Distributed Algorithms for U-statistics-based Empirical Risk Minimization

Chen, Lanjue; Wan, Alan T.-K.; Zhang, Shuyi; Zhou, Yong

Published in:
Journal of Machine Learning Research

Published: 01/01/2023

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Publication details:
Chen, L., Wan, A. T.-K., Zhang, S., & Zhou, Y. (2023). Distributed Algorithms for U-statistics-based Empirical Risk Minimization. *Journal of Machine Learning Research*, *24*, Article 263.

Journal of Machine Learning Research 24 (2023) 1-43 Submitted 8/21; Revised 2/23; Published 9/23

Distributed Algorithms for U-statistics-based Empirical Risk Minimization

Lanjue Chen CHENLANJUE15@163.COM
Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education, Academy of Statistics and Interdisciplinary Sciences and School of Statistics,
East China Normal University,
Shanghai, China

Alan T.K. Wan ALAN.WAN@CITYU.EDU.HK
Department of Management Sciences, School of Data Science and Department of Biostatistics
City University of Hong Kong,
Kowloon, Hong Kong

Shuyi Zhang SYZHANG@FEM.ECNU.EDU.CN
Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education, Academy of Statistics and Interdisciplinary Sciences and School of Statistics,
East China Normal University,
Shanghai, China

Yong Zhou YZHOU@FEM.ECNU.EDU.CN
Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education, Academy of Statistics and Interdisciplinary Sciences and School of Statistics,
East China Normal University,
Shanghai, China

Editor: Nicolas Vayatis

Abstract

Empirical risk minimization, where the underlying loss function depends on a pair of data points, covers a wide range of application areas in statistics including pairwise ranking and survival analysis. The common empirical risk estimator obtained by averaging values of a loss function over all possible pairs of observations is essentially a U-statistic. One well-known problem with minimizing U-statistic type empirical risks, is that the computational complexity of U-statistics increases quadratically with the sample size. When faced with big data, this poses computational challenges as the colossal number of observation pairs virtually prohibits centralized computing to be performed on a single machine. This paper addresses this problem by developing two computationally and statistically efficient methods based on the divide-and-conquer strategy on a decentralized computing system, whereby the data are distributed among machines to perform the tasks. One of these methods is based on a surrogate of the empirical risk, while the other method extends the one-step updating scheme in classical M-estimation to the case of pairwise loss. We show that the proposed estimators are as asymptotically efficient as the benchmark global U-estimator obtained under centralized computing. As well, we introduce two distributed iterative algorithms to facilitate the implementation of the proposed methods, and conduct extensive numerical experiments to demonstrate their merit.

1. Introduction

The past decade has witnessed the increasing availability of large data sets due to advances in information technology. Big data pose many challenges, including those of computation and storage. In particular, the storage problem makes it difficult for a single computer to handle a large data set. This renders many traditional statistical techniques inapplicable in the face of big data.

Many statistical problems can be expressed in terms of parameter estimation. Let θ^* be the parameter of interest. Typically, θ^* is the minimizer of the population risk $F(\theta) = \mathbb{E}_{Z\sim\mathcal{P}} f(\theta; Z)$, with f being a loss function, and Z a random vector following an unknown probability distribution, $\mathcal{P}$. Assume that z_i's, $i = 1, \ldots, N$, are i.i.d. One of the most common approaches to estimating θ^* is empirical risk minimization (ERM) (Bartlett and Mendelson, 2006). The estimator of θ^* obtained by minimizing the empirical risk function $F_N(\theta) = 1/N \sum_{i=1}^{N} f(\theta; z_i)$ is commonly referred to as the M-estimator, whose properties have been extensively studied under a variety of setups (van de Geer, 2000). Large scale data pose challenges for the implementation of M-estimation due to the computational and storage problems mentioned above. A number of authors, including Zhang et al. (2013), Shamir et al. (2014), Huang and Huo (2015), Smith et al. (2017), Fan et al. (2021) and Jordan et al. (2019), have considered M-estimation when the data are divided into blocks stored on different platforms.

The ERM studies cited above are all based on univariate error or loss functions. There exist many statistical problems where pairwise loss functions provide a more appropriate basis for evaluating estimators' efficiency (see Subsection 3.2 for examples). Under a pairwise decision problem, θ^* is the minimizer of the population risk $L(\theta) = \mathbb{E}_{(Z,Z')\sim\mathcal{P}\times\mathcal{P}} \ell(\theta; Z, Z')$, where Z and Z' are pairwise i.i.d. random vectors and $\ell(\theta; Z, Z')$ is the corresponding loss function. A common estimator of $L(\theta)$ is the empirical risk

$$L_N(\theta) = \frac{1}{N(N-1)} \sum_{i\neq j} \ell\left(\theta, z_i, z_j\right), \tag{1}$$

obtained by averaging $\ell(.)$ over all pairs of observations of z_i's. The estimator $L_N(\theta)$ is essentially a U-statistic, which has the smallest variance among all unbiased estimators (Korolyuk and Borovskich, 2013). The minimizer of U-statistic-based empirical risk is a generalization of the M-estimator, and is commonly referred to as the U-(or M_2-)estimator. The large sample properties of U-estimators under the centralized setting (that is, using all data on one machine) have been extensively studied. See, for example, Bose (1998), Song and Ma (2010) and Bose and Chatterjee (2018b). The U-statistic-based ERM approach has been applied to many statistical problems, including ranking problems (Clémençon et al., 2005, 2008; Agarwal and Niyogi, 2009), survival analysis (Brown and Wang, 2007; Chung et al., 2013), ROC analysis (Ying and Zhou, 2016), and others.

With N observations, there are $O(N^2)$ matched pairs. When the sample size becomes very large, this poses computational challenges that virtually prohibit centralized computing. A more viable approach is to adopt a decentralized, or distributed, computing system, whereby the data are distributed among machines to perform the tasks. When the empirical risk is just a simple average of the errors, distributed computing can usually be implemented without any major difficulty, and the complete information, including the empirical risk val-

ues at different points and their gradients, can be obtained by integrating the subset-based information via a centralized machine. However, it is difficult, if not impossible, to obtain a complete U-statistic due to the high computational cost associated with a colossal number of sample pairs. This is the major issue with distributed inference for pairwise problems, for which the objective function is inseparable across the observations and to which existing distributed computing methods under univariate loss functions cannot be directly applied. We need a new distributed algorithm for the U-statistic-based ERM. The objective of the present paper is to take steps in this direction.

The strategy of divide-and-conquer has long been regarded as an effective paradigm to reduce computational efforts and memory requirements, and has been used in conjunction with many distributed algorithms. Instead of processing all data on a single machine, this strategy divides the data into manageable subsets stored on different local machines, and constructs local statistics from each subset to be integrated at the final stage. In this paper, we develop two simple and reliable distributed methods that use the divide-and-conquer strategy for conducting U-statistics-based ERM. The first method is based on a surrogate empirical risk that can be calculated in a distributed manner, and the second method is a distributed variant of the usual M-estimation procedure. We denote the two methods as the SU-ERM and the OS-ERM methods respectively. These two methods share one common feature that local gradient information based on subsets of data is computed by the local machines. The local statistics are then transferred to the master machine to perform the remaining operations. The SU-ERM method uses a weighted aggregate of local gradients combined with the subset information for computing the surrogate empirical risk that yields the estimator. Let n be the size of the subset sample. The SU-ERM method has the advantage of reducing the computational complexity from $O(N^2)$ to $O(n^2)$, and avoids many memory-intensive operations brought about by the large number of observation pairs. The purpose of the OS-ERM method is to reduce computational cost. It replaces the optimization of the surrogate empirical risk in SU-ERM by a Newton-type algorithm involving the computation of the local Hessian matrix on the central machine. Being a Newton-type optimization algorithm, the OS-ERM method has a closed-form analytical solution, being its biggest advantage over the SU-ERM method. In addition, for the classical one-step M-estimator to achieve optimal efficiency, the initial estimator $\widehat{\theta}_0$ has to satisfy $\|\widehat{\theta}_0-\theta^*\|_2 = O_{\mathbb{P}}((p/N)^{1/2})$, but to achieve the same efficiency based on the OS-ERM method, $\widehat{\theta}_0$ only has to satisfy $\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}((p/n)^{1/2})$. We provide theoretical upper bounds for the approximation errors and the mean squared errors (MSE) of the estimators obtained from the SU-ERM and OS-ERM methods, and show that under some regularity conditions, these estimators are as asymptotically efficient as the global U-estimator. As well, we develop procedures of distributed statistical inference, including methods for estimating covariance matrices and constructing confidence intervals.

The rest of this paper is organized as follows. In Section 2, we review the literature of distributed algorithms and U-statistic-based ERM. Section 3 introduces the problem setup. The two distributed methods for conducting U-statistic-based ERM in the face of big data are introduced in Section 4, and their theoretical properties are analyzed in Section 5. In Section 6, we relax a technical condition about the number of machines and develop iterative versions of the methods for handling situations where the sample size is exceeded by the number of machines. Section 7 investigates the properties of the proposed methods in finite

samples. Section 8 concludes. The Appendix provides the proofs of the main theorems. Proofs of other theorems and results related to the degenerate U-Statistics and the naive estimator are relegated to the Online Supplementary Material.

2. A Review of Related Studies

Distributed algorithms related to M-estimation have been studied for sparse linear regression (Lee et al., 2017), generalized linear regression (Chen and Xie, 2014; Cai et al., 2019), M-estimators with cubic-rate (Shi et al., 2018), convex learning and optimization (Arjevani and Shamir, 2015; Chen et al., 2021; Fan et al., 2021), sparse Cox regression (Wang et al., 2021), linear support vector machine (Wang et al., 2019), and quantile regression (Volgushev et al., 2019; Chen et al., 2019; Chen and Zhou, 2019). Distributed algorithms not involving M-estimation have also been considered in the context of hypothesis testing (Battey et al., 2018), Bayesian estimation (Suchard et al., 2010; Wang and Dunson, 2013; Scott et al., 2016; Terenin et al., 2020), principal component analysis (Garber et al., 2017; Fan et al., 2019), and bootstraping (Kleiner et al., 2014).

Lin and Xi (2010) and Chen and Peng (2021) proposed surrogate versions of U-statistics for big data that can be computed in a distributed framework. The surrogate U-statistics are obtained by weighting and aggregating the local U-statistics based on different subsets of the full sample. They proved that the surrogate U-statistic is as efficient as the global U-statistic computed from the full sample. Xi and Lin (2016) introduced a distributed estimation method for the U-statistic-based functional regression model (U-FRM) for estimating higher-order moments. Although the estimating equations that underlie their U-FRM are U-statistics-based, they are different from the ERM framework being considered in the present paper. Also of relevance to our work are the studies of Vogel et al. (2019) and Wang et al. (2019). Vogel et al. (2019) proposed a distributed stochastic gradient descent (SGD) algorithm for pairwise empirical risk minimization that involves alternating between repartitioning the full data across local machines and in-parallel computation of gradients. An important shortcoming of the SGD method is that when there is a large amount of data, the method requires hundreds if not thousands of iterations in order to converge. As well, as each SGD iteration involves a divide-and-conquer procedure, the communication cost can be enormous. Determining the number of SGD iterations to balance the trade-off between accuracy and cost can also be difficult. Wang et al. (2019) investigated the distributed pairwise learning based on SGD under the framework of a reproducing kernel Hilbert space (RKHS).

3. Preliminaries and Problem Formulation

In this section, we describe the framework of our analysis. Let $\mathcal{P}$ be an unknown probability distribution over the sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a domain in $\mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}$. Let $\{\ell(\theta; z, z') : \Theta \times \mathcal{Z}^2 \mapsto \mathbb{R} \mid \theta \in \Theta \subseteq \mathbb{R}^p\}$ denote a collection of convex loss functions that are assumed to be continuously twice-differentiable with respect to θ. For each parameter θ, the loss function ℓ measures its performance on a pair of instances (z, z'). The corresponding population risk is $\mathcal{L}_0(\theta) := \mathbb{E}_{(Z,Z')\sim\mathcal{P}\times\mathcal{P}}[\ell(\theta; Z, Z')]$. Without loss of generality, we assume that ℓ is symmetric with respect to (z, z').

3.1 The U-statistics-based Empirical Risk Minimization

Our goal is to estimate the parameter of interest θ^* that minimizes the population risk, that is,

$$\theta^* = \arg\min_{\theta\in\Theta} \mathcal{L}_0(\theta) = \arg\min_{\theta\in\Theta} \mathbb{E}_{(Z,Z')\sim\mathcal{P}\times\mathcal{P}}[\ell(\theta; Z, Z')]. \tag{2}$$

In order to obtain the estimate of θ^* in (2), an estimation of the unknown $\mathcal{L}_0$ is required. Denote a sample of i.i.d. observations as $\mathcal{D}_N := \{Z_i = (X_i^\top, Y_i)^\top : i = 1, \ldots, N\}$. Let $\mathscr{C}_N^2$ denote the number of all possible pairs of observations in $\mathcal{D}_N$. A natural way to estimate $\mathcal{L}_0$ is to average the values of ℓ over $\mathscr{C}_N^2$. This leads to the following global empirical risk measure:

$$\mathcal{L}_N(\theta) = \frac{1}{\mathscr{C}_N^2} \sum_{1\le i<j\le N} \ell\left(\theta; Z_i, Z_j\right). \tag{3}$$

Clearly, $\mathcal{L}_N$ is a U-statistic, which has the favourable property of minimum variance in the class of all unbiased estimators of $\mathcal{L}_0$. The minimization with respect to (3) is known as empirical risk minimization (ERM). We denote the resultant minimizer as $\widehat{\theta}_N$, that is, $\widehat{\theta}_N = \arg\min_{\theta\in\Theta}\mathcal{L}_N(\theta)$.

Remark 1 *When the loss function is univariate, the estimators obtained by ERM are referred to as the M-estimators, which are a broad class of estimators for which the objective function corresponds to a sample average. For example, when the loss function is a negative log-likelihood, the M-estimator is the maximum likelihood estimator. We label the univariate loss-based ERM as the usual ERM and the pairwise loss-based ERM as the U-ERM.*

Remark 2 *Our analysis can be generalized to loss functions that average over all d-tuples ($d \ge 2$) of observations. For ease of illustration, we consider $d = 2$ in this paper.*

3.2 Examples

The U-ERM framework can cover a wide range of statistical problems. Some examples are provided below.

Example 1 (Ranking Problems). Ranking problems abound in information retrieval, credit-risk screening, and quality control. Let $Z_1 = (X_1, Y_1)$ and $Z_2 = (X_2, Y_2)$ be objects, where X_1 and X_2 are random vectors describing the objects' features, and Y_1 and Y_2 are random variables that define the ordering between the objects. For example, Z_1 is said to be "better" than Z_2 if $Y_1 > Y_2$. We can observe X but not Y. The goal of ranking is to determine the ordering based upon the observed features of the objects by finding a rule $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that Z_1 is ranked higher than Z_2 if $f(X_1, X_2) > 0$. Here, we focus on linear ranking rules, that is, from a collection of ranking rules $\{f_\theta(x_1, x_2) = \theta^\top(x_1 - x_2), \theta \in \mathbb{R}^p\}$, we seek the rule (or the optimal θ) that minimizes the following population ranking risk:

$$\mathcal{L}_0(\theta) = \mathbb{E}\phi\left[\operatorname{sign}(Y_1 - Y_2)\,\theta^\top(X_1 - X_2)\right],$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a loss function, for example, the logistic loss $\phi(u) = \ln(1 + e^{-u})$ as in Laporte et al. (2013), or the exponential loss $\phi(u) = \exp(-u)$ as in Freund et al. (2003). As

$\mathcal{D}_N$ contains only the observed features, the most natural estimator of $\mathcal{L}_0$ is the empirical risk defined by the U-statistic

$$\mathcal{L}_N(\theta) = \frac{1}{N(N-1)} \sum_{i \neq j} \phi \left[\text{sign}\left(Y_i - Y_j\right) \theta^\top \left(X_i - X_j\right) \right].$$

Example 2 (Rank-based coefficient estimation for censored regression models). A popular model used in survival analysis is the following accelerated failure time (AFT) model that assumes the logarithm of the failure time T_i to be linearly related to the associated covariates X_i:

$$\log T_i = X_i^\top \theta + \zeta_i, \tag{4}$$

where θ is a $p \times 1$ vector of unknown regression parameters and $\{\zeta_i, i = 1, \ldots, N\}$ are i.i.d. random errors with an unspecified distribution function independent of $\{X_i, i = 1, \ldots, N\}$. Usually, T_i is subject to censoring at time C_i, and the observed dataset is $\{Z_i = (\widetilde{T}_i, \delta_i, X_i), i = 1, \cdots, N\}$, where $\widetilde{T}_i = \min\left(T_i, C_i\right)$ is the observed failure time and $\delta_i = I\left(T_i \leq C_i\right)$ is the censoring indicator. One technique often used for estimating AFT models is the Gehan estimator (Fygenson and Ritov, 1994), obtained by minimizing the following objective function in the form of a U-statistic:

$$F_N(\theta) = \frac{1}{\mathscr{C}_N^2} \sum_{i \neq j}^{N} \delta_i \left(e_j(\theta) - e_i(\theta)\right) I\left(e_i(\theta) \leq e_j(\theta)\right), \tag{5}$$

where $e_i(\theta) = \log \widetilde{T}_i - X_i^\top \theta$. The non-smooth nature of (5) introduces challenges to the computation of coefficient estimates and their standard errors. To reconcile this difficulty, Brown and Wang (2007) developed an induced smoothing method that replaces the non-smooth objective function by a smooth approximation. Specifically, let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal cumulative distribution and density functions respectively. Brown and Wang (2007) obtained their estimator of θ by minimizing the following empirical risk:

$$\frac{1}{\mathscr{C}_N^2} \sum_{i \neq j}^{N} \delta_i \left[\left(e_j(\theta) - e_i(\theta)\right) \Phi\left(\frac{e_j(\theta) - e_i(\theta)}{r_{ij}}\right) + r_{ij} \phi\left(\frac{e_j(\theta) - e_i(\theta)}{r_{ij}}\right) \right], \tag{6}$$

where $r_{ij}^2 = \frac{1}{N}(X_i - X_j)^\top \Sigma (X_i - X_j)$, and Σ is a symmetric and positive definite matrix that satisfies $\|\Sigma^{1/2}\| = O(1)$.

