}
U_1(\bar\theta) \le& \left\|\left\|\frac{1}{k}\sum_{j=1}^k n\left(\nabla\mathcal{L}_j(\bar\theta) - \nabla\mathcal{L}^*(\bar\theta) - \nabla\mathcal{L}_j(\theta^*)\right)\left(\nabla\mathcal{L}_j(\bar\theta) - \nabla\mathcal{L}^*(\bar\theta) - \nabla\mathcal{L}_j(\theta^*)\right)^\top\right\|\right\|_{\max} \\
&+ 2\left\|\left\|\frac{1}{k}\sum_{j=1}^k n\nabla\mathcal{L}_j(\theta^*)\left(\nabla\mathcal{L}_j(\bar\theta) - \nabla\mathcal{L}^*(\bar\theta) - \nabla\mathcal{L}_j(\theta^*)\right)^\top\right\|\right\|_{\max} \\
:=& U_{11}(\bar\theta) + 2U_{12}(\bar\theta),
\end{aligned}$$

and

$$U_{12}(\bar\theta) \le \left\|\left\|\frac{1}{k}\sum_{j=1}^k n\nabla\mathcal{L}_j(\theta^*)\nabla\mathcal{L}_j(\theta^*)^\top\right\|\right\|_{\max}^{1/2} U_{11}(\bar\theta)^{1/2}.$$

Note that $\left\|\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|_{\max} = O(1)$ under Assumption (B3). Then, by the triangle inequality, we have that

$$\begin{aligned}
&\left\|\frac{1}{k}\sum_{j=1}^k n\nabla\mathcal{L}_j(\theta^*)\nabla\mathcal{L}_j(\theta^*)^\top\right\|_{\max} \\
&\leq \left\|\frac{1}{k}\sum_{j=1}^k n\nabla\mathcal{L}_j(\theta^*)\nabla\mathcal{L}_j(\theta^*)^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|_{\max} + \left\|\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|_{\max} \\
&= U_2 + \left\|\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|_{\max} \\
&= O_P\left(1+\sqrt{\frac{\log d}{k}}\right).
\end{aligned}$$

It remains to bound $U_{11}(\bar\theta)$. Note that

$$\begin{aligned}
\nabla\mathcal{L}_j(\bar\theta) - \nabla\mathcal{L}_j(\theta^*) &= \int_0^1 \nabla^2\mathcal{L}_j(\theta^* + t(\bar\theta-\theta^*))dt(\bar\theta-\theta^*) \\
&= \int_0^1 \frac{1}{n}\sum_{i=1}^n g''(y_{ij}, x_{ij}^\top(\theta^* + t(\bar\theta-\theta^*)))x_{ij}x_{ij}^\top dt(\bar\theta-\theta^*),
\end{aligned}$$

and

$$g''(y_{ij}, x_{ij}^\top(\theta^* + t(\bar\theta-\theta^*))) = g''(y_{ij}, x_{ij}^\top\theta^*) + \int_0^1 g'''(y_{ij}, x_{ij}^\top(\theta^* + st(\bar\theta-\theta^*)))ds x_{ij}^\top(t(\bar\theta-\theta^*)),$$

and then

$$\begin{aligned}
\nabla\mathcal{L}_j(\bar\theta) - \nabla\mathcal{L}_j(\theta^*) &= \frac{1}{n}\sum_{i=1}^n g''(y_{ij}, x_{ij}^\top\theta^*)x_{ij}x_{ij}^\top(\bar\theta-\theta^*) \\
&\quad + \int_0^1\int_0^1 \frac{1}{n}\sum_{i=1}^n g'''(y_{ij}, x_{ij}^\top(\theta^* + st(\bar\theta-\theta^*)))x_{ij}^\top t(\bar\theta-\theta^*)x_{ij}x_{ij}^\top dtds(\bar\theta-\theta^*).
\end{aligned}$$