Example 3 (Partially linear logit model). Consider the partially linear logit model

$$y_i = I\left\{ x_i^\top \theta + g\left(w_i\right) + \varepsilon_i \geq 0 \right\}, \quad i = 1, \ldots, N,$$

where $\{\varepsilon_i\}_{i=1}^N$ are i.i.d random errors from the logistic distribution and $g(\cdot)$ is an unknown function assumed to be sufficiently smooth. As $g(w_i)$ and $g(w_j)$ behave like fixed effects when w_i is close to w_j, Honoré and Powell (1997) suggested estimating θ by optimizing the following empirical risk based on pairwise differencing:

$$\frac{1}{\mathscr{C}_N^2} \sum_{\substack{i<j \\ y_i \neq y_j}} K\left(\frac{w_i - w_j}{h_N}\right) \left(y_i \ln\left(1 + \exp\left(\left(x_j - x_i\right)^\top \theta\right)\right) + y_j \ln\left(1 + \exp\left(\left(x_i - x_j\right)^\top \theta\right)\right)\right),$$

where the bandwidth h_N shrinks towards zero as N increases, and $K(\cdot)$ is a kernel function that gives a large weight to the pair (i, j) when w_i and w_j are close. Similar approaches exist for the estimation of other semiparametric models including the partially linear Tobit and Poisson regression models (Honoré and Powell, 1997).

There exist many other examples, including the AUC maximization (Gao et al., 2013) based on the regularized misranking loss function $\ell(\theta, z, z') = I_{\{(x-x')^\top\theta<0\}} I_{\{y=1\wedge y'=-1\}} + (\mu/2)\|\theta\|^2$ that has been widely used in binary classification and bipartite ranking. For computational convenience, ℓ is usually replaced by a surrogate loss function $\widetilde{\ell}(\theta, z, z') = (1-(x-x')^\top\theta)^2 I_{\{y=1\wedge y'=-1\}} + (\mu/2)\|\theta\|^2$. Another example is metric learning (Papa et al., 2015), where the empirical risk is defined by a U-statistic of degree 3.

4. Distributed Estimation for Large-scale Datasets

In this section, we focus on U-ERM in the context of big data, where N is exceedingly large and p grows with N. As mentioned, the computational complexity of the problem increases quadratically with the sample size due to the pairwise feature of the loss functions. The high volume of sample pairs poses formidable challenges to the implementation of U-ERM in terms of both memory requirements and computational cost. A single machine cannot accommodate the memory required for processing $O(N^2)$ sample pairs. Enormous computational operations are also needed for performing numerical iterative algorithms on such a high volume of data. The overall computational cost is therefore very high, even for moderate N. The way to proceed is to develop a computationally efficient and statistically valid divide-and-conquer algorithm for the U-ERM, to be conducted in a distributed manner rather than on a centralized machine.

To this end, let the data $\mathcal{D}_N$ be evenly divided into K smaller subsets $\{\mathcal{D}_k\}_{k=1}^K$, each of size n. The subset $\mathcal{D}_k := \left\{Z_{k,i} = (X_{k,i}^\top, Y_{k,i})^\top, i = 1, \ldots, n\right\}$ is stored on the k-th machine, $k = 1, \ldots, K$. One of these K machines is a master machine on which the summary statistics computed from local machines are aggregated. No machine other than the master machine can access data not produced by itself. Without loss of generality, we assume that the first machine is the master machine. In the following, we develop two simple and reliable distributed estimation methods for U-ERM.

4.1 The SU-ERM Estimator

The local empirical risk corresponding to the k-th machine is

$$\mathcal{L}_n^k(\theta) = \frac{1}{\mathscr{C}_n^2} \sum_{1\le i<j\le n} \ell\left(\theta; Z_{k,i}, Z_{k,j}\right) \quad \text{for} \quad k = 1, \ldots, K. \tag{7}$$

A natural way to obtain a distributed estimator with the same statistical efficiency as the global U-estimator $\widehat{\theta}_N$ is to perform ERM based on $\mathcal{S}_N$, a surrogate empirical risk computed in a distributed manner that is asymptotically equivalent to the global risk $\mathcal{L}_N$. Specifically, consider the Taylor's expansion of $\mathcal{L}_N$ around an initial value $\widehat{\theta}_0$:

$$\mathcal{L}_N(\theta) = \mathcal{L}_N(\widehat{\theta}_0) + \nabla\mathcal{L}_N(\widehat{\theta}_0)^\top \left(\theta - \widehat{\theta}_0\right) + \sum_{m=2}^{\infty} \frac{1}{m!} \nabla^m \mathcal{L}_N(\widehat{\theta}_0)(\theta - \widehat{\theta}_0)^{\otimes m}. \tag{8}$$

Here, $\widehat{\theta}_0$ can be any initial estimator such as the local U-estimator $\widehat{\theta}_n^k$ obtained through optimizing any local empirical risk $\mathcal{L}_n^k(\theta)$ for some k, or the naive estimator $\overline{\theta}_N = 1/K \sum_{k=1}^K \widehat{\theta}_n^k$ obtained by averaging all of $\widehat{\theta}_n^k$'s. For the usual ERM for which the loss function depends on a single sample of observations, one can readily obtain the exact values of $\mathcal{L}_N$ and the corresponding m-th order derivatives $\nabla^m \mathcal{L}_N$ $(m \geq 1)$ in a single round of communications. On the other hand, as $\mathcal{L}_N$ and $\nabla^m \mathcal{L}_N$ $(m \geq 1)$ are U-statistics that involve all pairs of sample observations, it is difficult to obtain all of their values even after multi-rounds of communications between the local machines. To reconcile this difficulty, we replace $\mathcal{L}_N$ and $\nabla^m \mathcal{L}_N$ $(m \geq 1)$ at $\widehat{\theta}_0$ by their surrogates

$$\widetilde{\mathcal{L}}_N(\widehat{\theta}_0) = \frac{1}{K}\sum_{k=1}^K \mathcal{L}_n^k(\widehat{\theta}_0) \quad \text{and} \quad \nabla^m \widetilde{\mathcal{L}}_N(\widehat{\theta}_0) = \frac{1}{K}\sum_{k=1}^K \nabla^m \mathcal{L}_n^k(\widehat{\theta}_0), \text{ for } m \geq 1 \tag{9}$$

respectively in the Taylor's expansion in (8). Given $\widehat{\theta}_0$, these surrogate U-statistics can be calculated in parallel in a distributed manner such that machine k computes $\mathcal{L}_n^k(\widehat{\theta}_0)$ or $\nabla^m \mathcal{L}_n^k(\widehat{\theta}_0)$ using the subset $\mathcal{D}_k$, $k = 1, \cdots, K$. These local U-statistics will then be transferred to the master machine for aggregation.

The surrogates defined in (9) can be viewed as incomplete U-statistics as in Clémençon et al. (2016), who derived uniform deviation results for the estimation errors using incomplete U-statistics and compared them with their complete counterparts; the objective is to reduce the computational burden by reducing the number of pairs involved in the summation of U-statistics, yet retaining the original convergence rate.

In a distributed environment, the data are stored in or artificially assigned to local machines that do not communicate with one another. Under this setting, it is difficult if not impossible to obtain the complete U-statistic $\mathcal{L}_N(\theta)$. On the other hand, the U-statistic $\mathcal{L}_n^k(\theta)$ based on data on a local machine can be computed with complexity equal to $O(n)$. Because $\mathcal{L}_n^k(\theta)$ is complete, information on all local machines is used in the aggregate statistic $\widetilde{\mathcal{L}}_N(\theta) = \sum_{k=1}^K \mathcal{L}_n^k(\theta)$; this approach also has the advantage of reducing the computational burden as the computation of $\mathcal{L}_n^k(\theta)$ uses significantly fewer observation pairs than the computation of $\mathcal{L}_N(\theta)$ in a centralized setting. If computational burden remains an issue, other incomplete U-statistics may be used to approximate $\mathcal{L}_n^k(\theta)$.

Note that the aggregation of the higher-order derivatives $\nabla^m \mathcal{L}_n^k$ $(m \geq 2)$ requires a vast amount of communication exchanges between the master and local machines. To reduce the communication cost, we can replace the surrogate higher-order derivatives $\nabla^m \widetilde{\mathcal{L}}_N(\widehat{\theta}_0)$ $(m \geq 2)$ by the local derivatives computed from the subset stored in the master machine. This leads to the following surrogate empirical risk:

$$\mathcal{S}_N(\theta) = \widetilde{\mathcal{L}}_N(\widehat{\theta}_0) + \nabla \widetilde{\mathcal{L}}_N(\widehat{\theta}_0)^\top (\theta - \widehat{\theta}_0) + \sum_{m=2}^{\infty} \frac{1}{m!} \nabla^m \mathcal{L}_n^1(\widehat{\theta}_0)(\theta - \widehat{\theta}_0)^{\otimes m}. \tag{10}$$

Again, by using the Taylor's expansion of $\mathcal{L}_n^1(\theta)$ around $\widehat{\theta}_0$, the infinite sum of the higher-order terms in (10) can be replaced by the term $\mathcal{L}_n^1(\theta) - \mathcal{L}_n^1(\widehat{\theta}_0) - \nabla \mathcal{L}_n^1(\widehat{\theta}_0)^\top (\theta - \widehat{\theta}_0)$, yielding

$$\mathcal{S}_N(\theta) = \widetilde{\mathcal{L}}_N(\widehat{\theta}_0) + \nabla \widetilde{\mathcal{L}}_N(\widehat{\theta}_0)^\top (\theta - \widehat{\theta}_0) + \mathcal{L}_n^1(\theta) - \mathcal{L}_n^1(\widehat{\theta}_0) - \nabla \mathcal{L}_n^1(\widehat{\theta}_0)^\top (\theta - \widehat{\theta}_0). \tag{11}$$

After omitting the constant term in (11), the surrogate empirical risk reduces to

$$\mathcal{S}_N(\theta) = \mathcal{L}_n^1(\theta) - \left(\nabla\mathcal{L}_n^1(\widehat{\theta}_0) - \nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0)\right)^\top\theta. \tag{12}$$

We abbreviate the ERM based on this surrogate empirical risk as the SU-ERM, and denote the following estimator obtained by minimizing $\mathcal{S}_N(\theta)$ as the SU-ERM estimator:

$$\widetilde{\theta}_N = \arg\min_{\theta\in\Theta}\mathcal{S}_N(\theta). \tag{13}$$

In summary, by the SU-ERM procedure, the gradients are computed based on $K-1$ local machines with the rest of the computational operations performed on the master machine. Compared with the U-ERM method conducted on a single machine, the SU-ERM method reduces the computational complexity from $O(N^2)$ to $O(n^2)$, which in turn reduces the risk of memory overflow caused by the large number of sample pairs.

4.2 The OS-ERM Estimator

Despite the advantages of the SU-ERM method, the computational complexity of optimizing the surrogate empirical risk $\mathcal{S}_N$ is still high even for moderate sample sizes. To further reduce the computational cost, we consider another distributed method based on a one-step approach. For classical M-estimation, Bickel (1975) developed a one-step estimator based on the Newton's method that improves the initial estimator. van der Vaart (1998) proved that the same one-step estimator can be as efficient as the M-estimator. In the following, we investigate this one-step method for U-ERM under the distributed setting.

The procedure entails a quadratic approximation to the surrogate empirical risk $\mathcal{S}_N$. By applying the Taylor's expansion, we can write

$$\mathcal{S}_N(\theta) \approx \mathcal{S}_N(\widehat{\theta}_0) + \nabla\mathcal{S}_N(\widehat{\theta}_0)^\top(\theta-\widehat{\theta}_0) + \frac{1}{2}(\theta-\widehat{\theta}_0)^\top\nabla^2\mathcal{S}_N(\widehat{\theta}_0)(\theta-\widehat{\theta}_0). \tag{14}$$

Recognizing that $\nabla\mathcal{S}_N(\widehat{\theta}_0) = \nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0)$ and $\nabla^2\mathcal{S}_N(\widehat{\theta}_0) = \nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)$, and omitting the constant terms in (14), we obtain the following quadratic surrogate empirical risk function:

$$\mathcal{S}_N^Q(\theta) = \nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0)^\top(\theta-\widehat{\theta}_0) + \frac{1}{2}(\theta-\widehat{\theta}_0)^\top\nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)(\theta-\widehat{\theta}_0). \tag{15}$$

The OS-ERM estimator $\widetilde{\theta}_N^Q$ is the minimizer of (15) and has the following closed-form expression:

$$\widetilde{\theta}_N^Q = \widehat{\theta}_0 - \nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)^{-1}\nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0). \tag{16}$$

Similar to the SU-ERM procedure, the OS-ERM method uses $K-1$ machines to compute the local gradients, which are then transferred to the master machine for the remaining processing. However, unlike the SU-ERM method, the OS-ERM method is based on an empirical risk function with a closed form expression. This is an important advantage as no optimization would be required at the master machine. This significantly reduces the computational effort and offers a fast solution compared to U-ERM in a distributed setting. Furthermore, as will be shown in Section 5, the resultant OS-ERM estimator $\widetilde{\theta}_N^Q$ enjoys the same statistical efficiency as the SU-ERM estimator $\widetilde{\theta}_N$. These credentials strongly favour the use of the OS-ERM method in practice.

5. Theoretical Results

This section is devoted to an analysis of the theoretical properties of the SU-ERM and OS-ERM estimators. Let $U(\theta^*, \rho) = \{\theta \in \mathbb{R}^p : \|\theta - \theta^*\|_2 < \rho\}$ be a ball around θ^* with radius $\rho \in (0, 1)$. Our theoretical analysis requires the following regularity conditions:

Condition 1 The dimension p, the number of machines K and the total sample size N satisfy $p = o(N)$ and $\log K = o(N)$ as $\min(p, n, K) \to \infty$.

Condition 2 The parameter space $\Theta \subset \mathbb{R}^p$ is a compact convex set, and θ^* is an interior point of Θ.

Condition 3 The Hessian matrix $\nabla^2 \mathcal{L}_0(\theta)$ of the population risk at θ^* is positive definite, and there exist two positive parameters (λ_-, λ_+) such that $\lambda_- I_{p\times p} \preceq \nabla^2 \mathcal{L}_0(\theta^*) \preceq \lambda_+ I_{p\times p}$.

Condition 4 There exists a compact neighborhood $K \subset \Theta$ of θ^* such that

$$\inf_{\theta \in \Theta \setminus K} \mathcal{L}_0(\theta) > \mathcal{L}_0(\theta^*), \quad a.s.$$

Condition 5 There exists a function $M(z, z')$ such that for all $\theta, \theta' \in U(\theta^*, \rho)$, the loss function $\ell(\theta; z, z')$ and $\ell(\theta'; z, z')$ satisfy

$$\left\|\nabla^2 \ell(\theta; z, z') - \nabla^2 \ell\left(\theta'; z, z'\right)\right\|_2 \leq M(z, z') \left\|\theta - \theta'\right\|_2.$$

As well, it is assumed that $\mathbb{E}\left[M^8(Z, Z')\right] \leq M^8$ for some constant $M > 0$.

Condition 6 There exist a constant $\tau > 0$, and functions $G(\cdot)$ and $H(\cdot)$ satisfying $\mathbb{E}G^{16}(Z) \leq G^{16}$ and $\mathbb{E}H^{16}(Z) \leq H^{16}$ for some constants G and H, such that for all $\theta \in U(\theta^*, \rho)$.

$$\mathbb{E}\left[\|\nabla \ell(\theta; Z, z_0)\|_2^{16}\right] \leq p^8 G^{16}(z_0) \quad \text{and} \quad \mathbb{E}\left[\left\|\nabla^2 \ell(\theta; Z, z_0)\right\|_2^{16}\right] \leq p^\tau H^{16}(z_0),$$

Condition (C1) guarantees the consistency of the global estimator as $\|\widehat{\theta}_N - \theta^*\|_2 = O_{\mathbb{P}}(\sqrt{p/N})$. Note that under Condition (C1), $\log p = o(n)$ on any local machine. Condition (C2) informs the relationship between θ^* and the parameter space. Condition (C3) assumes local convexity of the population risk, which guarantees θ^* to be a local minimum. Condition (C4) is a standard model identifiability condition that guarantees the consistency of estimators. Conditions (C2)-(C4) are standard conditions for U-estimation (Honoré and Powell, 1994; Wang et al., 2009; Bose and Chatterjee, 2018a). Condition (C5) restricts the first and second derivatives of the loss function ℓ to be bounded, and places a Lipschitz condition for ∇^2 to hold at least in a small neighbourhood of θ^*. Condition (C6) pertains to the projection of ℓ, and is necessary for establishing the moment inequalities of the empirical risk and its derivatives; see Lemma 14 in the Appendix. Note that Condition (C6) implies $\mathbb{E}\left[\|\nabla \ell(\theta; Z, Z')\|_2^{16}\right] \leq p^8 G^{16}$ and $\mathbb{E}\left[\left\|\nabla^2 \ell(\theta; Z, Z')\right\|_2^{16}\right] \leq p^\tau H^{16}$. The first part of Condition (C6) holds if the high-order moments of the p-dimensional random vector $\nabla \ell(\theta; Z, Z')$ is bounded. In particular, Condition (C6) results in the condition of $\mathbb{E}\left[\|\nabla \ell(\theta; Z, z_0)\|_2^2\right] \leq pG^2(z_0)$, which is satisfied if all the coordinates of $\nabla \ell(\theta; Z, z_0)$ are

uniformly bounded, as in Tu et al. (2021). The second part of Condition (C6) introduces a constant τ to allow flexibility for the rate of $\mathbb{E}\left[\left\|\nabla^2\ell(\theta;Z,z_0)\right\|_2^{16}\right]$, so that our framework can cover a diverse range of situations. Under Examples 1-3 above, we can set τ to 16, which is a common choice of τ. The following is an example where other values of τ may be selected.

Example 4 (U-means). Let $\mathcal{K}(\cdot):\mathbb{R}^2\to\mathbb{R}^p$ be a p-dimensional symmetric kernel function. Define the loss function

$$\ell(\theta;x_1,x_2)=(\mathcal{K}(x_1,x_2)-\theta)^\top\Sigma^{-1}(\mathcal{K}(x_1,x_2)-\theta).$$

where Σ is a positive definite matrix. If Σ is selected as the covariance matrix of $\mathcal{K}(X_1,X_2)$, we obtain the so-called Mahalanobis distance between $\mathcal{K}(x_1,x_2)$ and θ. It can be shown that $\nabla^2\ell(\theta;X_1,X_2)=\Sigma^{-1}$. Hence the parameter τ in Condition (C6) is determined by the rate of the largest eigenvalue of Σ^{-1}.

Heuristically, under the above conditions, given that $\nabla\mathcal{S}_N(\widetilde{\theta}_N)\approx\nabla\mathcal{S}_N(\theta^*)+\nabla^2\mathcal{S}_N(\theta^*)(\widetilde{\theta}_N-\theta^*)$, the errors of the SU-ERM estimator $\widetilde{\theta}_N$ with respect to the true parameter θ^* may be approximated by

$$\begin{aligned}
&\|\widetilde{\theta}_N-\theta^*\|_2\\
\approx&O_{\mathbb{P}}(\|\nabla\mathcal{S}_N(\theta^*)\|_2)\\
=&O_{\mathbb{P}}(\|\nabla\mathcal{L}_n^1(\theta^*)-\nabla\mathcal{L}_n^1(\widehat{\theta}_0)-(\nabla\widetilde{\mathcal{L}}_N(\theta^*)-\nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0))+\nabla\widetilde{\mathcal{L}}_N(\theta^*)\|_2)\\
=&O_{\mathbb{P}}(\|(\nabla^2\mathcal{L}_n^1(\theta^*)-\nabla^2\widetilde{\mathcal{L}}_N(\theta^*))(\theta^*-\widehat{\theta}_0)\|_2)+O_{\mathbb{P}}((p/n)^{1/2}\|\theta^*-\widehat{\theta}_0\|_2^2)+O_{\mathbb{P}}((p/N)^{1/2})\\
=&O_{\mathbb{P}}((p/n)^{1/2}\|\widehat{\theta}_0-\theta^*\|_2)+O_{\mathbb{P}}((p/N)^{1/2}),
\end{aligned}\tag{17}$$

provided that the initial estimator satisfies $\widehat{\theta}_0-\theta^*=O_{\mathbb{P}}((p/n)^{1/2})$. The statistical efficiency of $\widetilde{\theta}_N$ is of the same order of magnitude as that of the global U-estimator $\widehat{\theta}_N$ (that is, $\|\widehat{\theta}_N-\theta^*\|_2=O_{\mathbb{P}}((p/N)^{1/2})$), if $pK=O(n)$. In particular, under the case of $pK<n$, this condition is satisfied when the local U-estimator $\widehat{\theta}_n^k$ obtained from the k-th machine or the naive estimator $\overline{\theta}_N$ is used as the initial estimator. We now present a result on the approximated errors of $\widetilde{\theta}_N$.

Proposition 3 *Let Conditions (C1)-(C6) be satisfied. In particular, there exists $\zeta>0$ such that $p=O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau>0$. Assume the initial estimator $\widehat{\theta}_0$ lies in $U(\theta^*,\tilde{\rho})$, where $\tilde{\rho}=\min\{(1-\rho)\lambda_-\delta_\rho/(32\lambda_+),\sqrt{(1-\rho)\lambda_-\delta_\rho/(32M)}\}$ with $\delta_\rho=\min\{\rho,\rho\lambda_-/(4M)\}$. Then the SU-ERM estimator $\widetilde{\theta}_N$ satisfies*

$$\begin{aligned}
\|\widetilde{\theta}_N-\widehat{\theta}_N\|_2\leq&C\left(\|\widehat{\theta}_0-\widehat{\theta}_N\|_2+\|\widehat{\theta}_N-\theta^*\|_2+\|\nabla^2\mathcal{L}_n^1(\theta^*)-\nabla^2\mathcal{L}_N(\theta^*)\|_2\right)\|\widehat{\theta}_0-\widehat{\theta}_N\|_2\\
&+C\left(\|\widehat{\theta}_0-\theta^*\|_2+\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\widetilde{\mathcal{L}}_N(\theta^*)\|_2\right)\|\widehat{\theta}_0-\theta^*\|_2\\
&+\|\nabla\mathcal{L}_N(\theta^*)-\nabla\widetilde{\mathcal{L}}_N(\theta^*)\|_2
\end{aligned}\tag{18}$$

with probability no smaller than $1-C'\left(K(\log p)^8/n^8+p^8/N^8\right)$, where C and C' are independent of (K,n,N,p).

Note that the parameter τ in Condition (C6) determines the condition on p in Proposition 3. The smaller the value of τ, the faster the convergence rate of p. For example, if $\tau = 16$, Proposition 3 requires $p = O(n^{(1-\frac{1+\zeta}{8})/2})$ for some $\zeta > 0$, but if $\tau = 8$, then $p = O(n^{(1-\frac{1+\zeta}{8})})$.

As will be shown in the Appendix, $\|\nabla^2\mathcal{L}_n^1(\theta^*) - \nabla^2\mathcal{L}_N(\theta^*)\|_2 = O_{\mathbb{P}}[\{(\log p)/n\}^{1/2}]$, $\|\nabla^2\mathcal{L}_N(\theta^*) - \nabla^2\widetilde{\mathcal{L}}_N(\theta^*)\|_2 = O_{\mathbb{P}}[\{(\log p)/n\}^{1/2}]$, and $\|\nabla\mathcal{L}_N(\theta^*) - \nabla\widetilde{\mathcal{L}}_N(\theta^*)\|_2 = O_{\mathbb{P}}[\{p/(nN)\}^{1/2}]$. A key takeaway of Proposition 3 is that the SU-ERM estimator's performance can be improved by updating the estimates iteratively. For example, if $\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}((p/n)^{1/2})$, and when $\widehat{\theta}_n^1$ is used as the initial estimator, we have $\|\widetilde{\theta}_N - \widehat{\theta}_N\|_2 = O_{\mathbb{P}}((p/n)^{1/2})\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 + O_{\mathbb{P}}[\{p/(nN)\}^{1/2}]$. This process can reduce the estimator's errors by a factor of $O((p/n)^{1/2})$. The following theorem presents the upper bound of the MSE of $\widetilde{\theta}_N$ for estimating θ^*.

Theorem 4 *Let Conditions (C1)-(C6) be satisfied. In particular, there exists $\zeta > 0$ such that $p = O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau > 0$. Assume the initial estimator $\widehat{\theta}_0$ lies in $U(\theta^*, \tilde{\rho})$, where $\tilde{\rho} = \min\{(1-\rho)\lambda_-\delta_\rho/(32\lambda_+), \sqrt{(1-\rho)\lambda_-\delta_\rho/(32M)}\}$ with $\delta_\rho = \min\{\rho, \rho\lambda_-/(4M)\}$. Then the MSE of the SU-ERM estimator $\widetilde{\theta}_N$ is bounded above by*

$$\mathbb{E}\left[\left\|\widetilde{\theta}_N - \theta^*\right\|_2^2\right] \leq \frac{4A}{N} + C\left(\gamma_{n,K,p}^2(\widehat{\theta}_0) + \frac{\gamma_{n,K,p}(\widehat{\theta}_0)\sqrt{A}}{\sqrt{N}} + \frac{p\log N\log\log N}{N^2}\right), \tag{19}$$

where $A = \mathbb{E}[\left\|\nabla^2\mathcal{L}_0(\theta^*)^{-1} g(\theta^*; Z)\right\|_2^2]$ with $g(\theta; Z) = \mathbb{E}[\nabla\ell(\theta; Z, Z')\,|Z]$, C is some constant independent of (K, n, N, p), and

$$\gamma_{n,K,p}^2(\widehat{\theta}_0) = \sqrt{\mathbb{E}\left[\|\widehat{\theta}_0 - \theta^*\|_2^4\right]}\max\left\{\frac{p}{N}, \frac{\log p}{n}, \sqrt{\mathbb{E}\left[\|\widehat{\theta}_0 - \theta^*\|_2^4\right]}\right\} + \left(\frac{p}{N}\max\left\{\frac{p}{N}, \frac{\log p}{n}\right\} + \frac{K(\log p)^8}{n^8}\right).$$

Theorem 4 implies the consistency of $\widetilde{\theta}_N$, which is formalized in the following theorem.

Theorem 5 *Let Conditions (C1)-(C6) be satisfied. In particular, there exists $\zeta > 0$ such that $p = O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau > 0$. Assume the initial estimator $\widehat{\theta}_0$ lies in $U(\theta^*, \tilde{\rho})$, where $\tilde{\rho} = \min\{(1-\rho)\lambda_-\delta_\rho/(32\lambda_+), \sqrt{(1-\rho)\lambda_-\delta_\rho/(32M)}\}$ with $\delta_\rho = \min\{\rho, \rho\lambda_-/(4M)\}$, and $\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}(\alpha_{n,K,p})$, where $\alpha_{n,K,p} = \sqrt{1/K}$ or $\sqrt{p/n}$. Then the SU-ERM estimator satisfies*

$$\|\widetilde{\theta}_N - \theta^*\|_2 = O_{\mathbb{P}}\left(\alpha_{n,K,p}^2 + \sqrt{\frac{p}{N}} + \frac{\sqrt{K}(\log p)^4}{n^4}\right).$$

Furthermore, if $\log p = o(nK^{-1/8})$ and $\alpha_{n,K,p} = o(1)$, the SU-ERM estimator $\widetilde{\theta}_N$ is consistent.

The convergence rates $\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}(K^{-1/2})$ and $\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}((p/n)^{1/2})$ can be achieved by setting $\widehat{\theta}_0$ to be the naive estimator $\overline{\theta}_N = 1/K\sum_{k=1}^K \widehat{\theta}_n^k$ and the local U-estimator $\widehat{\theta}_n^1$ respectively. Theorem 5 provides the conditions for the consistency of $\widetilde{\theta}_N$

under the case of diverging p. If p is fixed, and $\alpha_{n,K,p} = K^{-1/2}$ or $\alpha_{n,K,p} = (p/n)^{1/2}$, the consistency conditions are equivalent to $K = o(n^8)$. Specifically, if $\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}(K^{-1/2})$, then $\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|_2 = O\left(\frac{1}{K^2} + \frac{1}{N} + \frac{K}{n^8}\right)$; additionally, if $C_1 n \leq K \leq C_2 n^{7/2}$ for some constants $C_1, C_2 > 0$, the leading term of $\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|_2^2$ is $O(1/N)$, which is of the same order as that of $\mathbb{E}\|\widehat{\theta}_N - \theta^*\|_2^2$. This indicates that the SU-ERM estimator enjoys the same efficiency as the global U-estimator. Similarly, if $\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}((p/n)^{1/2})$, then $\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|_2 = O\left(\frac{1}{n^2} + \frac{1}{N} + \frac{K}{n^8}\right)$, and additionally, if $K = O(n)$, then $\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|_2^2 = O(1/N)$. For the case of $K = O(n^\gamma)$ with $\gamma > 1$, the SU-ERM estimator may exhibit slightly worse performance than the global U-estimator. The execution of the right-hand-side of (19) requires an iterative algorithm with $\widetilde{\theta}_N$ as the initial estimate to successively refine the result. Section 6 discusses this algorithm. For the case of diverging p, the global U-estimator satisfies $\mathbb{E}\|\widehat{\theta}_N - \theta^*\|_2^2 = O(p/N)$. The following corollary presents the conditions for $\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|_2$ to achieve the optimal convergence rate.

Corollary 6 *Assume the conditions in Theorem 4 hold. Then (i) when* $\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}(K^{-1/2})$*, and if*

$$n = O(pK) \quad \textit{and} \quad \log p / p^{1/8} = O(n^{-7/8} K^{1/4}),$$

the SU-ERM estimator achieves the optimal convergence rate with respect to MSE, in the sense that $\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|_2^2 = O(p/N)$*, which is also the convergence rate of the global U-estimator. If one assumes the stronger conditions of* $n = o(pK)$ *and* $\log p / p^{1/8} = O(n^{-7/8} K^{1/4})$*, then the MSE of the SU-ERM estimator achieves the optimal upper bound of* $\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|_2^2 = 4A/N + o(p/N)$*; (ii) when* $\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}((p/n)^{1/2})$*, and if*

$$pK = O(n),$$

then $\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|_2^2 = O(p/N)$*, or if* $pK = o(n)$*, then* $\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|_2^2 = 4A/N + o(p/N)$*, which is the optimal upper bound of the MSE.*

Now, let us consider the OS-ERM estimator $\widetilde{\theta}_N^Q$. As $\widetilde{\theta}_N^Q$ is obtained by minimizing a quadratic approximation to the surrogate empirical risk $\mathcal{S}_N$ (that is, $\mathcal{S}_N^Q$ and $\mathcal{S}_N$ agree up to the second-order Taylor's expansion), they share similar theoretical properties. The analogues of Proposition 3 and Theorem 4 that pertain to $\widetilde{\theta}_N^Q$ are presented below.

Proposition 7 *Let Conditions (C1)-(C6) be satisfied. In particular, there exists* $\zeta > 0$ *such that* $p = O(n^{\frac{8}{\tau}(1 - \frac{1+\zeta}{8})})$ *if* $\tau > 0$*. Assume that the initial estimator* $\widehat{\theta}_0$ *lies in* $U(\theta^*, \rho')$*, where* $\rho' = \min\left\{\rho, (16M)^{-1}(1-\rho)\lambda_-\right\}$*. Then the OS-ERM estimator* $\widetilde{\theta}_N^Q$ *satisfies*

$$\begin{aligned}
\|\widetilde{\theta}_N^Q - \widehat{\theta}_N\|_2 \leq & C_1 \left(\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 + \|\widehat{\theta}_N - \theta^*\|_2 + \|\nabla^2 \mathcal{L}_n^1(\theta^*) - \nabla^2 \mathcal{L}_N(\theta^*)\|_2 \right) \|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 \\
& + C_1 \left(\|\widehat{\theta}_0 - \theta^*\|_2 + \|\nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \widetilde{\mathcal{L}}_N(\theta^*)\|_2 \right) \|\widehat{\theta}_0 - \theta^*\|_2 \\
& + \|\nabla \mathcal{L}_N(\theta^*) - \nabla \widetilde{\mathcal{L}}_N(\theta^*)\|_2,
\end{aligned} \tag{20}$$

with probability no smaller than $1 - C_1' \left(K(\log p)^8 / n^8 + p^8 / N^8 \right)$*, where* C_1 *and* C_1' *are independent of* (K, n, N, p)*.*

Theorem 8 *Let Conditions (C1)-(C6) be satisfied. In particular, there exists $\zeta > 0$ such that $p = O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau > 0$. Assume that the initial estimator $\widehat{\theta}_0$ lies in $U(\theta^*, \rho')$, where $\rho' = \min\{\rho, (16M)^{-1}(1-\rho)\lambda_-\}$. Then the MSE of the OS-ERM estimator $\widetilde{\theta}_N^Q$ is bounded above by*

$$\mathbb{E}\left[\left\|\widetilde{\theta}_N^Q - \theta^*\right\|_2^2\right] \leq \frac{4A}{N} + C\left(\gamma_{n,K,p}^2(\widehat{\theta}_0) + \frac{\gamma_{n,K,p}(\widehat{\theta}_0)\sqrt{A}}{\sqrt{N}} + \frac{p\log N\log\log N}{N^2}\right), \tag{21}$$

where $A = \mathbb{E}[\|\nabla^2\mathcal{L}_0(\theta^*)^{-1} g(\theta^*; Z)\|_2^2]$ with $g(\theta; Z) = \mathbb{E}[\nabla\ell(\theta; Z, Z')|Z]$, C is some constant independent of (K, n, N, p), and

$$\gamma_{n,K,p}^2(\widehat{\theta}_0) = \sqrt{\mathbb{E}\left[\|\widehat{\theta}_0 - \theta^*\|_2^4\right]}\max\left\{\frac{p}{N}, \frac{\log p}{n}, \sqrt{\mathbb{E}\left[\|\widehat{\theta}_0 - \theta^*\|_2^4\right]}\right\} + \left(\frac{p}{N}\max\left\{\frac{p}{N}, \frac{\log p}{n}\right\} + \frac{K(\log p)^8}{n^8}\right).$$

Analogous to Theorem 5 and Corollary 6, the OS-ERM estimator $\widetilde{\theta}_N^Q$ is a consistent estimator of θ^* and achieves the same convergence rate to the optimal MSE as the global estimator $\widehat{\theta}_N$ under similar conditions.

We have also derived the theoretical properties of the naive estimator $\bar{\theta}_N = K^{-1}\sum_{k=1}^K \widehat{\theta}_n^k$, including the bound of its MSE, consistency properties and optimal convergence rate. These results facilitate a formal comparison between the naive estimator and our proposed estimators. They are presented in the Online Supplementary Material.

It is well-known that under some regularity conditions (Honoré and Powell, 1994; Wang et al., 2009), the global U-estimator $\widehat{\theta}_N$ satisfies

$$\widehat{\theta}_N - \theta^* = -\nabla^2\mathcal{L}_0(\theta^*)^{-1}\left\{\frac{2}{N}\sum_{i=1}^N g(\theta^*; Z_i) + \frac{2}{N^2}\sum_{1\leq i<j\leq N} h(\theta^*; Z_i, Z_j)\right\} + \varpi_{N,p},$$

where $g(\theta; z) = \mathbb{E}[\nabla\ell(\theta; Z, Z')|Z = z]$ as defined in Theorems 4 and 8, $h(\theta^*; z, z') = \nabla\ell(\theta^*; z, z') - g(\theta^*; z) - g(\theta^*; z')$, and $\|\varpi_{N,p}\|_2 = o_{\mathbb{P}}(\|\widehat{\theta}_N - \theta^*\|_2)$. Let $V_0 = \mathbb{E}[g(\theta^*; Z)g(\theta^*; Z)^\top]$ be the dispersion matrix of the first projection of $\nabla\ell(\theta^*; Z, Z')$. Assume that V_0 is non-degenerate. If V_0 is positive definite, then for any $v_0 \in \mathbb{R}^p$,

$$\sqrt{N}v_0^\top(\widehat{\theta}_N - \theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, 4v_0^\top\nabla^2\mathcal{L}_0(\theta^*)^{-1}V_0\nabla^2\mathcal{L}_0(\theta^*)^{-1}v_0\right). \tag{22}$$

Similar results hold for the SU-ERM estimator $\widetilde{\theta}_N$ and the OS-ERM estimator $\widetilde{\theta}_N^Q$, meaning that $\widetilde{\theta}_N$ and $\widetilde{\theta}_N^Q$ are asymptotically equivalent to $\widehat{\theta}_N$ when $V_0 > 0$.