In a similar way, we have that

$$\begin{aligned}
\nabla\mathcal{L}^*(\bar\theta) &= \nabla\mathcal{L}^*(\bar\theta) - \nabla\mathcal{L}^*(\theta^*) \\
&= \mathbb{E}\left[g''(y, x^\top\theta^*)xx^\top\right](\bar\theta-\theta^*) \\
&\quad + \int_0^1\int_0^1 \mathbb{E}_{x,y}\left[g'''(y, x^\top(\theta^* + st(\bar\theta-\theta^*)))x^\top t(\bar\theta-\theta^*)xx^\top\right] dtds(\bar\theta-\theta^*),
\end{aligned}$$

and then,

$$\begin{aligned}\nabla\mathcal{L}_j(\bar\theta)-\nabla\mathcal{L}^*(\bar\theta)-\nabla\mathcal{L}_j(\theta^*) &= \left(\frac{1}{n}\sum_{i=1}^n g''(y_{ij},x_{ij}^\top\theta^*)x_{ij}x_{ij}^\top-\mathbb{E}\left[g''(y,x^\top\theta^*)xx^\top\right]\right)(\bar\theta-\theta^*)\\ &\quad+\bigg(\int_0^1\int_0^1\frac{1}{n}\sum_{i=1}^n g'''(y_{ij},x_{ij}^\top(\theta^*+st(\bar\theta-\theta^*)))x_{ij}^\top t(\bar\theta-\theta^*)x_{ij}x_{ij}^\top\\ &\quad-\mathbb{E}_{x,y}\left[g'''(y,x^\top(\theta^*+st(\bar\theta-\theta^*)))x^\top t(\bar\theta-\theta^*)xx^\top\right]dtds(\bar\theta-\theta^*)\bigg)\\ &:=U_{111,j}(\bar\theta)+U_{112,j}(\bar\theta).\end{aligned}$$

Then, we have by the triangle inequality that

$$\begin{aligned}U_{11}(\bar\theta) &= \left\|\frac{1}{k}\sum_{j=1}^k n\left(U_{111,j}(\bar\theta)+U_{112,j}(\bar\theta)\right)\left(U_{111,j}(\bar\theta)+U_{112,j}(\bar\theta)\right)^\top\right\|_{\max}\\ &\le\frac{1}{k}\sum_{j=1}^k n\left\|\left(U_{111,j}(\bar\theta)+U_{112,j}(\bar\theta)\right)\left(U_{111,j}(\bar\theta)+U_{112,j}(\bar\theta)\right)^\top\right\|_{\max}\\ &=\frac{1}{k}\sum_{j=1}^k n\left\|U_{111,j}(\bar\theta)+U_{112,j}(\bar\theta)\right\|_\infty^2\\ &\le\frac{2}{k}\sum_{j=1}^k n\left(\left\|U_{111,j}(\bar\theta)\right\|_\infty^2+\left\|U_{112,j}(\bar\theta)\right\|_\infty^2\right)\end{aligned}$$

Using the argument for obtaining (45), we have that

$$\begin{aligned}\left\|U_{111,j}(\bar\theta)\right\|_\infty &= \left\|\left(\nabla^2\mathcal{L}_j(\theta^*)-\nabla^2\mathcal{L}^*(\theta^*)\right)(\bar\theta-\theta^*)\right\|_\infty\\ &\le\left\|\nabla^2\mathcal{L}_j(\theta^*)-\nabla^2\mathcal{L}^*(\theta^*)\right\|_{\max}\left\|\bar\theta-\theta^*\right\|_1\\ &=O_P\left(\sqrt{\frac{\log d}{n}}\right)O_P\left(r_{\bar\theta}\right)\\ &=O_P\left(\sqrt{\frac{\log d}{n}}r_{\bar\theta}\right).\end{aligned}$$

Under Assumptions (B1) and (B2), we have that

$$\begin{aligned}\left\|U_{112,j}(\bar\theta)\right\|_\infty &\le\int_0^1\int_0^1\frac{1}{n}\sum_{i=1}^n\left|g'''(y_{ij},x_{ij}^\top(\theta^*+st(\bar\theta-\theta^*)))\right|\left\|x_{ij}\right\|_\infty t\left\|\bar\theta-\theta^*\right\|_1\left\|x_{ij}\right\|_\infty^2\\ &\quad+\mathbb{E}_{x,y}\left[\left|g'''(y,x^\top(\theta^*+st(\bar\theta-\theta^*)))\right|\left\|x\right\|_\infty t\left\|\bar\theta-\theta^*\right\|_1\left\|x\right\|_\infty^2\right]dtds\left\|\bar\theta-\theta^*\right\|_1\\ &\lesssim\left\|\bar\theta-\theta^*\right\|_1^2\\ &=O_P\left(r_{\bar\theta}^2\right).\end{aligned}$$