Theorem 9 *Let Conditions (C1)-(C6) be satisfied. In particular, there exists $\zeta > 0$ such that $p = O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau > 0$. Assume that the initial estimator $\widehat{\theta}_0$ lies in $U(\theta^*, \tilde{\rho})$, where $\tilde{\rho} = \min\{(1-\rho)\lambda_-\delta_\rho/(32\lambda_+), \sqrt{(1-\rho)\lambda_-\delta_\rho/(32M)}\}$ with $\delta_\rho = \min\{\rho, \rho\lambda_-/(4M)\}$, and*

$\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}\left((p/n)^{1/2}\right)$. *Then the SU-ERM estimator* $\widetilde{\theta}_N$ *satisfies*

$$\widetilde{\theta}_N - \theta^* = -\nabla^2 \mathcal{L}_0\left(\theta^*\right)^{-1} \left\{ \frac{2}{N} \sum_{i=1}^{N} g\left(\theta^*; Z_i\right) + \frac{2K}{N^2} \sum_{k=1}^{K} \sum_{1 \le i < j \le n} h(\theta^*; Z_{k,i}, Z_{k,j}) \right\} + G_{n,K,p}(\widehat{\theta}_0) + \omega_{n,K,p}, \tag{23}$$

where $g(\theta; z) = \mathbb{E}\left[\nabla \ell\left(\theta; Z, Z'\right) | Z = z\right]$, $h(\theta^*; z, z') = \nabla \ell(\theta^*; z, z') - g(\theta^*; z) - g(\theta^*; z')$, $G_{n,K,p}(\widehat{\theta}_0)$ *is a random function of* $\widehat{\theta}_0$ *satisfying* $\|G_{n,K,p}(\widehat{\theta}_0)\|_2 = O_{\mathbb{P}}((p/n)^{1/2}\|\widehat{\theta}_0 - \theta^*\|_2)$, *and* $\|\omega_{n,K,p}\|_2 = o_{\mathbb{P}}(\|\widetilde{\theta}_N - \theta^*\|_2)$. *If we also assume* $pK = o(n)$, *then* $\widetilde{\theta}_N$ *is asymptotically equivalent to* $\widehat{\theta}_N$, *and for any* $v_0 \in \mathbb{R}^p$ $(v_0 \neq 0)$,

$$\frac{\sqrt{N} v_0^\top (\widetilde{\theta}_N - \theta^*)}{\sqrt{4 v_0^\top \nabla^2 \mathcal{L}_0(\theta^*)^{-1} V_0 \nabla^2 \mathcal{L}_0(\theta^*)^{-1} v_0}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

This theorem also holds for the OS-ERM estimator. One can obtain the analogous result for the OS-ERM estimator by replacing $\widetilde{\theta}_N$ *by* $\widetilde{\theta}_N^Q$ *everywhere in the statements of the theorem and replacing* $\widehat{\theta}_0 \in U(\theta^*, \tilde{\rho})$ *by* $\widehat{\theta}_0 \in U(\theta^*, \rho')$, *where* ρ' *is defined in Proposition 3.*

Remark 10 *The asymptotic normality in Theorem 9 is valid for* $V_0 > 0$, *which is referred to as the non-degenerate case. If* V_0 *is degenerate, then* $\widetilde{\theta}_N$ *is still asymptotically normally distributed, a result different from that of the global estimator* $\widehat{\theta}_N$, *whose asymptotic distribution is an infinite series of independent Chi-square random vectors. As well, the asymptotic variance of* $\widetilde{\theta}_N$ *is* $2KN^{-2}\nabla^2 \mathcal{L}_0(\theta^*)^{-1}(V_1 - 2V_0)\nabla^2 \mathcal{L}_0(\theta^*)^{-1}$, *which is* K *times of the asymptotic variance of* $\widehat{\theta}_N$, *where* $V_1 = \mathbb{E}[\nabla \ell(\theta^*; Z, Z')\nabla^\top \ell(\theta^*; Z, Z')]$. *The asymptotic results relating to the case of degenerate* V_0 *are contained in the Online Supplementary Material.*

From Theorem 9, provided that $\|\widehat{\theta}_0 - \theta^*\|_2 = o_{\mathbb{P}}(\sqrt{1/K})$, the SU-ERM estimator and the OS-ERM estimators are as asymptotically efficient as the global U-estimator $\widehat{\theta}_N$ when V_0 is non-degenerate. As mentioned, this condition is satisfied when $pK = o(n)$, and $\widehat{\theta}_n^k$ is chosen to be the initial estimator. We suggest an iterative algorithm to deal with the situation where the number of local machines or the dimension of covariates exceed the size of the sample subset. This is discussed in Section 6.

Our next task is to develop a valid distributed statistical inference procedure. By Theorem 9, the inference procedure necessitates the estimation of $\nabla^2 \mathcal{L}_0(\theta^*)$ and V_0 in a distributed setting. From results of Honoré and Powell (1994), a consistent estimator of $\nabla^2 \mathcal{L}_0(\theta^*)$ is $\nabla^2 \mathcal{L}_N(\widetilde{\theta}_N)$. The following may be used as a consistent estimator of V_0:

$$\widehat{V}_N(\widetilde{\theta}_N) = \frac{1}{N^3} \sum_{i} \sum_{j \neq i} \sum_{l \neq i,j} \nabla \ell(\widetilde{\theta}_N; Z_i, Z_j) \nabla \ell(\widetilde{\theta}_N; Z_i, Z_l)^\top. \tag{24}$$

The problem here is that $\nabla^2 \mathcal{L}_N(\widetilde{\theta}_N)$ and $\widehat{V}_N(\widetilde{\theta}_N)$ cannot be computed in the distributed setting because they cannot be calculated when the data are not all used in one go. On the other hand, $\nabla^2 \mathcal{L}_N(\widetilde{\theta}_N)$ and $\widehat{V}_N(\widetilde{\theta}_N)$ are U-statistics. From Chen and Peng (2021), provided

that $K = o(N)$, the surrogate U-statistic $\nabla^2\widetilde{\mathcal{L}}_N(\widetilde{\theta}_N) = 1/K\sum_{k=1}^K \nabla^2\mathcal{L}_n^k(\widetilde{\theta}_N)$ has the same asymptotic efficiency as $\nabla^2\mathcal{L}_N(\widetilde{\theta}_N)$ and is therefore a consistent estimator of $\nabla^2\mathcal{L}_0(\theta^*)$. By an analogous argument, $\widehat{V}_{N,K}(\widetilde{\theta}_N) = 1/K\sum_{k=1}^K \widehat{V}_{n,k}(\widetilde{\theta}_N)$ is a consistent estimator of V_0, where $\widehat{V}_{n,k}(\widetilde{\theta}_N)$ is computed from the subset of data $\mathcal{D}_k$ on the k^{th} machine. The consistency of variance estimators is shown in the following theorem.

Theorem 11 *Let Conditions (C1)-(C6) be satisfied. In particular, there exists $\zeta > 0$ such that $p = O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau > 0$. Assume that the initial estimator $\widehat{\theta}_0$ lies in $U(\theta^*, \tilde{\rho})$, where $\tilde{\rho} = \min\{(1-\rho)\lambda_-\delta_\rho/(32\lambda_+), \sqrt{(1-\rho)\lambda_-\delta_\rho/(32M)}\}$ with $\delta_\rho = \min\{\rho, \rho\lambda_-/(4M)\}$, and $\widehat{\theta}_0$ also satisfies $\mathbb{E}\|\widehat{\theta}_0 - \theta^*\|_2^{16} = O(\alpha_{n,K,p}^{16})$, where $\alpha_{n,K,p} = \sqrt{1/K}$ or $\sqrt{p/n}$. Then the MSE of the estimators $\nabla^2\widetilde{\mathcal{L}}_N(\widetilde{\theta}_N)$ and $\widehat{V}_{N,K}(\widetilde{\theta}_N)$ are bounded above by*

$$\mathbb{E}[\|\nabla^2\widetilde{\mathcal{L}}_N(\widetilde{\theta}_N) - \nabla^2\mathcal{L}_0(\theta^*)\|_2^2] = O\left(\frac{\alpha_{n,K,p}^4}{K} + \frac{p}{nK^2} + \frac{(\log p)^4}{n^4\sqrt{K}} + \frac{\log p}{N}\right) \quad \text{and}$$

$$\mathbb{E}[\|\widehat{V}_{N,K}(\widetilde{\theta}_N) - V_0\|_2^2] = O\left(\frac{\alpha_{n,K,p}^4}{K} + \frac{\alpha_{n,K,p}^8}{K} + \frac{p}{nK^2} + \frac{(\log p)^4}{n^4\sqrt{K}} + \frac{(\log p)^8}{n^8}\right)$$

respectively. In particular, the statistics $\nabla^2\widetilde{\mathcal{L}}_N(\widetilde{\theta}_N)$ and $\widehat{V}_{N,K}(\widetilde{\theta}_N)$ are consistent estimators of $\nabla^2\mathcal{L}_0(\theta^*)$ and V_0 respectively.

It can be shown using Theorem 11 that the conditions on (n, K, p) required for the consistency of the variance estimators may be weaker than the analogous conditions required for the parameter estimators presented in Theorem 5. This will be the case, for example, when the upper bound are expressed in terms of the spectral norm $\|\cdot\|_2$. However, the conditions are different when the Frobenius norm $\|\cdot\|_F$ is used, since $\|\Omega\|_F \le \sqrt{p}\|\Omega\|_2$ for any $p\times p$ symmetric matrix Ω. It can be shown that if we require $\mathbb{E}[\|\nabla^2\widetilde{\mathcal{L}}_N(\widetilde{\theta}_N) - \nabla^2\mathcal{L}_0(\theta^*)\|_F^2] = o(1)$ and $\mathbb{E}[\|\widehat{V}_{N,K}(\widetilde{\theta}_N) - V_0\|_F^2] = o(1)$, the restrictions of (n, K, p) are in the form of

$$p = o(\min\{n^{2/3}K^{1/3}, n^{1/2}K, n^{4/5}K^{1/5}\}) \text{ and } \log p = o(n\min\{p^{-1/8}, K^{1/8}p^{-1/4}, Kp^{-1}\}). \tag{25}$$

To further compare Condition (25) with the analogous condition in Theorem 5, let us consider the special case where $K = O(n^\gamma)$ and $p = O(n^\eta)$ for some $A, \eta > 0$. Then Condition (25) holds if and only if

$$3\eta - \gamma < 2, \quad \eta - \gamma < 1/2, \quad 5\eta - \gamma < 4 \quad \text{and} \quad \eta < 8.$$

Figure 1 shows the feasible region for the proposed coefficient estimators and the variance estimator to be consistent, and the region for the proposed estimators to achieve the optimal rate of convergence. From Figure 1, both the coefficient and variance estimators are consistent in the region $R_4 \cup R_5$ that corresponds to $3\eta - \gamma < 2$, $\eta - \gamma < 1/2$, $\eta < 1$ and $\gamma < 8$. In particular, if $\eta - \gamma < 1/2$ and $\eta + \gamma < 1$ (say $(\gamma, \eta) \in R_5$), the SU-ERM and OS-ERM estimators achieve the optimal rate and the variance estimators are consistent. We label R_5 as the ideal region.

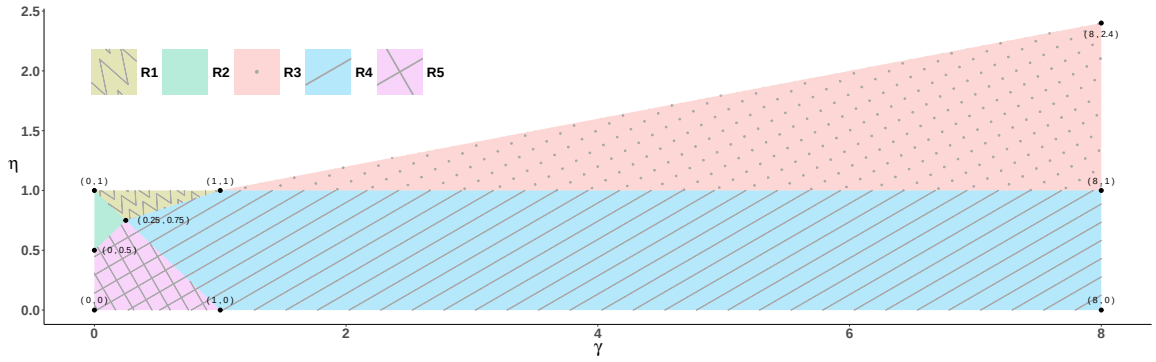


Figure 1: The consistency regions (CRs), that is, the feasible regions for the SU-ERM/OS-ERM estimator and the variance estimator to be consistent as a function of γ and η. To conserve space, we only report the CRs for $\gamma \leq 8$. The CR of the SU-ERM/OS-ERM estimator is $R_1 \cup R_2 \cup R_4 \cup R_5$, while the CR of the variance estimator is $R_3 \cup R_4 \cup R_5$. In the region $R_2 \cup R_5$, the SU-ERM/OS-ERM estimator achieves the optimal convergence rate.

Hence there is a justification to use the estimated covariance matrix of $\widetilde{\theta}_N$ as a plug-in estimator, that is, $\widehat{\mathrm{Var}}(\widetilde{\theta}_N) = \nabla^2 \widetilde{\mathcal{L}}_N(\widetilde{\theta}_N)^{-1} \widehat{V}_{N,K}(\widetilde{\theta}_N) \nabla^2 \widetilde{\mathcal{L}}_N(\widetilde{\theta}_N)^{-1}$. We can construct the $100(1-\alpha)\%$ confidence interval of $v_0^\top \theta^*$ as follows:

$$\left[v_0^\top \widetilde{\theta}_N - N^{-1/2} \sqrt{v_0^\top \widehat{\mathrm{Var}}(\widetilde{\theta}_N) v_0} z_{1-\alpha/2}, v_0^\top \widetilde{\theta}_N + N^{-1/2} \sqrt{v_0^\top \widehat{\mathrm{Var}}(\widetilde{\theta}_N) v_0} z_{1-\alpha/2} \right], \tag{26}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th percentile of the standard Normal distribution.

6. Distributed Iterative Algorithm

Given the limited capacity of any single machine, distributed data processing will likely require many machines when there is a massive number of observations. This can result in the number of machines K exceeding the sample size n of the data subset. In such a circumstance, the SU-ERM and OS-ERM approaches may not result in the same asymptotic efficiency as the global U-estimator. We reconcile this difficulty by implementing the method within a distributed iterative algorithm that has the effect of progressively improving the estimator's performance. For each of the SU-ERM and OS-ERM methods, we introduce one such algorithm, and label them as DiaSER and DiaOSE respectively. The steps of the algorithms are presented in Algorithms 1 and 2 below.

6.1 Properties of Estimators Following the Algorithm

The following comments are based on the DiaOSE algorithm. The same comments apply to the DiaSER algorithm. Assume that at the t-th iteration ($t \geq 1$), the estimate $\widetilde{\theta}^Q_{N,t}$ is obtained, where $\widetilde{\theta}^Q_{N,1} = \widetilde{\theta}^Q_N$. Then at the $(t+1)$-th iteration, DiaOSE treats $\widetilde{\theta}^Q_{N,t}$ as the (new) initial estimate, and refines it by the one-step updating mechanism $\widetilde{\theta}^Q_{N,t+1} = \widetilde{\theta}^Q_{N,t} - \nabla^2 \mathcal{L}^t_n(\widetilde{\theta}^Q_{N,t})^{-1} \nabla \widetilde{\mathcal{L}}_N(\widetilde{\theta}^Q_{N,t})$ as in (16). If the initial estimator $\widehat{\theta}_0$ satisfies $\mathbb{E}\|\widehat{\theta}_0 - \theta^*\|^2 = O(p/n)$ and $p = o(n)$, Proposition 7 implies that the OS-ERM estimate $\widetilde{\theta}^Q_{N,t}$, obtained from the t-th

Initialization: Compute $\widehat{\theta}_n^k$ at Machine k, where $\widehat{\theta}_n^k = \arg\min_{\theta\in\Theta} \mathcal{L}_n^k(\theta)$, $k = 1, 2, \cdots, K$. Transfer the local estimates to the master machine, on which the naive estimator $\overline{\theta}_N = 1/K \sum_{k=1}^K \widehat{\theta}_n^k$ is calculated. Set $\widehat{\theta}_0 = \overline{\theta}_N$;
for *Iteration* $t = 1, \ldots, T$ **do**
  The master machine transfers $\widehat{\theta}_0$ to each local machine;
  for *Machine* $k = 1, \ldots, K$ **do**
    Machine k computes the local gradient $\nabla\mathcal{L}_n^k(\widehat{\theta}_0)$ based on data subset $\mathcal{D}_k$;
    Machine k sends $\nabla\mathcal{L}_n^k(\widehat{\theta}_0)$ back to the master machine ;
  end
  The master machine computes the surrogate global gradient $\nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0) = 1/K \sum_{k=1}^K \nabla\mathcal{L}_n^k(\widehat{\theta}_0)$;
  The master machine conducts the empirical risk minimization based on the surrogate empirical risk $\mathcal{S}_N^t(\theta) = \mathcal{L}_n^t(\theta) - (\nabla\mathcal{L}_n^t(\widehat{\theta}_0) - \nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0))^\top\theta$, and obtains the t-th SU-ERM estimator $\widetilde{\theta}_{N,t} = \arg\min_{\theta\in\Theta} \mathcal{S}_N^t(\theta)$;
  Set $\widehat{\theta}_0 = \widetilde{\theta}_{N,t}$;
end
return $\widetilde{\theta}_{N,T}$;

Algorithm 1: Distributed iterative algorithm based on surrogate empirical risk (DiaSER)

Initialization: Compute $\widehat{\theta}_n^k$ at Machine k, where $\widehat{\theta}_n^k = \arg\min_{\theta\in\Theta} \mathcal{L}_n^k(\theta)$, $k = 1, 2, \cdots, K$. Transfer the local estimates to the master machine, on which the naive estimator $\overline{\theta}_N = 1/K \sum_{k=1}^K \widehat{\theta}_n^k$ is calculated. Set $\widehat{\theta}_0 = \overline{\theta}_N$;
for *Iteration* $t = 1, \ldots, T$ **do**
  The master machine transfers $\widehat{\theta}_0$ to each local machine;
  for *Machine* $k = 1, \ldots, K$ **do**
    Machine k computes the local gradient $\nabla\mathcal{L}_n^k(\widehat{\theta}_0)$ with subset $\mathcal{D}_k$;
    Machine k sends $\nabla\mathcal{L}_n^k(\widehat{\theta}_0)$ back to the master machine ;
  end
  The master machine computes the surrogate global gradient $\nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0) = 1/K \sum_{k=1}^K \nabla\mathcal{L}_n^k(\widehat{\theta}_0)$;
  The master machine computes the local Hessian matrix $\nabla^2\mathcal{L}_n^t(\widehat{\theta}_0)$ with subset $\mathcal{D}_t$, and obtains the t-th OS-ERM estimator $\widetilde{\theta}_{N,t}^Q = \widehat{\theta}_0 - \nabla^2\mathcal{L}_n^t(\widehat{\theta}_0)^{-1}\nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0)$;
  Set $\widehat{\theta}_0 = \widetilde{\theta}_{N,t}^Q$;
end
return $\widetilde{\theta}_{N,T}^Q$;

Algorithm 2: Distributed iterative algorithm based on one-step estimation (DiaOSE)

iteration satisfies the recursive formula

$$\|\widetilde{\theta}^{Q}_{N,t+1} - \widehat{\theta}_N\|_2 = O_{\mathbb{P}}((p/n)^{1/2})\|\widetilde{\theta}^{Q}_{N,t} - \widehat{\theta}_N\|_2 + o_{\mathbb{P}}((p/N)^{1/2}).$$

Hence, after performing multiple rounds of iterations, we obtain the Bahadur representations of $\widehat{\theta}^{Q}_{N,t}$ and $\widetilde{\theta}_{N,t}$ in the following theorem:

Theorem 12 *Assume that the conditions of Proposition 7 holds and the initial estimator $\widehat{\theta}_0$ satisfies $\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}(\alpha_{n,K,p})$ with $\alpha_{n,K,p} = o(1)$. Then the OS-ERM estimate $\widetilde{\theta}^{Q}_{N,t}$ obtained from the t-th iteration satisfies*

$$\|\widetilde{\theta}^{Q}_{N,t} - \widehat{\theta}_N\|_2 = O_{\mathbb{P}}\left(\left(\alpha_{n,K,p} + \sqrt{\frac{p}{N}} + \sqrt{\frac{\log p}{n}}\right)^t\right)\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 + O_{\mathbb{P}}\left(\sqrt{\frac{p}{N}}\max\left\{\alpha_{n,K,p}, \sqrt{\frac{\log p}{N}}\right\}\right).$$

Also, for any $v_0 \in \mathbb{R}^p$,

$$v_0^\top \widetilde{\theta}^{Q}_{N,t} - v_0^\top \theta^* = -v_0^\top \nabla^2 \mathcal{L}_0(\theta^*)^{-1}\left\{\frac{2}{N}\sum_{i=1}^{N} g(\theta^*; Z_i) + \frac{2K}{N^2}\sum_{k=1}^{K}\sum_{1\le i<j\le n} h(\theta^*; Z_{k,i}, Z_{k,j})\right\} + r_{n,K,p},$$

where

$$\begin{aligned} r_{n,K,p} = O_{\mathbb{P}}&\left(\left(\alpha_{n,K,p} + \sqrt{\frac{p}{N}}\right)^{t+1} + \left(\sqrt{\frac{\log p}{n}}\right)^t\left(\alpha_{n,K,p} + \sqrt{\frac{p}{N}}\right)\right) \\ &+ O_{\mathbb{P}}\left(\sqrt{\frac{p}{N}}\max\left\{\alpha_{n,K,p}, \sqrt{\frac{\log p}{N}}\right\}\right). \end{aligned}$$

This theorem also holds for the SU-ERM estimator, with $\widetilde{\theta}_N$ replacing $\widetilde{\theta}^{Q}_N$ everywhere in the statement of the theorem.

Remark 13 *If $\alpha_{n,K,p} = \sqrt{p/n}$, Theorem 12 implies*

$$\|\widetilde{\theta}^{Q}_{N,t} - \widehat{\theta}_N\|_2 = O_{\mathbb{P}}((p/n)^{t/2})\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 + o_{\mathbb{P}}((p/N)^{1/2}). \tag{27}$$

Note that no more than $\lceil \log K/\log(n/p)\rceil$ iterations are required to achieve the same accuracy as $\widehat{\theta}_N$ (recall that $\|\widehat{\theta}_N - \theta^*\|_2 = O_{\mathbb{P}}(\sqrt{p/N})$). Hence, by implementing the DiaOSE and DiaSER algorithms, our proposed methods no longer require the condition of $pK = o(n)$ to achieve the optimal convergence rates. This leads to weaker conditions on K and p of the proposed estimators compared with those of the naive estimator $\bar{\theta}_N$. In particular, when $pK = o(n)$, we require only one iteration to achieve the desired accuracy.

When implementing the algorithms, we use the naive estimator $\overline{\theta}_N$ obtained by simple averaging as the initial estimator. Although the naive estimator $\overline{\theta}_N$ and the local U-estimator $\widehat{\theta}^k_n$ share the same order of magnitude of bias, the naive estimator has a smaller variance than $\widehat{\theta}^k_n$. The lower variability can enhance the numerical stability of the algorithms. There is only a minimal extra communication cost of $O((K-1)p)$ of using the naive estimator at the start.

6.2 Time Complexity

Here, we derive bounds on the running times of the DiaSER and DiaOSE algorithms and the optimization methods used. The proposed SU-ERM and OS-ERM estimators are special cases for DiaSER and DiaOSE algorithms respectively with T set to 1. Instead of focusing on a specific case, we conduct a general investigation with respect to the proposed SU-ERM and OS-ERM estimators, the naive estimator $\bar{\theta}_N = K^{-1}\sum_{k=1}^{K}\widehat{\theta}_n^k$ and the global empirical risk minimization (GL-ERM) estimator $\widehat{\theta}_N = \arg\min \mathcal{L}_N(\theta)$ under a general loss function. The computational complexity of the GL-ERM estimator is $O(pn^2K^2)$, which is significantly higher than that of the proposed distributed iterative algorithms. We present the details below.

Let us first consider the DiaSER algorithm. Assume that the optimization of (12) is implemented via some algorithms, resulting in an approximated minimizer $\widetilde{\theta}_N^*$. Then the error of $\widetilde{\theta}_N^*$ is upper-bounded by the sum of the estimation error of $\widetilde{\theta}_N$ and the approximation error of the optimization algorithm, that is,

$$\|\widetilde{\theta}_N^* - \theta^*\|_2 \leq \|\widetilde{\theta}_N - \theta^*\|_2 + \|\widetilde{\theta}_N^* - \widetilde{\theta}_N\|_2. \tag{28}$$

Analogously, let $\widehat{\theta}_n^1$ be the initial estimator and $\widehat{\theta}_n^{1*}$ the corresponding approximated minimizer obtained by some optimization algorithm. The error bound for $\widehat{\theta}_n^{1*}$ is also represented by (28).

We assume that all estimates, including the initial estimate obtained by $\widetilde{\theta}_n^1$ and the final estimate based on $\widetilde{\theta}_N$ or $\widetilde{\theta}_N^Q$ are calculated by the same optimization method. Let the approximation error be $\epsilon > 0$. One may adopt a gradient method or the Newton-Raphson method as the optimization method. These methods usually provide satisfactory results. Two popular gradient methods are the gradient descent (GD) and the stochastic gradient descent (SGD) methods. If one uses the SGD to obtain the initial estimator, then $O(\log^2(n/p)/\rho^2)$ iterations are required to guarantee $\widetilde{\theta}_n^{1*} \in U(\theta^*, \rho)$ for some $\rho > 0$. This results in a total computational cost of $O(pn^2\log^2(n/p))$. If GD is used to optimize (12), we need $O(\log(nK))$ iterations to reduce the approximation error to $\|\widetilde{\theta}_N^* - \widetilde{\theta}_N\|_2 = O_{\mathbb{P}}((p/N)^{1/2})$ with the cost per iteration in the order of $O(pn^2)$. Hence the total computational cost associated with the GD approach is $O(pn^2T\log(nK))$. If SGD is used, the cost per iteration is reduced to $O(p)$ but the number of iterations increases to $O(nK)$. Hence the SGD approach results in a computational cost of $O(pnKT)$. As well, the Broyden-Fletcher=Goldfarb-Shanno (BFGS) method (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) is a commonly used Newton-Raphson-type method. The BFGS methods incurs a computational cost of $O(Tp^2\log\log(1/\epsilon))$ when being implemented to optimize (12).

As in the case of the DiaSER algorithm, the DiaOSE algorithm is implemented by computing (16) in multiple rounds. If the SGD approach is used to obtain the initial estimator, then the total computational cost amounts to $O(pn^2\log^2(n/p))$. In practice, instead of directly deriving the inverse matrix $\nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)^{-1}$, we usually compute the LU decomposition of the matrix $\nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)$, and solve the linear system

$$\nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)(\widetilde{\theta}_N^Q - \widehat{\theta}_0) = -\nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0). \tag{29}$$

The computational costs associated with the LU decomposition and the steps for solving the linear system are $O(p^3)$ and $O(p^2)$ respectively. To further reduce the computing time,

we update $\nabla^2\mathcal{L}_n^1(\widehat{\theta}_{N,t}^Q)$ every p steps. These result in a total computing time of $O(n^2T + p^2T + p^3T/p + pn^2\log^2(n/p)) = O((p^2+n^2)T + pn^2\log^2(n/p))$. We summarize the running time bounds of different methods in Table 1.

Distributed Algorithm	Optimization Algorithm	Computational Cost
DiaOSE	Direct	$O((p^2+n^2)T + \mathcal{T}(n,K,p))$
DiaSER	GD	$O(pn^2K^2T\log(1/\epsilon))$
DiaSER	SGD	$O(Tp/\epsilon)$
DiaSER	BGFS	$O(Tp^2\log\log(1/\epsilon))$
Naive	GD	$O(pn^2\log(1/\epsilon) + pK)$
Naive	SGD	$O(p/\epsilon + pK)$
Naive	BGFS	$O(p^2\log\log(1/\epsilon) + pK)$

Table 1: Computational cost of the DiaOSE and DiaSER algorithms. We set the approximation accuracy to some known $\epsilon > 0$. The DiaOSE algorithm has a closed-form expression and usually requires no optimization (except for $\widetilde{\theta}_n^{1*}$). Hence in the table, we refer to the optimization algorithm associated with the DiaOSE algorithm as "Direct". With the DiaSER algorithm, the cost of estimating $\widehat{\theta}_0$ is given by $\mathcal{T}(n,K,p) = np\log(1/\epsilon), p/\epsilon$ and $p^2\log\log(1/\epsilon)$ under the GD, SGD and the BGFS methods respectively.

From Table 1, other things being equal, the SGD and the BGFS methods result in the smallest time bounds for the large and small p cases respectively. We implement the BFGS method via the R package *optim* with $J = 1000$ iterations. Under the BGFS method, the costs of computing the DiaOSE algorithm, DiaSER algorithm and the naive estimator are $O((p^2+n^2)T + p^2J)$, $O(p^2TJ)$ and $O(p^2J + pK)$ respectively.

6.3 Guidance on Practice

In practice, a proper choice of the parameter K is required to balance statistical efficiency and computational complexity. When N is fixed, the number of machines K must not be very large. This is because when K is exceedingly large, the sample size of each machine will be small, resulting in poor approximations of the high-order gradients in the surrogate empirical risk $\mathcal{S}_N(\theta)$ in (10). As well, from the asymptotic expansion of $\widetilde{\theta}_N$, the parameter K only has an effect on the order of the remainder term. To avoid the cumbersome task of computing the inverse matrix, we consider the equivalent expansion

$$[\nabla^2\mathcal{L}_0(\theta^*)](\widetilde{\theta}_N - \theta^*) = -\frac{2}{N}\sum_{i=1}^N g(\theta^*; Z_i) + e_{n,K,p},$$

where $\|e_{n,K,p}\|_2 = o_{\mathbb{P}}(\|N^{-1}\sum_{i=1}^N g(\theta^*; Z_i)\|_2)$. A good candidate of K is to minimize the MSE of the remainder term.

Let the initial estimator be $\widehat{\theta}_0 = \widetilde{\theta}_n^1$. For a given K, let $\widetilde{\theta}_{N,K}^{-k}$ be the SU-ERM estimator obtained by excluding the data stored on the k-th machine ($k \geq 2$). Specifically, $\widetilde{\theta}_{N,K}^{-k} = \arg\min_{\theta\in\Theta}\mathcal{S}_{N,K}^{-k}(\theta)$, where $\mathcal{S}_{N,K}^{-k}(\theta)$ is constructed by replacing $\nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0)$ by $\nabla\widetilde{\mathcal{L}}_{N,K}^{-k}(\widehat{\theta}_0) = \frac{1}{K}\sum_{k'\neq k}\mathcal{L}_n^{k'}(\widehat{\theta}_0)$ in (10). Denote $\mathbb{G}_N(\theta^*) = \sum_{i=1}^N g(\theta^*; Z_i)/N$ and $\mathbb{V}(\theta^*) = \nabla^2\mathcal{L}(\theta^*)$. The

matrix $\mathbb{V}(\theta^*)$ can be estimated by $\mathbb{V}_n = \frac{1}{n^2}\sum_{i\neq j}\nabla^2\ell(\widehat{\theta}_0; Z_{1,i}, Z_{1,j})$ using the data stored on the first machine. For a given $k \geq 2$, let

$$\mathbb{G}_{N,K}^{-k} = \frac{1}{n(N-n)}\sum_{k'\neq k}\sum_{i\neq j}\nabla\ell(\widetilde{\theta}_{N,K}^{-k}; Z_{k',i}, Z_{k',j})$$

as an approximation of $\mathbb{G}_N(\theta^*)$. Write the error as

$$\epsilon_{N,K}^{-k} = \mathbb{V}_n(\widetilde{\theta}_{N,K}^{-k} - \bar{\theta}) + 2\mathbb{G}_{N,K}^{-k},$$

where $\bar{\theta} = \sum_{k=2}^{K}\widetilde{\theta}_{N,K}^{-k}/(K-1)$. The parameter K is chosen by minimising the MSE of $\{\epsilon_{N,K}^{-k}\}_{k=1}^{K}$ such that $\widehat{K} = \arg\min_K K^{-1}\sum_{k=1}^{K}(\epsilon_{N,K}^{-k}-\bar{\epsilon})(\epsilon_{N,K}^{-k}-\bar{\epsilon})'$, where $\bar{\epsilon} = \sum_{k=1}^{K}\epsilon_{N,K}^{-k}/K$. As the OS-ERM and SU-ERM estimators share the same asymptotic properties, the parameter K for the OS-ERM estimator can be chosen analogously.

It is instructive to note that in practice, when one applies the DiaOSE and DiaSER algorithms under a pre-determined K, the number of iterations essentially becomes the tuning parameter. According to Theorem 12 and Remark 13, it suffices to iterate $\lceil \log K/\log(n/p)\rceil$ times when $\widehat{\theta}_0 = \widetilde{\theta}_n^1$. Also, our simulation results show that the proposed algorithms are not sensitive to the iteration number T, and in most situations, reasonably accurate results are obtained after a relatively small number of iterations.

7. Simulation Studies

The purpose of this section is to examine the finite sample performance of the proposed methods via a large scale simulation study. We focus on the U-ERM problems of pairwise ranking and smoothed rank-based estimation under the accelerated failure time model, both being introduced in Section 3.2. For comparison purposes, we also consider the gold-standard method that uses the full set of data all in one go to obtain the global U-estimates, and the naive method that averages local U-estimates obtained across different subsets of data.

7.1 Pairwise Ranking Problem

Our first experiment considers the ranking problem with the logistic loss function, the details of which are presented in Example 1.

7.1.1 Simulation Setting

Note that statistical inference for ranking problems requires only the ordering information but not the exact values of Y_i's. Thus, in our experiment, for any two objects Z_i and Z_j, based on their corresponding features X_i and X_j, we simulate the pairwise ordering from the following model:

$$Y_i = J(X_i^\top\theta^* + \varepsilon_i), \quad i = 1, \ldots, n, \tag{30}$$

where θ^* is the true value of the parameter satisfying $\|\theta^*\|_2 = 1$, X_i is the predictor and ε_i is the noise variable. The function $J(\cdot)$ is defined as

$$J(t) = \sum_{k=1}^{5} kI(t_{k-1} \leq t < t_k),$$

where $t_0 = -\infty$, $t_5 = \infty$, and $t_1, \ldots, t_4$ are fixed such that there are approximately equal number of elements in each class. The estimator of θ^* is the solution to the empirical risk minimization problem based on the following logistic loss function:

$$\mathcal{L}_N(\theta) = \frac{1}{\mathcal{C}_N^2} \sum_{i<j} \ln[1 + \exp\{-\text{sign}(Y_i - Y_j)\theta^\top (X_i - X_j)\}],$$

subject to the constraint $\|\theta\|_2 = 1$.

We set $\theta^* = (1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5})^\top$ and generate X_i from a multivariate $\mathcal{N}(0, \Sigma)$ distribution, where $\Sigma_{ij} = 1$ when $i = j$ and $\Sigma_{ij} = 0.5$ when $i \neq j$, and $\varepsilon \sim \mathcal{N}(0, 1)$. We generate the ordering between Z_i and Z_j from (30). To examine the performance of the proposed methods for the cases of $K < n$ and $K > n$, we fix the sample size of each subset to be $n = 50$ and vary K, the number of machines, from 25 to 200. We set the number of replications of the experiment to $S = 200$. It is worth noting that the constraint $\|\theta\|_2 = 1$ places additional demands on the minimization problem, but it is needed for model identifiability purpose. Indeed, when $K = 200$, to obtain the global U-estimate, one has to simultaneously process more than 49995000 sample pairs on a single machine, and solve the corresponding constrained optimization problem. This is virtually impossible to compute. Hence, we only present the simulation results for the proposed estimators and the naive estimator.

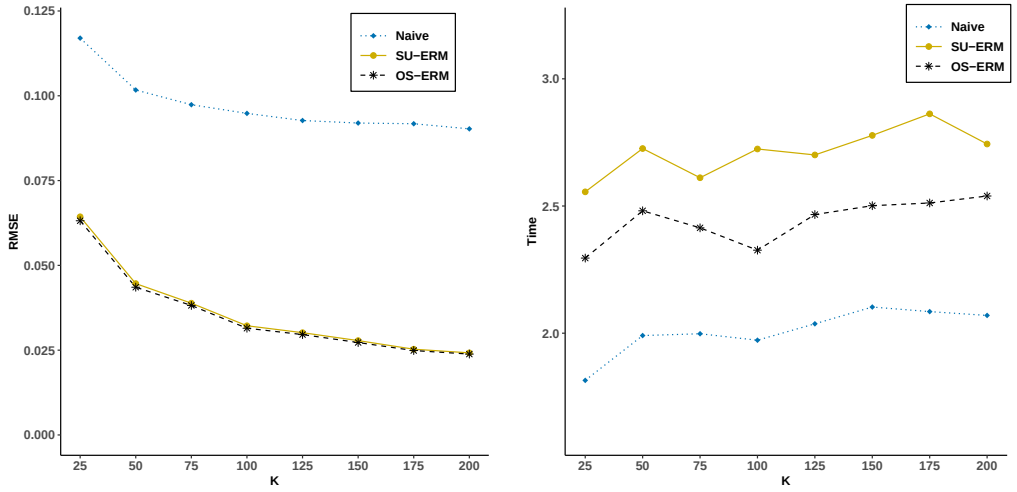
To gauge the performance of the methods, we use the following empirical root mean square error (RMSE) of an estimator $\widehat{\theta}$ of θ^*:

$$\frac{1}{S} \sum_{s=1}^{S} \sqrt{\sum_{l=1}^{p} (\widehat{\theta}_l^s - \theta^*)^2}, \tag{31}$$

where $\widehat{\theta}_l^s$ is the estimates of the l-th component of θ based on the s-th simulation. As mentioned, the execution of the DiaSER and DiaOSE algorithms requires only a very small number of iterations. Under the current simulation setup, it is found that both algorithms converge after no more than five iterations. The results to be presented are based on the SU-ERM and OS-ERM estimates obtained at the end of the fifth iteration. We label them as $\widetilde{\theta}_{N,5}$ and $\widetilde{\theta}^Q_{N,5}$ respectively.

7.1.2 Overall Results

Figure 2(a) provides the plot of the RMSE of the methods in terms of K, the number of machines. Note that an increase in K is equivalent to an increase in N as $N = K \times n$ and n is fixed. The figure shows that in all cases and by both yardsticks, the SU-ERM and OS-ERM methods outperform the naive method. This is because although simple averaging yields smaller variability than the local U-estimates, it also results in large biases that cannot be compensated for by the lower variance. It can be seen that when K increases, the RMSE of the naive estimator $\overline{\theta}_N$ decreases at a relatively slow rate. In contrast, the RMSEs of the SU-ERM and OS-ERM estimators decrease markedly when K increases. In general, the SU-ERM and OS-ERM methods exhibit very similar performance, a feature that corroborates our theoretical results.//



(a) (b)

Figure 2: (a) RMSEs and (b) computational times of different methods for the pairwise ranking problem.