Hence, we have that

$$U_{11}(\bar{\theta}) = n\left(O_P\left(\frac{\log d}{n}r_{\bar{\theta}}^2\right) + O_P\left(r_{\bar{\theta}}^4\right)\right) = O_P\left(\left(\log d + nr_{\bar{\theta}}^2\right)r_{\bar{\theta}}^2\right).$$

Putting all the preceding bounds together, we obtain that

$$U_{12}(\bar{\theta}) = O_P\left(\left(1+\left(\frac{\log d}{k}\right)^{1/4}\right)\left(\sqrt{\log d}+\sqrt{n}r_{\bar{\theta}}\right)r_{\bar{\theta}}\right),$$

$$U_1(\bar{\theta}) = O_P\left(\left(1+\left(\frac{\log d}{k}\right)^{1/4}\right)\left(\sqrt{\log d}+\sqrt{n}r_{\bar{\theta}}\right)r_{\bar{\theta}} + \left(\log d + nr_{\bar{\theta}}^2\right)r_{\bar{\theta}}^2\right),$$

and finally,

$$\begin{aligned}&\left|\left|\left|\frac{1}{k}\sum_{j=1}^k n\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right|\right|\right|_2 \\ &= O_P\left(\sqrt{\frac{\log d}{k}}+\frac{\log d}{k}+\left(1+\left(\frac{\log d}{k}\right)^{1/4}\right)\left(\sqrt{\log d}+\sqrt{n}r_{\bar{\theta}}\right)r_{\bar{\theta}} + \left(n+\log d + nr_{\bar{\theta}}^2\right)r_{\bar{\theta}}^2\right).\end{aligned}$$

■

Lemma 34 *In sparse GLM, under Assumptions (B1)–(B3), provided that* $\left\|\bar{\theta}-\theta^*\right\|_1 = O_P(r_{\bar{\theta}})$*, we have that*

$$\begin{aligned}&\left|\left|\left|\frac{1}{n+k-1}\left(\sum_{i=1}^n\left(\nabla\mathcal{L}(\bar{\theta};Z_{i1})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}(\bar{\theta};Z_{i1})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top\right.\right.\right.\right. \\ &\qquad\left.\left.\left.\left.+\sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta})-\nabla\mathcal{L}_N(\bar{\theta})\right)^\top\right) - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right|\right|\right|_{\max} \\ &= O_P\Bigg(\sqrt{\frac{\log d}{n+k}}+\frac{\log d}{n+k}+\frac{nk}{n+k}r_{\bar{\theta}}^2+\left(1+\left(\frac{\log d}{n}\right)^{1/4}\right)\frac{n}{n+k}\left(r_{\bar{\theta}}+r_{\bar{\theta}}^2\right)+\frac{n}{n+k}r_{\bar{\theta}}^4 \\ &\qquad+\left(1+\left(\frac{\log d}{k}\right)^{1/4}\right)\frac{k\sqrt{\log d}+k\sqrt{n}r_{\bar{\theta}}}{n+k}r_{\bar{\theta}}+\frac{k\log d+knr_{\bar{\theta}}^2}{n+k}r_{\bar{\theta}}^2\Bigg).\end{aligned}$$

Proof of Lemma 34. By Lemma 31, it suffices to bound $V_1(\bar{\theta})$, $V_1'(\bar{\theta})$, V_2, V_2', and $V_3(\bar{\theta})$. By the proof of Lemma 33, we have that under Assumptions (B1)–(B3), assuming that $\left\|\bar{\theta}-\theta^*\right\|_1 = O_P(r_{\bar{\theta}})$,

$$\begin{aligned}V_1(\bar{\theta}) &= \frac{k-1}{n+k-1}O_P\left(\left(1+\left(\frac{\log d}{k}\right)^{1/4}\right)\left(\sqrt{\log d}+\sqrt{n}r_{\bar{\theta}}\right)r_{\bar{\theta}}+\left(\log d+nr_{\bar{\theta}}^2\right)r_{\bar{\theta}}^2\right) \\ &= O_P\left(\left(1+\left(\frac{\log d}{k}\right)^{1/4}\right)\frac{k\sqrt{\log d}+k\sqrt{n}r_{\bar{\theta}}}{n+k}r_{\bar{\theta}}+\frac{k\log d+knr_{\bar{\theta}}^2}{n+k}r_{\bar{\theta}}^2\right),\end{aligned}$$

$$V_2 = \frac{k-1}{n+k-1} O_P\left(\sqrt{\frac{\log d}{k}}\right) = O_P\left(\frac{\sqrt{k \log d}}{n+k}\right),$$

and

$$V_3(\bar{\theta}) = \frac{nk}{n+k-1} O_P\left(r_{\bar{\theta}}^2 + \frac{\log d}{N}\right) = O_P\left(\frac{nk}{n+k} r_{\bar{\theta}}^2 + \frac{\log d}{n+k}\right).$$

It remains to bound $V_1'(\bar{\theta})$ and V_2'.

To bound V_2', we note that each $\nabla \mathcal{L}(\theta^*; Z_{i1})_l \nabla \mathcal{L}(\theta^*; Z_{i1})_{l'} = g'(y_{i1}, x_{i1}^\top \theta^*)^2 x_{i1,l} x_{i1,l'}$ is bounded under Assumptions (B1) and (B2). Applying Hoeffding's inequality, we obtain that for any $t > 0$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}(\theta^*; Z_{i1})_l \nabla \mathcal{L}(\theta^*; Z_{i1})_{l'} - \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z)_l \nabla \mathcal{L}(\theta^*; Z)_{l'}\right]\right| > t\right) \le 2\exp\left(-\frac{nt^2}{c}\right),$$

for some constant c, or, for any $\delta \in (0,1)$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}(\theta^*; Z_{i1})_l \nabla \mathcal{L}(\theta^*; Z_{i1})_{l'} - \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z)_l \nabla \mathcal{L}(\theta^*; Z)_{l'}\right]\right| > \sqrt{\frac{c \log \frac{2d^2}{\delta}}{n}}\right) \le \frac{\delta}{d^2},$$

and by the union bound, with probability at least $1 - \delta$,

$$\left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}(\theta^*; Z_{i1}) \nabla \mathcal{L}(\theta^*; Z_{i1})^\top - \mathbb{E}\left[\nabla \mathcal{L}(\theta^*; Z) \nabla \mathcal{L}(\theta^*; Z)^\top\right]\right\|_{\max} \le \sqrt{\frac{c \log \frac{2d^2}{\delta}}{n}},$$

which implies that

$$V_2' = \frac{n}{n+k-1} O_P\left(\sqrt{\frac{\log d}{n}}\right) = O_P\left(\sqrt{\frac{n \log d}{n+k}}\right).$$

Lastly, we bound $V_1'(\bar{\theta})$. As in the proof of Lemma 32, we have that

$$\begin{aligned}
\frac{n+k-1}{n} V_1'(\bar{\theta}) \le & \left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) - \nabla \mathcal{L}(\theta^*; Z_{i1})\right)\left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) - \nabla \mathcal{L}(\theta^*; Z_{i1})\right)^\top\right\|_{\max} \\
& + 2\left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}(\theta^*; Z_{i1})\left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) - \nabla \mathcal{L}(\theta^*; Z_{i1})\right)^\top\right\|_{\max} \\
:= & V_{11}'(\bar{\theta}) + 2V_{12}'(\bar{\theta}),
\end{aligned}$$

and

$$V_{12}'(\bar{\theta}) \le \left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}(\theta^*; Z_{i1}) \nabla \mathcal{L}(\theta^*; Z_{i1})^\top\right\|_{\max}^{1/2} V_{11}'(\bar{\theta})^{1/2}.$$

Note that $\left\|\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|_{\max} = O(1)$ under Assumption (B3). Then, by the triangle inequality, we have that

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n \nabla\mathcal{L}(\theta^*;Z_{i1})\nabla\mathcal{L}(\theta^*;Z_{i1})^\top\right\|_{\max} \\
&\le \left\|\frac{1}{n}\sum_{i=1}^n \nabla\mathcal{L}(\theta^*;Z_{i1})\nabla\mathcal{L}(\theta^*;Z_{i1})^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|_{\max} + \left\|\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|_{\max} \\
&= \frac{n+k-1}{n}V_2' + \left\|\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|_{\max} \\
&= O_P\left(1+\sqrt{\frac{\log d}{n}}\right).
\end{aligned}$$