Table 2 presents the coverage probabilities (CPs) of the 95% confidence intervals based on different methods computed from 200 trials. The construction of the naive method-based CPs uses the same formula (26), with $\widetilde{\theta}_N$ replaced by $\overline{\theta}_N$. The results show that the SU-ERM- and OS-ERM-based CPs are close to the nominal 95% level. This suggests that the proposed variance estimation method is consistent. On the other hand, the CPs based on the naive method are unsatisfactory, especially when K is large. This can be explained by the fact that the naive method results in much larger RMSEs than the SU-ERM and OS-ERM methods, and the gap in RMSEs produced by the naive and the two proposed methods widens as K increases, as shown in Figure 2(a). This indicates that the variance estimation method that works well for the SU-ERM and OS-ERM methods likely underestimates the variance of the naive estimator. Because the naive method requires $pK = o(n)$, the naive estimator is not consistent when K is large. The poor empirical showing of the naive estimator with respect to RMSE and CP corroborates the derived theoretical results on its properties.

K	Naive					SU-ERM					OS-ERM				
	θ_1^*	θ_2^*	θ_3^*	θ_4^*	θ_5^*	θ_1^*	θ_2^*	θ_3^*	θ_4^*	θ_5^*	θ_1^*	θ_2^*	θ_3^*	θ_4^*	θ_5^*
25	0.755	0.660	0.780	0.805	0.810	0.980	0.955	0.950	0.960	0.965	0.970	0.955	0.950	0.955	0.965
50	0.670	0.595	0.565	0.610	0.625	0.975	0.955	0.965	0.960	0.985	0.975	0.960	0.970	0.970	0.990
75	0.490	0.450	0.475	0.415	0.450	0.935	0.965	0.935	0.955	0.945	0.935	0.970	0.945	0.960	0.955
100	0.370	0.330	0.325	0.430	0.380	0.965	0.955	0.965	0.970	0.955	0.970	0.955	0.970	0.965	0.950
125	0.260	0.250	0.295	0.280	0.300	0.975	0.965	0.955	0.925	0.965	0.970	0.980	0.950	0.940	0.965
150	0.155	0.215	0.235	0.220	0.185	0.955	0.970	0.940	0.965	0.940	0.955	0.965	0.940	0.975	0.955
175	0.165	0.195	0.170	0.135	0.115	0.955	0.945	0.975	0.970	0.950	0.955	0.955	0.975	0.970	0.955
200	0.115	0.125	0.125	0.100	0.065	0.970	0.940	0.945	0.935	0.945	0.965	0.940	0.945	0.940	0.940

Table 2: The coverage probabilities (CPs) of CIs corresponding to the nominal 95% of various methods under the pairwise ranking model.

7.1.3 Time Complexity Analysis

Figure 2(b) reports the computational time of the four methods corresponding to the experimental setup in Subsection 7.1.1.

The computational time of a given method is obtained by first averaging the times required for the method to estimate the parameters across all machines and then adding this average and the time required for the optimization. Further details on the calculation of the computational times are given in the Online Supplementary Material. We are primarily interested in the running times of the methods as K varies (that is, with the increase of the total sample size N). The running times reported for the SU-ERM and OS-ERM estimates correspond to the times recorded at the end of the fifth iteration of the DiaSER or DiaOSE algorithm.

The running times of the proposed SU-ERM and OS-ERM methods are not substantially more than the time required for computing the naive estimate used in the initialization stage of the DiaSER and DiaOSE algorithms. This indicates that it does not take long to complete the steps of the algorithms subsequent to initiation. As well, the computational time of the two algorithms remains more or less static irrespective of the value of K. We consider $K = 100$ to be an appropriate choice of K. Generally speaking, the SU-ERM method is a slightly more time consuming procedure than the OS-ERM method because SU-ERM entails solving an optimization problem at each conquer step, whereas the OS-ERM possesses a closed-form solution. It is worth noting that while the OS-ERM estimate takes marginally longer time to compute than the naive estimate, its performance is way better than the latter.

7.1.4 Impact of the Iterative Algorithm

This subsection is devoted to an investigation of the impact of the iterative algorithm on the estimator's properties. We use DiaSER as an example and report the changes in the estimator's RMSE. It is seen that in all cases, the errors of the SU-ERM method decrease rapidly within the initial two iterations and remain steady thereafter, meaning that the algorithm converges very quickly after a small number of iterations. This finding is consistent with our theoretical results.

7.2 Accelerated Failure Time Model

The following experiment, performed in accordance with the smoothed rank-based estimation of the AFT model introduced in Example 2, allows us to examine the efficiency of the procedures with respect to inference.

7.2.1 Simulation Setting

We generate X_i from $\mathcal{N}(0, \Sigma)$ and the random errors ζ_i from the standard extreme value distribution with $\Sigma_{ij} = 0.5^{|i-j|}$. We let $\theta^* = (1, 1, 1)$, and generate the failure time T_i based on the AFT model in (4), with the censoring time C_i generated from the $U[0, \tau]$ distribution, and τ selected to achieve a 25% censoring percentage.

As in the last experiment, we let the sample size of each subset be $n = 50$ and vary the number of subsets K in $\{25, 50, 75, 100, 125, 150, 175, 200\}$. Each experiment is based on

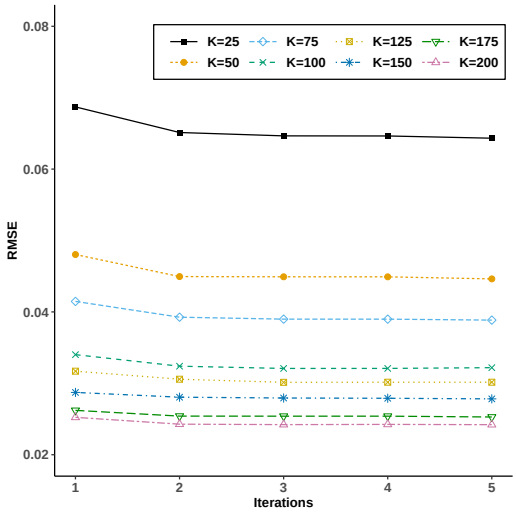
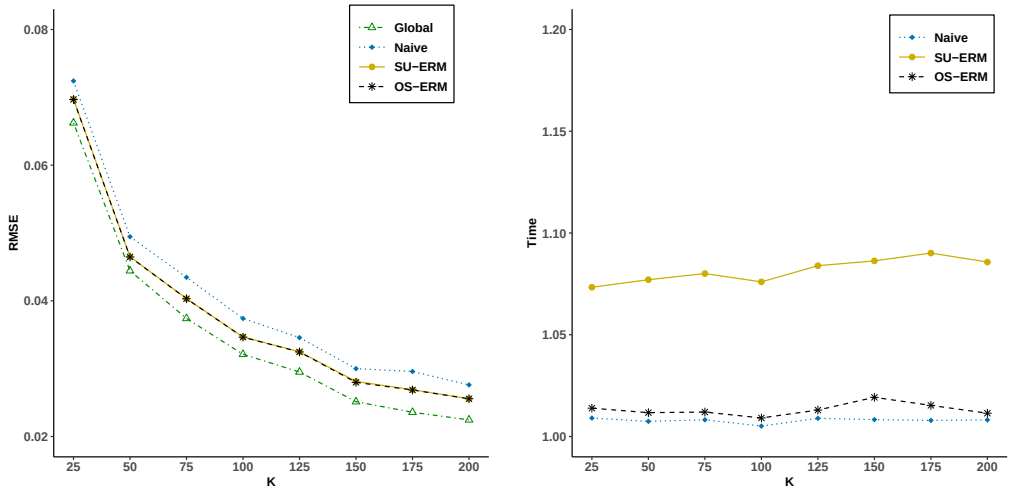


Figure 3: RMSE after varying numbers of iterations for different K's under the DiaSER algorithm.

$S = 200$ replications. We measure the performance of the procedures by the CP associated with a 95% confidence interval for each coordinate of θ^*. Table 3 reports the CP, and Figure 4 shows the RMSE of the four methods. The results reported for the SU-ERM and OS-ERM estimators are those delivered after five iterations.



(a) (b)

Figure 4: (a) RMSEs and (b) computational times of different methods for the AFT model when $n = 50$.

7.2.2 Overall Results

Figure 4 and Table 3 show that in all cases, by the yardsticks of CP and RMSE, the SU-ERM and OS-ERM methods yield estimates comparable to each other's and to those

(K, N)	Naive			SU-ERM			OS-ERM		
	θ_1^*	θ_2^*	θ_3^*	θ_1^*	θ_2^*	θ_3^*	θ_1^*	θ_2^*	θ_3^*
(25, 1250)	0.900	0.975	0.925	0.950	0.945	0.930	0.950	0.945	0.930
(50, 2500)	0.925	0.950	0.925	0.955	0.955	0.965	0.955	0.955	0.965
(75, 3750)	0.910	0.945	0.920	0.940	0.955	0.940	0.940	0.955	0.940
(100, 5000)	0.905	0.925	0.920	0.920	0.960	0.925	0.920	0.960	0.925
(125, 6250)	0.915	0.925	0.910	0.935	0.940	0.925	0.935	0.940	0.925
(150, 7500)	0.920	0.935	0.905	0.930	0.955	0.925	0.930	0.960	0.930
(175, 8750)	0.915	0.890	0.925	0.935	0.905	0.965	0.935	0.905	0.965
(200, 10000)	0.950	0.870	0.915	0.940	0.915	0.950	0.940	0.915	0.950

Table 3: The coverage probabilities (CPs) of CIs corresponding to the nominal 95% of various methods when $n = 50$ under the AFT model.

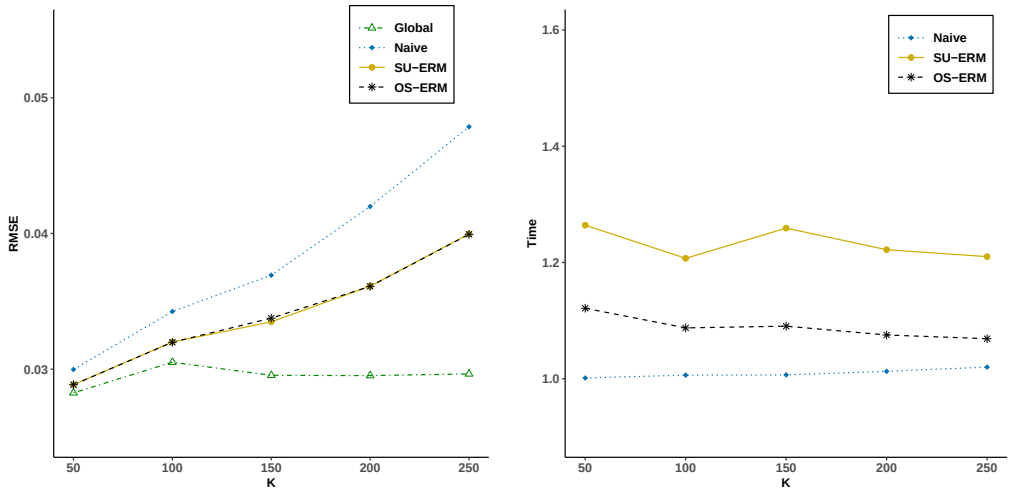
delivered by the global method. This is consistent with the conclusions reached under the previous experiment. It is noteworthy that in all cases, the CPs of both the SU-ERM and OS-ERM methods are close to the nominal level of 95%. This affirms that the efficiency of the proposed methods carries over to aspects of inference. The performance of the naive method, especially with respect to the criterion of CP, is sub-optimal.

7.2.3 Time Complexity Analysis

The computational time is presented in Figure 4(b), which shows that the naive method is the winner with respect to running time and the OS-ERM method is a close second. Although the SU-ERM method takes the longest to execute, the differences in running times between the three methods in absolute terms are in fact very small. Similar to the observation under the ranking problem, the value of K has no impact on the running times of all three methods. The running times of the global estimator, which are not shown in Figure 4(b), are 102.34, 378.78, 821.06, 1301.67, 2002.11, 2375.44, 2910.37 and 3270.46 seconds for K =25,50,75,100,125, 150, 175 and 200 respectively. The time-curve of using the global method therefore mimics a quadratic function of K, for example, if K (or N) increases by 10 times, the running time increases by nearly 100-fold. Thus, when there is a large volume of data, it will be difficult if not impossible to apply the global method.

(K, n)	Naive			SU-ERM			OS-ERM		
	θ_1^*	θ_2^*	θ_3^*	θ_1^*	θ_2^*	θ_3^*	θ_1^*	θ_2^*	θ_3^*
(50, 120)	0.970	0.940	0.955	0.975	0.945	0.975	0.975	0.945	0.975
(100, 60)	0.885	0.925	0.920	0.925	0.935	0.935	0.930	0.935	0.935
(150, 40)	0.870	0.875	0.890	0.923	0.943	0.933	0.922	0.938	0.933
(200, 30)	0.850	0.880	0.875	0.940	0.925	0.885	0.940	0.925	0.885
(250, 24)	0.855	0.805	0.795	0.875	0.875	0.885	0.880	0.875	0.890

Table 4: The coverage probabilities (CPs) of CIs corresponding to a 95% nominal level delivered by different methods when $N = 6000$.



(a) (b)

Figure 5: (a) RMSEs and (b) computational times of different methods when $N = 6000$. The computational time of the global estimator is about 2×10^3 for all K's.

7.2.4 Analysis for fixed N

Here, we investigate the empirical performance of the SU-ERM and OS-ERM methods for $K = 50$, 100, 150, 200, and 250, while fixing the total sample size to $N = 60000$. These values of K and N result in $n = 120$, 60, 40, 30, and 24.

Figure 5(a), which presents the RMSEs of estimators under above the setting, shows that the RMSE of the global estimator is generally insensitive to values of K (or equivalently, n). This is expected because the RMSE of the global estimator $\widehat{\theta}_N$ has a convergence rate that satisfies $\|\widehat{\theta}_N - \theta^*\|_2 = O_{\mathbb{P}}(\sqrt{p/N})$, and the latter's magnitude remains unchanged when N and p are fixed. On the other hand, the RMSEs of the naive, SU-ERM and OS-ERM estimators increase as n decreases. As well, for a given value of K, the SU-ERM and OS-ERM estimators produce very similar RMSEs that compare favourably with the RMSE arising from the naive method. Figure 5(b) shows that K has little effect on the running times of all methods, which in turn shows that n, the size of local samples, does not impact the running times of the methods under the AFT model.

Table 4 presents the coverage probabilities (CPs) associated with the nominal 95% confidence interval of different methods. When n is large, all methods deliver CPs that are reasonably close to the nominal 95% level. Other things being equal, a decrease in n has the effect of decreasing the CPs, but the rate of decrease varies across the methods. Generally speaking, the naive method produces CPs that experience the most rapid decline as n decreases; in particular, when $n < 60$, the naive method-based CPs are below 90%. On the other hand, the CPs based on the SU-ERM and OS-ERM methods fall below 90% only when $(K, n) = (250, 24)$. These results show the proposed methods are clearly superior to the naive method, especially when the local sample size is small.

8. Conclusions

U-statistic type ERM has ample applications in a wide range of statistical problems including bipartite ranking, survival analysis, metric learning, pairwise clustering and others, but this method usually requires summation over all pairs of observations, rendering computation difficult if not impossible on a single machine in the face of big data. Even with moderate sample sizes, performing an optimization routine on a single machine can be costly. There is a large collection of literature on distributed algorithms for ERM with univariate loss functions, but the results are inapplicable to pairwise cases. This is because under univariate losses, the objective function is separable across observations, and when applying the divide-and-conquer strategy, this separability feature allows the data to be stored in a distributed environment (for example, in a sensor network) and local statistics such as the local empirical risks and their gradients from the different subsets to be computed in parallel. In contrast, under bivariate losses, obtaining a complete U-statistic is difficult because the computation requires communications between all pairs of machines. The presence of massive data sets, as is often the case, will only worsen an already difficult situation.

This paper is an attempt to address the above issue. Based on the divide-and-conquer strategy, we propose two computationally and statistically efficient distributed estimation methods for the U-ERM: the SU-ERM and the OS-ERM methods. The SU-ERM method entails the construction of a surrogate empirical risk that can be computed in a distributed manner. The surrogate empirical risk consists of a local empirical risk computed only on one subset. Optimization based on this surrogate empirical risk can substantially reduce the computational complexity from $O(pN^2)$ to $O(pn^2)$. We have also developed a quadratic approximation to this surrogate empirical risk as an extension of the traditional one-step M-estimator to the case of U-ERM. This quadratic approximation has the advantage of yielding a closed-form analytical solution. We have shown that the resultant OS-ERM estimator has the same asymptotic efficiency as SU-ERM and the global U-estimators under the condition of $pK = o(n)$, and provided the theoretical upper bounds of the approximation errors and MSE of the OS-ERM and SU-ERM estimators. In addition, for the case of $K = O(n^A)$ $(A \geq 1)$, we have developed iterative algorithms to improve the performance of estimators, and demonstrated both theoretically and empirically that only a small number of iterations are required for the proposed estimators to achieve the same efficiency as the global U-estimator.

Our results can be readily extended to M_m estimators (Bose and Chatterjee, 2018a). Throughout the analysis we assume twice-differentiability of the loss functions, as the distributed estimation procedures being considered are Newton-type methods. When this assumption is unsustainable, the loss function can be substituted by a surrogate that is sufficiently smooth and the proposed methods can still be applied. One can further relax the restrictions about the smoothness of loss functions and develop a new method for U-ERM along the lines of the distributed first-order Newton-type estimator (FONE) proposed by Chen et al. (2021) that considers large-scale ERM with univariate loss functions. As FONE approximates the Newton step only through stochastic sub-gradients, it can accommodate non-differentiable loss functions. These remain for future research.

Acknowledgements

The authors wish to acknowledge financial support from the following funding bodies: National Natural Science Foundation of China (State Key Program: 71931004 (Zhou)), National Key R&D Program of China: 2021YFA1000100 (Zhou & Zhang), the Hong Kong Research Grant Council (General Research Fund: 11501522 (Wan)), National Natural Science Foundation of China (General Program: 72273129 (Wan), State Key Program: 72331005 (Zhang), Young Scientist Fund: 72201101 (Zhang)), Ministry of Education of Humanities and Social Sciences (Youth Fund: 22YJC910013 (Zhang)), and Shanghai Pujiang Program: 21PJC034 (Zhang). The authorship is in alphabetical order. The authors thank the editor, associate editor, and four referees for their comments and suggestions on an earlier version of the paper. Any errors that remain are the sole responsibility of the authors.