It remains to bound $V_{11}'(\bar{\theta})$. Using the same argument for analyzing $\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}^*(\bar{\theta}) - \nabla\mathcal{L}_j(\theta^*)$ in the proof of Lemma 33, we obtain that

$$\begin{aligned}
\nabla\mathcal{L}(\theta;Z_{i1}) - \nabla\mathcal{L}^*(\theta) - \nabla\mathcal{L}(\theta^*;Z_{i1}) &= \left(g''(y_{i1},x_{i1}^\top\theta^*)x_{i1}x_{i1}^\top - \mathbb{E}\left[g''(y,x^\top\theta^*)xx^\top\right]\right)(\bar{\theta}-\theta^*) \\
&\quad + \bigg(\int_0^1\int_0^1 g'''(y_{i1},x_{i1}^\top(\theta^*+st(\bar{\theta}-\theta^*)))x_{i1}^\top t(\bar{\theta}-\theta^*)x_{i1}x_{i1}^\top \\
&\quad - \mathbb{E}_{x,y}\left[g'''(y,x^\top(\theta^*+st(\bar{\theta}-\theta^*)))x^\top t(\bar{\theta}-\theta^*)xx^\top\right]dtds(\bar{\theta}-\theta^*)\bigg) \\
&:= V_{111,i}'(\bar{\theta}) + V_{112,i}'(\bar{\theta}),
\end{aligned}$$

and

$$\begin{aligned}
V_{11}'(\bar{\theta}) &= \left\|\frac{1}{n}\sum_{i=1}^n \left(V_{111,i}'(\bar{\theta})+V_{112,i}'(\bar{\theta})\right)\left(V_{111,i}'(\bar{\theta})+V_{112,i}'(\bar{\theta})\right)^\top\right\|_{\max} \\
&\le \frac{1}{n}\sum_{i=1}^n \left\|\left(V_{111,i}'(\bar{\theta})+V_{112,i}'(\bar{\theta})\right)\left(V_{111,i}'(\bar{\theta})+V_{112,i}'(\bar{\theta})\right)^\top\right\|_{\max} \\
&= \frac{1}{n}\sum_{i=1}^n \left\|V_{111,i}'(\bar{\theta})+V_{112,i}'(\bar{\theta})\right\|_\infty^2 \\
&\le \frac{2}{n}\sum_{i=1}^n \left(\left\|V_{111,i}'(\bar{\theta})\right\|_\infty^2 + \left\|V_{112,i}'(\bar{\theta})\right\|_\infty^2\right).
\end{aligned}$$

Moreover, under Assumptions (B1)–(B3), we have that

$$\begin{aligned}
\left\|V_{111,i}'(\bar{\theta})\right\|_\infty &= \left\|\left(\nabla^2\mathcal{L}(\theta^*;Z_{i1}) - \nabla^2\mathcal{L}^*(\theta^*)\right)(\bar{\theta}-\theta^*)\right\|_\infty \\
&\le \left\|\nabla^2\mathcal{L}(\theta^*;Z_{i1}) - \nabla^2\mathcal{L}^*(\theta^*)\right\|_{\max}\left\|\bar{\theta}-\theta^*\right\|_1 \\
&\le \left(\left|g''(y_{i1},x_{i1}^\top\theta^*)\right|\left\|x_{i1}\right\|_\infty^2 + \left\|\nabla^2\mathcal{L}^*(\theta^*)\right\|_{\max}\right)\left\|\bar{\theta}-\theta^*\right\|_1 \\
&= O_P\left(r_{\bar{\theta}}\right),
\end{aligned}$$

and

$$\begin{aligned}\left\|V'_{112,i}(\bar\theta)\right\|_\infty \le& \int_0^1\int_0^1 \left|g'''(y_{i1}, x_{i1}^\top(\theta^*+st(\bar\theta-\theta^*)))\right| \|x_{i1}\|_\infty\, t \left\|\bar\theta-\theta^*\right\|_1 \|x_{i1}\|_\infty^2 \\ &+ \mathbb{E}_{x,y}\left[\left|g'''(y, x^\top(\theta^*+st(\bar\theta-\theta^*)))\right| \|x\|_\infty\, t \left\|\bar\theta-\theta^*\right\|_1 \|x\|_\infty^2\right] dtds \left\|\bar\theta-\theta^*\right\|_1 \\ \lesssim& \left\|\bar\theta-\theta^*\right\|_1^2 \\ =& O_P\left(r_{\bar\theta}^2\right),\end{aligned}$$

and hence,

$$V'_{11}(\bar\theta) = O_P\left(r_{\bar\theta}^2 + r_{\bar\theta}^4\right).$$

Putting all the preceding bounds together, we obtain that

$$V'_{12}(\bar\theta) = O_P\left(\left(1+\left(\frac{\log d}{n}\right)^{1/4}\right)\left(r_{\bar\theta}+r_{\bar\theta}^2\right)\right),$$