Appendix A.

This section gives the proofs of theorems. In the following, we let $\widehat{\theta}_n^1 = \widehat{\theta}_0$, that is, we assume that the estimator produced by the first machine is the initial estimator. It is instructive to note that our proofs also hold for the case where the naive estimator $\overline{\theta}_N$, which has superior properties to $\widehat{\theta}_n^1$, is used as the initial estimator.

A.1 Proofs of Theorems Related to SU-ERM

Let $\delta_\rho = \min\{\rho, \rho\lambda_-/4M\}$, $M_N = 1/(N(N-1))\sum_{i\neq j} M(z_i, z_j)$, and $M_k = 1/(n(n-1))\sum_{i\neq j} M(z_{ik}, z_{jk})$, $k = 1, \ldots, K$. Let us define the following "good" events:

$$\mathcal{E}_0 = \left\{ M_N \leq 2M, \left\|\nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}_0(\theta^*)\right\|_2 \leq \frac{\rho\lambda_-}{2}, \|\nabla \mathcal{L}_N(\theta^*)\|_2 \leq \frac{(1-\rho)\lambda_-}{2}\delta_1(\rho, \lambda_-, \lambda_+, M) \right\}.$$

and

$$\mathcal{E}_k = \left\{ M_k \leq 2M, \|\nabla^2 \mathcal{L}_n^k(\theta^*) - \nabla^2 \mathcal{L}_0(\theta^*)\|_2 \leq \frac{\rho\lambda_-}{4} \right\}$$

for $k = 1, \ldots, K$, where

$$\delta_1(\rho, \lambda_-, \lambda_+, M) = \min\left\{ \frac{(1-\rho)\lambda_-\delta_\rho}{32\lambda_+}, \sqrt{\frac{(1-\rho)\lambda_-\delta_\rho}{32M}} \right\}.$$

The proof of Proposition 3 requires the following three lemmas.

Lemma 14 *Under Conditions (C1), (C2), (C5) and (C6), there exist constants C and C' that depend only on ν, such that for $\nu \in \{1, \ldots, 8\}$ and $k \in \{1, \ldots, K\}$,*

$$\mathbb{E}\left[\left\|\nabla \mathcal{L}_n^k(\theta^*)\right\|_2^{2\nu}\right] \leq C\frac{p^\nu G^{2\nu}}{n^\nu} \quad \text{and} \quad \mathbb{E}\left[\|\nabla \mathcal{L}_N(\theta^*)\|_2^{2\nu}\right] \leq C\frac{p^\nu G^{2\nu}}{N^\nu}.$$

Furthermore, assume that there exists $\zeta > 0$ such that $p = O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau > 0$. Then

$$\mathbb{E}\left[\left\|\nabla^2 \mathcal{L}_n^k(\theta^*) - \nabla^2 \mathcal{L}_0(\theta^*)\right\|_2^{2\nu}\right] \leq C'\frac{(\log p)^\nu H^{2\nu}}{n^\nu} \quad \text{and}$$
$$\mathbb{E}\left[\left\|\nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}_0(\theta^*)\right\|_2^{2\nu}\right] \leq C'\frac{(\log p)^\nu H^{2\nu}}{N^\nu}.$$

Lemma 15 *Suppose that Conditions (C1)-(C6) hold. Assume that there exists $\zeta > 0$ such that $p = O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau > 0$. Then there exists a constant C'' independent of (n, K, N, p) such that*

$$\mathbb{P}\left(\bigcap_{k=0}^K \mathcal{E}_k\right) \geq 1 - C''\left(\frac{K(\log p)^8}{n^8} + \frac{p^8}{N^8}\right).$$

Lemma 16 *Suppose that Conditions (C1)-(C6) hold, $\log p = O(n^{7/8})$ and the initial estimator $\widehat{\theta}_0$ satisfies $\|\widehat{\theta}_0 - \theta^*\|_2 \leq \delta_1(\rho, \lambda_-, \lambda_+, M)$. Then, under the event $\bigcap_{k=0}^K \mathcal{E}_k$,*

$$\|\widetilde{\theta}_N - \widehat{\theta}_N\|_2 \leq \frac{2\|\nabla \mathcal{S}_N(\widehat{\theta}_N)\|_2}{(1-\rho)\lambda_-}.$$

Proof [Proof of Proposition 3] By Lemma 16, it suffices to prove the error bound for $\|\nabla \mathcal{S}_N(\widehat{\theta}_N)\|_2$ under the event $\bigcap_{k=0}^{K} \mathcal{E}_k$. Note that

$$\nabla \mathcal{S}_N(\widehat{\theta}_N) = \nabla \mathcal{L}_n^1(\widehat{\theta}_N) - \nabla \mathcal{L}_n^1(\widehat{\theta}_0) + \nabla \widetilde{\mathcal{L}}_N(\widehat{\theta}_0). \tag{32}$$

By the first-order optimality condition, the solution $\widehat{\theta}_N$ for the global empirical risk $\mathcal{L}_N$ satisfies $\nabla \mathcal{L}_N(\widehat{\theta}_N) = 0$. As well, by subtracting $\nabla \mathcal{L}_N(\widehat{\theta}_N)$ from the l.h.s. and r.h.s. of (32), we can write

$$\nabla \mathcal{S}_N(\widehat{\theta}_N) = \left(\nabla \mathcal{L}_n^1(\widehat{\theta}_N) - \nabla \mathcal{L}_n^1(\widehat{\theta}_0)\right) - \left(\nabla \mathcal{L}_N(\widehat{\theta}_N) - \nabla \widetilde{\mathcal{L}}_N(\widehat{\theta}_0)\right).$$

Using the integral form of the Taylor's expansion, we have, for $k = 1, \ldots, K$,

$$\nabla \mathcal{L}_n^k(\widehat{\theta}_N) - \nabla \mathcal{L}_n^k(\widehat{\theta}_0) = H_k(\widehat{\theta}_N - \widehat{\theta}_0),$$

where $H_k = \int_0^1 \nabla^2 \mathcal{L}_n^k(\widehat{\theta}_0 + t(\widehat{\theta}_N - \widehat{\theta}_0))dt$, and H_k satisfies

$$\begin{aligned}\left\|H_k - \nabla^2 \mathcal{L}_n^k(\theta^*)\right\|_2 &\leq \left\|H_k - \nabla^2 \mathcal{L}_n^k(\widehat{\theta}_N)\right\|_2 + \left\|\nabla^2 \mathcal{L}_n^k(\widehat{\theta}_N) - \nabla^2 \mathcal{L}_n^k(\theta^*)\right\|_2 \\ &\leq 2M\left(\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 + \|\widehat{\theta}_N - \theta^*\|_2\right).\end{aligned}$$

Similarly, $\nabla \mathcal{L}_N(\widehat{\theta}_N) - \nabla \mathcal{L}_N(\widehat{\theta}_0) = H_N(\widehat{\theta}_N - \widehat{\theta}_0)$, where $H_N = \int_0^1 \nabla^2 \mathcal{L}_N(\widehat{\theta}_0 + t(\widehat{\theta}_N - \widehat{\theta}_0))dt$ and satisfies

$$\left\|H_N - \nabla^2 \mathcal{L}_N(\theta^*)\right\|_2 \leq 2M\left(\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 + \|\widehat{\theta}_N - \theta^*\|_2\right).$$

Therefore, we can write

$$\begin{aligned}&\left\|\nabla \mathcal{S}_N(\widehat{\theta}_N)\right\|_2 \\ =&\left\|\left(\nabla \mathcal{L}_n^1(\widehat{\theta}_N) - \nabla \mathcal{L}_n^1(\widehat{\theta}_0)\right) - \left(\nabla \mathcal{L}_N(\widehat{\theta}_N) - \nabla \mathcal{L}_N(\widehat{\theta}_0)\right) - \left(\nabla \mathcal{L}_N(\widehat{\theta}_0) - \nabla \widetilde{\mathcal{L}}_N(\widehat{\theta}_0)\right)\right\|_2 \\ \leq&\left\|\left(\nabla \mathcal{L}_n^1(\widehat{\theta}_N) - \nabla \mathcal{L}_n^1(\widehat{\theta}_0)\right) - \left(\nabla \mathcal{L}_N(\widehat{\theta}_N) - \nabla \mathcal{L}_N(\widehat{\theta}_0)\right)\right\|_2 + \left\|\nabla \mathcal{L}_N(\widehat{\theta}_0) - \nabla \widetilde{\mathcal{L}}_N(\widehat{\theta}_0)\right\|_2 \\ =:&R_1 + R_2.\end{aligned}$$

Note that

$$\begin{aligned}\|R_1\|_2 \leq&\left\|\left(H_1 - \nabla^2 \mathcal{L}_n^1(\theta^*)\right)(\widehat{\theta}_N - \widehat{\theta}_0)\right\|_2 + \left\|\left(H_N - \nabla^2 \mathcal{L}_N(\theta^*)\right)(\widehat{\theta}_N - \widehat{\theta}_0)\right\|_2 \\ &+ \left\|\left(\nabla^2 \mathcal{L}_n^1(\theta^*) - \nabla^2 \mathcal{L}_N(\theta^*)\right)(\widehat{\theta}_N - \widehat{\theta}_0)\right\|_2 \\ \leq&\left(4M\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 + 4M\|\widehat{\theta}_N - \theta^*\|_2 + \|\nabla^2 \mathcal{L}_n^1(\theta^*) - \nabla^2 \mathcal{L}_N(\theta^*)\|_2\right)\|\widehat{\theta}_N - \widehat{\theta}_0\|_2\end{aligned}$$

and

$$R_2 = \left(\nabla \mathcal{L}_N(\widehat{\theta}_0) - \nabla \mathcal{L}_N(\theta^*)\right) - \left(\nabla \widetilde{\mathcal{L}}_N(\widehat{\theta}_0) - \nabla \widetilde{\mathcal{L}}_N(\theta^*)\right) + \left(\nabla \mathcal{L}_N(\theta^*) - \nabla \widetilde{\mathcal{L}}_N(\theta^*)\right).$$

Similar to the expansion for $\|R_1\|_2$ above, we can write

$$\begin{aligned}&\left\|\left(\nabla \mathcal{L}_N(\widehat{\theta}_0) - \nabla \mathcal{L}_N(\theta^*)\right) - \left(\nabla \widetilde{\mathcal{L}}_N(\widehat{\theta}_0) - \nabla \widetilde{\mathcal{L}}_N(\theta^*)\right)\right\|_2 \\ \leq&\left(4M\|\widehat{\theta}_0 - \theta^*\|_2 + \left\|\nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \widetilde{\mathcal{L}}_N(\theta^*)\right\|_2\right)\|\widehat{\theta}_0 - \theta^*\|_2.\end{aligned}$$

Therefore,

$$\begin{aligned}\|R_2\|_2 \leq & \left(4M\|\widehat{\theta}_0-\theta^*\|_2+\left\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\widetilde{\mathcal{L}}_N(\theta^*)\right\|_2\right)\|\widehat{\theta}_0-\theta^*\|_2 \\ & +\left\|\nabla\mathcal{L}_N(\theta^*)-\nabla\widetilde{\mathcal{L}}_N(\theta^*)\right\|_2.\end{aligned}$$

Combining these results and Lemmas 15 and 16, we obtain

$$\begin{aligned}\|\widetilde{\theta}_N-\widehat{\theta}_N\|_2 \leq & C\left(\|\widehat{\theta}_0-\widehat{\theta}_N\|_2+\|\widehat{\theta}_N-\theta^*\|_2+\left\|\nabla^2\mathcal{L}_n^1(\theta^*)-\nabla^2\mathcal{L}_N(\theta^*)\right\|_2\right)\|\widehat{\theta}_0-\widehat{\theta}_N\|_2 \\ & +C\left(\|\widehat{\theta}_0-\theta^*\|_2+\left\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\widetilde{\mathcal{L}}_N(\theta^*)\right\|_2\right)\|\widehat{\theta}_0-\theta^*\|_2 \\ & +\left\|\nabla\mathcal{L}_N(\theta^*)-\nabla\widetilde{\mathcal{L}}_N(\theta^*)\right\|_2\end{aligned}$$

with probability no smaller than $1-C'\left(K(\log p)^8/n^8+p^8/N^8\right)$, where the constants C and C' are independent of (K,n,N,p). ■

Proof [Proof of Theorem 4] First, denote C as an arbitrary constant independent of (K,n,N,p). Applying the integral form of Taylor's expansion with respect to the global empirical risk minimizer $\widehat{\theta}_N$, we can write

$$0=\nabla\mathcal{L}_N(\widehat{\theta}_N)=\nabla\mathcal{L}_N(\theta^*)+H_N(\widehat{\theta}_N-\theta^*),$$

where $H_N=\int_0^1\nabla^2\mathcal{L}_N(\theta^*+t(\widehat{\theta}_N-\theta^*))dt$. By a straightforward algebraic operation, we can show that

$$\widehat{\theta}_N-\theta^*=-\nabla^2\mathcal{L}_0(\theta^*)^{-1}\left(\nabla\mathcal{L}_N(\theta^*)+U_N(\widehat{\theta}_N-\theta^*)+V_N(\widehat{\theta}_N-\theta^*)\right),$$

where $U_N=H_N-\nabla^2\mathcal{L}_N(\theta^*)$ and $V_N=\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)$. Note that under the event $\mathcal{E}_0$, we have $\|\widehat{\theta}_N-\theta^*\|_2\leq\frac{2\|\nabla\mathcal{L}_N(\theta^*)\|_2}{(1-\rho)\lambda_-}$. This leads to

$$\begin{aligned}& \left\|\widehat{\theta}_N-\theta^*+\nabla^2\mathcal{L}_0(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_2 \\ \leq & \frac{1}{\lambda_-}\|U_N(\widehat{\theta}_N-\theta^*)\|_2+\frac{1}{\lambda_-}\|V_N(\widehat{\theta}_N-\theta^*)\|_2 \\ \leq & \frac{2M}{\lambda_-}\|\widehat{\theta}_N-\theta^*\|_2^2+\frac{1}{\lambda_-}\left\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\right\|_2\|\widehat{\theta}_N-\theta^*\|_2 \\ \leq & \frac{8M}{(1-\rho)^2\lambda_-^3}\|\nabla\mathcal{L}_N(\theta^*)\|_2^2+\frac{2}{(1-\rho)\lambda_-^2}\left\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\right\|_2\|\nabla\mathcal{L}_N(\theta^*)\|_2.\end{aligned}$$

Recognizing the above inequality and Proposition 3, we have, under the event $\bigcap_{k=0}^{K}\mathcal{E}_k$,

$$
\begin{aligned}
&\|\widetilde{\theta}_N-\theta^*+\nabla^2\mathcal{L}_0(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\|_2\\
\leq&\|\widehat{\theta}_N-\theta^*+\nabla^2\mathcal{L}_0(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\|_2+\|\widetilde{\theta}_N-\widehat{\theta}_N\|_2\\
\leq&C\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\|_2\|\nabla\mathcal{L}_N(\theta^*)\|_2+C\|\nabla\mathcal{L}_N(\theta^*)\|_2^2\\
&+C\left(\|\widehat{\theta}_0-\widehat{\theta}_N\|_2+\|\widehat{\theta}_N-\theta^*\|_2+\|\nabla^2\mathcal{L}_n^1(\theta^*)-\nabla^2\mathcal{L}_N(\theta^*)\|_2\right)\|\widehat{\theta}_0-\widehat{\theta}_N\|_2\\
&+C\left(\|\widehat{\theta}_0-\theta^*\|_2+\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\widetilde{\mathcal{L}}_N(\theta^*)\|_2\right)\|\widehat{\theta}_0-\theta^*\|_2+\|\nabla\mathcal{L}_N(\theta^*)-\nabla\widetilde{\mathcal{L}}_N(\theta^*)\|_2\\
\leq&C\left\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\right\|_2\|\nabla\mathcal{L}_N(\theta^*)\|_2+C\|\nabla\mathcal{L}_N(\theta^*)\|_2^2\\
&+C\Bigg(\|\widehat{\theta}_0-\theta^*\|_2+\|\nabla\mathcal{L}_N(\theta^*)\|_2+\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\|_2\\
&\qquad+\|\nabla^2\mathcal{L}_n^1(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\|_2+\frac{1}{K-1}\sum_{k\neq 1}\|\nabla^2\mathcal{L}_n^k(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\|_2\Bigg)\|\widehat{\theta}_0-\theta^*\|_2\\
&+C\|\nabla^2\mathcal{L}_n^1(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\|_2\|\nabla\mathcal{L}_N(\theta^*)\|_2+\|\nabla\mathcal{L}_N(\theta^*)-\nabla\widetilde{\mathcal{L}}_N(\theta^*)\|_2.
\end{aligned}
$$

Also, using Hölder's inequality and Lemma 14 together, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\left(\widetilde{\theta}_N-\theta^*+\nabla^2\mathcal{L}_0(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right)I\left(\bigcap_{k=0}^{K}\mathcal{E}_k\right)\right\|_2^2\right]\\
\leq&C\sqrt{\mathbb{E}\left[\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\|_2^4\right]\mathbb{E}\left[\|\nabla\mathcal{L}_N(\theta^*)\|_2^4\right]}+C\mathbb{E}\left[\|\nabla\mathcal{L}_N(\theta^*)\|_2^4\right]+C\mathbb{E}\left[\|\widehat{\theta}_0-\theta^*\|_2^4\right]\\
&+C\Bigg(\sqrt{\mathbb{E}[\|\nabla\mathcal{L}_N(\theta^*)\|_2^4]}+\sqrt{\mathbb{E}[\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\|_2^4]}+\sqrt{\mathbb{E}[\|\nabla^2\mathcal{L}_n^1(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\|_2^4]}\\
&\qquad+\frac{1}{K-1}\sum_{k\neq 1}\sqrt{\mathbb{E}[\|\nabla^2\mathcal{L}_n^k(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\|_2^4]}\Bigg)\sqrt{\mathbb{E}[\|\widehat{\theta}_0-\theta^*\|_2^4]}\\
&+C\sqrt{\mathbb{E}\left[\|\nabla^2\mathcal{L}_n^1(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\|_2^4\right]\mathbb{E}\left[\|\nabla\mathcal{L}_N(\theta^*)\|_2^4\right]}+\mathbb{E}\left[\left\|\nabla\mathcal{L}_N(\theta^*)-\nabla\widetilde{\mathcal{L}}_N(\theta^*)\right\|_2^2\right]\\
\leq&\frac{Cp}{N}\max\left\{\frac{p}{N},\frac{\log p}{n}\right\}+C\sqrt{\mathbb{E}\left[\|\widehat{\theta}_0-\theta^*\|_2^4\right]}\max\left\{\frac{p}{N},\frac{\log p}{n},\sqrt{\mathbb{E}\left[\|\widehat{\theta}_0-\theta^*\|_2^4\right]}\right\}\\
&\quad+\mathbb{E}\left[\left\|\nabla\mathcal{L}_N(\theta^*)-\nabla\widetilde{\mathcal{L}}_N(\theta^*)\right\|_2^2\right].
\end{aligned}
$$