$$\begin{aligned}V'_1(\bar\theta) &= \frac{n}{n+k-1}O_P\left(\left(1+\left(\frac{\log d}{n}\right)^{1/4}\right)\left(r_{\bar\theta}+r_{\bar\theta}^2\right)+r_{\bar\theta}^2+r_{\bar\theta}^4\right) \\ &= O_P\left(\left(1+\left(\frac{\log d}{n}\right)^{1/4}\right)\frac{n}{n+k}\left(r_{\bar\theta}+r_{\bar\theta}^2\right)+\frac{n}{n+k}r_{\bar\theta}^4\right),\end{aligned}$$

and finally,

$$\begin{aligned}&\left\|\frac{1}{n+k-1}\left(\sum_{i=1}^n\left(\nabla\mathcal{L}(\bar\theta;Z_{i1})-\nabla\mathcal{L}_N(\bar\theta)\right)\left(\nabla\mathcal{L}(\bar\theta;Z_{i1})-\nabla\mathcal{L}_N(\bar\theta)\right)^\top\right.\right. \\ &\quad\left.\left.+\sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\bar\theta)-\nabla\mathcal{L}_N(\bar\theta)\right)\left(\nabla\mathcal{L}_j(\bar\theta)-\nabla\mathcal{L}_N(\bar\theta)\right)^\top\right)-\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|_{\max} \\ &= O_P\left(\sqrt{\frac{\log d}{n+k}}+\frac{\log d}{n+k}+\frac{nk}{n+k}r_{\bar\theta}^2+\left(1+\left(\frac{\log d}{n}\right)^{1/4}\right)\frac{n}{n+k}\left(r_{\bar\theta}+r_{\bar\theta}^2\right)+\frac{n}{n+k}r_{\bar\theta}^4\right. \\ &\quad\left.+\left(1+\left(\frac{\log d}{k}\right)^{1/4}\right)\frac{k\sqrt{\log d}+k\sqrt{n}r_{\bar\theta}}{n+k}r_{\bar\theta}+\frac{k\log d+knr_{\bar\theta}^2}{n+k}r_{\bar\theta}^2\right).\end{aligned}$$

■

Lemma 35 *In high-dimensional linear model, under Assumption (A1), if* $n \gg s^*\log d$, *we have that*

$$\left\|\widetilde\Theta\right\|_\infty = O_P\left(\sqrt{s^*}\right), \quad \left\|\widetilde\Theta-\Theta\right\|_\infty = O_P\left(s^*\sqrt{\frac{\log d}{n}}\right),$$

$$\left\|\widetilde\Theta\frac{X_1^\top X_1}{n}-I_d\right\|_{\max} = O_P\left(\sqrt{\frac{\log d}{n}}\right), \quad \textit{and} \quad \max_l\left\|\widetilde\Theta_l-\Theta_l\right\|_2 = O_P\left(\sqrt{\frac{s^*\log d}{n}}\right).$$

Proof of Lemma 35. In the high-dimensional setting, $\widetilde{\Theta}$ is constructed using nodewise Lasso. We obtain the bounds in the lemma from the proof of Lemma 5.3 and Theorem 2.4 of van de Geer et al. (2014). ■

Lemma 36 *In high-dimensional GLM, under Assumptions (B1)–(B3), if* $n \gg s_0^2 \log^2 d + s^{*2} \log d$*, we have that*

$$\left\|\left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\right\|\right\|_\infty = O_P\left(\sqrt{s^*}\right), \quad \left\|\left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)}) - \Theta\right\|\right\|_\infty = O_P\left((s_0 + s^*)\sqrt{\frac{\log d}{n}}\right),$$

$$\left\|\left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)})\nabla^2 \mathcal{L}_1(\widetilde{\theta}^{(0)}) - I_d\right\|\right\|_{\max} = O_P\left(\sqrt{\frac{\log d}{n}}\right), \quad \textit{and}$$

$$\max_l \left\|\widetilde{\Theta}(\widetilde{\theta}^{(0)})_l - \Theta_l\right\|_2 = O_P\left(\sqrt{\frac{(s_0 + s^*)\log d}{n}}\right).$$

Proof of Lemma 36. In the high-dimensional setting, $\widetilde{\Theta}(\widetilde{\theta}^{(0)})$ is constructed using nodewise Lasso. We obtain the bounds in the lemma from Theorem 3.2 and the proof of Theorems 3.1 and 3.3 of van de Geer et al. (2014). ■