Let us consider the term $\nabla\mathcal{L}_N(\theta^*)$ in $\mathbb{E}[\|\nabla\mathcal{L}_N(\theta^*)-\nabla\widetilde{\mathcal{L}}_N(\theta^*)\|_2^2]$. As $\nabla\mathcal{L}_N(\theta^*)$ is a U-statistic, we denote it as $U_N=2\{N(N-1)\}^{-1}\sum_{1\leq i<j\leq N}h(X_i,X_j)$ for simplicity. By applying the Hoeffding's Decomposition (see the Supplementary Material of Chen and Peng (2021)), we can write

$$
U_N=\theta_U+N^{-1}\sum_{i=1}^{N}\alpha_U(Z_i)+N^{-2}\sum_{1\leq i<j\leq N}\beta_U(Z_i,Z_j)+R_{UN},
$$

where $\theta_U = \mathrm{E}(U_N)$, $\alpha_U(z) = 2[\mathrm{E}\{h(z, Z_2)\} - \theta_U]$, $\beta_U(z_1, z_2) = 2h(z_1, z_2) - \alpha_U(z_1) - \alpha_U(z_2) - 2\theta_U$, and R_{UN} is a remainder term that satisfies $\mathrm{E}(R_{UN}) = 0$ and $\mathrm{E}\left(\|R_{UN}\|_2^2\right) = O\left(p/N^4\right)$. As well, Proposition 1 of Chen and Peng (2021) demonstrates that

$$E\left\|N^{-2}\sum_{1\le i<j\le N}\beta_U(Z_i, Z_j)\right\|_2^2 = O(p/N^2).$$

Similarly, for $\nabla\widetilde{\mathcal{L}}_N(\theta^*)$, which is a distributed U-statistic, we denote it as $U_{N,K} = K^{-1}\sum_{j=1}^K U^{(k)} = K^{-1}\sum_{k=1}^K\{\frac{2}{n(n-1)}\sum_{1\le i<j\le n} h(Z_{k,i}, Z_{k,j})\}$ for simplicity. Following the same procedure as above, we have

$$U_{N,K} = \theta_U + N^{-1}\sum_{i=1}^N \alpha_U(Z_i) + N^{-1}\sum_{k=1}^K n^{-1}\sum_{1\le i<j\le n}\beta_U(Z_{k,i}, Z_{k,j}) + R_{UN,K},$$

where the remainder term $R_{UN,K}$ satisfies $\mathrm{E}(R_{UN,K}) = 0$ and $\mathrm{E}\left(\|R_{UN,K}\|_2^2\right) = O\left(pK^3N^{-4}\right)$. Also, Theorem 3.1 of Chen and Peng (2021) shows that

$$E\left\|N^{-1}\sum_{k=1}^K n^{-1}\sum_{1\le i<j\le n}\beta_U(Z_{k,i}, Z_{k,j})\right\|_2^2 = O(Kp/N^2).$$

Combining the above results, we can write

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla\mathcal{L}_N(\theta^*) - \nabla\widetilde{\mathcal{L}}_N(\theta^*)\right\|_2^2\right] \le& 4\mathbb{E}\left\|N^{-2}\sum_{1\le i<j\le N}\beta_U(Z_i, Z_j)\right\|_2^2 + 4\mathbb{E}\|R_{UN}\|_2^2 \\ &+ 4\mathbb{E}\left\|N^{-1}\sum_{k=1}^K n^{-1}\sum_{1\le i<j\le n}\beta_U(Z_{k,i}, Z_{k,j})\right\|_2^2 + 4\mathbb{E}\|R_{UN,K}\|_2^2 \\ \le& \frac{Cp}{N^2} + \frac{Cp}{N^4} + \frac{CKp}{N^2} + \frac{CpK^3}{N^4} \le \frac{Cp}{nN}. \end{aligned} \tag{33}$$

Therefore,

$$\begin{aligned}&\mathbb{E}\left[\left\|\left(\widetilde{\theta}_N - \theta^* + \nabla^2\mathcal{L}_0(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right) I\left(\bigcap_{k=0}^K \mathcal{E}_k\right)\right\|_2^2\right] \\ \le& \frac{Cp}{N}\max\left\{\frac{p}{N}, \frac{\log p}{n}\right\} + C\sqrt{\mathbb{E}\left[\|\widehat{\theta}_0 - \theta^*\|_2^4\right]}\max\left\{\frac{p}{N}, \frac{\log p}{n}, \sqrt{\mathbb{E}\left[\|\widehat{\theta}_0 - \theta^*\|_2^4\right]}\right\}\end{aligned}$$

Under Lemma 15 and Conditions (C2) and (C3), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\theta}_N-\theta^*+\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}\nabla\mathcal{L}_N\left(\theta^*\right)\right\|_2^2\right]\\
\leq& C\sqrt{\mathbb{E}\left[\|\widehat{\theta}_0-\theta^*\|_2^4\right]}\max\left\{\frac{p}{N},\frac{\log p}{n},\sqrt{\mathbb{E}\left[\|\widehat{\theta}_0-\theta^*\|_2^4\right]}\right\}\\
&+C\left(\frac{Cp}{N}\max\left\{\frac{p}{N},\frac{\log p}{n}\right\}+\frac{K(\log p)^8}{n^8}+\frac{p^8}{N^8}\right).
\end{aligned}$$

It can be derived, through a generalization of the properties of U-estimation (Bose and Chatterjee, 2018b) to the case of diverging dimension case, that

$$\mathbb{E}\left[\left\|\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}\nabla\mathcal{L}_N\left(\theta^*\right)\right\|_2^2\right]=\frac{4}{N}\mathbb{E}\left[\left\|\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}g(\theta^*;Z)\right\|_2^2\right]+O(p\log N\log\log N/N^2).$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\left\|\widetilde{\theta}_N-\theta^*\right\|_2^2\right]\leq&\mathbb{E}\left[\left\|\widetilde{\theta}_N-\theta^*+\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}\nabla\mathcal{L}_N\left(\theta^*\right)\right\|_2^2\right]+\mathbb{E}\left[\left\|\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}\nabla\mathcal{L}_N\left(\theta^*\right)\right\|_2^2\right]\\
&+2\sqrt{\mathbb{E}\left[\left\|\widetilde{\theta}_N-\theta^*+\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}\nabla\mathcal{L}_N\left(\theta^*\right)\right\|_2^2\right]\mathbb{E}\left[\left\|\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}\nabla\mathcal{L}_N\left(\theta^*\right)\right\|_2^2\right]}\\
\leq&\frac{4A}{N}+C\left(\gamma_{n,K,p}^2(\widehat{\theta}_0)+\frac{\gamma_{n,K,p}(\widehat{\theta}_0)\sqrt{A}}{\sqrt{N}}+\frac{p\log N\log\log N}{N^2}\right),
\end{aligned}$$

where $A=\mathbb{E}[\left\|\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}g\left(\theta^*;Z\right)\right\|_2^2]$ and

$$\begin{aligned}
\gamma_{n,K,p}(\widehat{\theta}_0)=&\sqrt{\mathbb{E}\left[\|\widehat{\theta}_0-\theta^*\|_2^4\right]}\max\left\{\frac{p}{N},\frac{\log p}{n},\sqrt{\mathbb{E}\left[\|\widehat{\theta}_0-\theta^*\|_2^4\right]}\right\}\\
&+\left(\frac{p}{N}\max\left\{\frac{p}{N},\frac{\log p}{n}\right\}+\frac{K(\log p)^8}{n^8}\right).
\end{aligned}$$

This completes the proof. ■

Proof [Proof of Theorem 9] Write $\widetilde{\theta}_N-\theta^*$ as $\widetilde{\theta}_N-\theta^*=\widetilde{\theta}_N-\widehat{\theta}_N+\widehat{\theta}_N-\theta^*$. It follows from properties of U-estimation (Wang et al., 2009) that

$$\widehat{\theta}_N-\theta^*=-\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}\frac{2}{N}\sum_{i=1}^N g\left(\theta^*;Z_i\right)+o_{\mathbb{P}}\left(\sqrt{p/N}\right).$$

Also, from Proposition 3, if $\|\widehat{\theta}_0-\theta^*\|_2=O_{\mathbb{P}}(\sqrt{p/n})$, it can be shown that

$$\|\widetilde{\theta}_N-\widehat{\theta}_N\|_2=O_{\mathbb{P}}\left((p/n)^{1/2}\|\widehat{\theta}_0-\theta^*\|_2\right)+O_{\mathbb{P}}\left(\|\nabla\mathcal{L}_N\left(\theta^*\right)-\nabla\widetilde{\mathcal{L}}_N\left(\theta^*\right)\|_2\right).$$

By (33), $\|\nabla\mathcal{L}_N(\theta^*) - \nabla\widetilde{\mathcal{L}}_N(\theta^*)\|_2 = O_{\mathbb{P}}(\sqrt{p/(nN)})$. Hence, for any $v_0 \in \mathbb{R}^p$,

$$v_0^\top\widetilde{\theta}_N - v_0^\top\theta^* = -v_0'\nabla^2\mathcal{L}_0(\theta^*)^{-1}\frac{2}{N}\sum_{i=1}^{N} g(\theta^*; Z_i) + O_{\mathbb{P}}\left((p/n)^{1/2}\|\widehat{\theta}_0 - \theta^*\|_2\right) + o_{\mathbb{P}}(\sqrt{p/N}).$$

If $pK = o(n)$, then $\|\widehat{\theta}_0 - \theta^*\|_2 = o_{\mathbb{P}}(\sqrt{1/K})$, and we have

$$\frac{\sqrt{N}v_0^\top(\widetilde{\theta}_N - \theta^*)}{\sqrt{4v_0^\top\mathcal{L}_0(\theta^*)^{-1}V_0\mathcal{L}_0(\theta^*)^{-1}v_0}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

This completes the proof.

■

A.2 Proofs of Theorems Related to OS-ERM

As the surrogate empirical risks $\mathcal{S}_N^Q$ and $\mathcal{S}_N$ are identical in the two highest orders of the Taylor's expansion, the OS-ERM estimator $\widetilde{\theta}_N^Q$ and the SU-ERM estimator $\widetilde{\theta}_N$ enjoy similar asymptotic properties. Therefore, we only give the probability bound on $\|\widetilde{\theta}_N^Q - \widehat{\theta}_N\|_2$ and the proof of Theorem 8. Theorem 9 can be similarly derived.

Let us define the following group of "good" events:

$$\mathcal{E}_0' = \left\{M_N \le 2M, \left\|\nabla^2\mathcal{L}_N(\theta^*) - \nabla^2\mathcal{L}_0(\theta^*)\right\|_2 \le \frac{\rho\lambda_-}{4}, \|\nabla\mathcal{L}_N(\theta^*)\|_2 \le \frac{(1-\rho)^2\lambda_-^2}{32M}\right\},$$

and

$$\mathcal{E}_k' = \left\{M_k \le 2M, \left\|\nabla^2\mathcal{L}_n^k(\theta^*) - \nabla^2\mathcal{L}_0(\theta^*)\right\|_2 \le \frac{\rho\lambda_-}{4}\right\}$$

for $k = 1, \ldots, K$. Note that under the event $\mathcal{E}_0'$, we have $\|\widehat{\theta}_N - \theta^*\|_2 \le (1-\rho)\lambda_-/(16M)$. Consider the following lemma.

Lemma 17 *Assume the conditions in Proposition 7 hold. Then under the event* $\bigcap_{k=0}^K \mathcal{E}_k'$, *we have*

$$\lambda_{\min}\left[\nabla^2\mathcal{L}_N(\widehat{\theta}_N)\right] \ge \frac{1}{2}(1-\rho)\lambda_-, \quad \lambda_{\min}\left[\nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)\right] \ge \frac{1}{2}(1-\rho)\lambda_-,$$

$$\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 \le \Delta := \frac{(1-\rho)\lambda_-}{8M},$$

$$V_N := \max_{\theta\in(\widehat{\theta}_N-\Delta, \widehat{\theta}_N+\Delta)}\left\|\nabla^2\mathcal{L}_N(\theta)\right\|_2 \le V := 4M\Delta + \frac{\rho\lambda_-}{4} + \lambda_+,$$

and

$$\left\|\nabla^2\mathcal{L}_N(\widehat{\theta}_0)^{-1} - \nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)^{-1}\right\|_2 \le C\left(\left\|\nabla^2\mathcal{L}_N(\theta^*) - \nabla^2\mathcal{L}_n^1(\theta^*)\right\|_2 + 4M\|\widehat{\theta}_0 - \theta^*\|_2\right),$$

where $C = 2M\rho/\lambda_-^2 + (\rho+4)/(4\lambda_-)$, *and* $\lambda_{\min}[A]$ *denotes the minimal eigenvalue of a matrix* A.

Proof [Proof of Lemma 17] The proofs of all of the inequalities stated above, except for the inequality on $\lambda_{\min}[\nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)]$, are given in Lemma C.4 of Jordan et al. (2019). These proofs can be easily extended to the case of pairwise losses. Here, we verify the inequality on $\lambda_{\min}[\nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)]$ stated above. Note that under Conditions (C3) and (C5) and the event $\mathcal{E}_1'$,

$$\begin{aligned}\lambda_{\min}\left[\nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)\right] \geq & \lambda_{\min}\left[\nabla^2\mathcal{L}_0\left(\theta^*\right)\right] - \left\|\nabla^2\mathcal{L}_n^1\left(\theta^*\right) - \nabla^2\mathcal{L}_0\left(\theta^*\right)\right\|_2 \\ & - \left\|\nabla^2\mathcal{L}_n^1(\widehat{\theta}) - \nabla^2\mathcal{L}_n^1\left(\theta^*\right)\right\|_2 \\ \geq & \lambda_- - \frac{\rho\lambda_-}{4} - 2M\|\widehat{\theta}_0 - \theta^*\|_2.\end{aligned}$$

Since $\|\widehat{\theta}_0 - \theta^*\|_2 \leq (2+\rho)\lambda_-^2/(8M)$, we have, under the event $\mathcal{E}_1'$,

$$\lambda_{\min}\left[\nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)\right] \geq \frac{1}{2}(1-\rho)\lambda_-.$$

This completes the proof. ■

Proof [Proof of Proposition 7] Analogous to the proof of Lemma 15, we can prove that under Conditions (C2)-(C6),

$$\mathbb{P}\left(\bigcup_{k=0}^{K}(\mathcal{E}_k')^c\right) \leq C''\left(\frac{K(\log p)^8}{n^8} + \frac{p^8}{N^8}\right).$$

where C'' is some constant independent of (n, K, N, p). Define the following global one-step estimator:

$$\widehat{\theta}_N^Q = \widehat{\theta}_0 - \nabla^2\mathcal{L}_N^{-1}(\widehat{\theta}_0)\nabla\mathcal{L}_N(\widehat{\theta}_0).$$

Note that

$$\begin{aligned}\widetilde{\theta}_N^Q - \widehat{\theta}_N^Q &= \left(\widehat{\theta}_0 - \nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)^{-1}\nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0)\right) - \left(\widehat{\theta}_0 - \nabla^2\mathcal{L}_N(\widehat{\theta}_0)^{-1}\nabla\mathcal{L}_N(\widehat{\theta}_0)\right) \\ &= \left(\nabla^2\mathcal{L}_N(\widehat{\theta}_0)^{-1} - \nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)^{-1}\right)\nabla\mathcal{L}_N(\widehat{\theta}_0) + \nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)^{-1}\left(\nabla\mathcal{L}_N(\widehat{\theta}_0) - \nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0)\right) \\ &= \left(\nabla^2\mathcal{L}_N(\widehat{\theta}_0)^{-1} - \nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)^{-1}\right)\left(\nabla\mathcal{L}_N(\widehat{\theta}_0) - \nabla\mathcal{L}_N(\widehat{\theta}_N)\right) \\ &\quad + \nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)^{-1}\left(\nabla\mathcal{L}_N(\widehat{\theta}_0) - \nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0)\right) \\ &=: R_1 + R_2.\end{aligned}$$

By Lemma 17, we can show that

$$\begin{aligned}\|R_1\|_2 &\leq V_N\left\|\nabla^2\mathcal{L}_N(\widehat{\theta}_0)^{-1} - \nabla^2\mathcal{L}_n^1(\widehat{\theta}_0)^{-1}\right\|_2\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 \\ &\leq CV_N\left(\left\|\nabla^2\mathcal{L}_N\left(\theta^*\right) - \nabla^2\mathcal{L}_n^1\left(\theta^*\right)\right\|_2 + 4M\|\widehat{\theta}_0 - \theta^*\|_2\right)\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2,\end{aligned}$$

and

$$\begin{aligned}\|R_2\|_2 \leq & \frac{2}{(1-\rho)\lambda_-}\left(4M\|\widehat{\theta}_0 - \theta^*\|_2 + \left\|\nabla^2\mathcal{L}_N\left(\theta^*\right) - \nabla^2\widetilde{\mathcal{L}}_N\left(\theta^*\right)\right\|_2\right)\|\widehat{\theta}_0 - \theta^*\|_2 \\ & + \frac{2}{(1-\rho)\lambda_-}\left\|\nabla\mathcal{L}_N\left(\theta^*\right) - \nabla\widetilde{\mathcal{L}}_N\left(\theta^*\right)\right\|_2.\end{aligned}$$

Lemma 17 also implies that

$$\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 \leq \frac{(1-\rho)\lambda_-}{8M} \leq \frac{\lambda_{\min}\left[\nabla^2 \mathcal{L}_N(\widehat{\theta}_N)\right]}{2M_N},$$

which is the condition for Theorem 5.3 of Bubeck (2015), resulting in

$$\|\widehat{\theta}_N^Q - \widehat{\theta}_N\|_2 \leq \frac{4M}{(1-\rho)\lambda_-}\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2^2.$$

Combining the above results, we obtain, under the event $\bigcap_{k=0}^K \mathcal{E}'_k$,

$$\begin{aligned}
\|\widetilde{\theta}_N^Q - \widehat{\theta}_N\|_2 \leq & \|\widetilde{\theta}_N^Q - \widehat{\theta}_N^Q\|_2 + \|\widehat{\theta}_N^Q - \widehat{\theta}_N\|_2 \\
\leq & C_1\left(\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 + \left\|\widehat{\theta}_N - \theta^*\right\|_2 + \left\|\nabla^2 \mathcal{L}_n^1(\theta^*) - \nabla^2 \mathcal{L}_N(\theta^*)\right\|_2\right)\|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 \\
& + C_1\left(\|\widehat{\theta}_0 - \theta^*\|_2 + \left\|\nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \widetilde{\mathcal{L}}_N(\theta^*)\right\|_2\right)\|\widehat{\theta}_0 - \theta^*\|_2 \\
& + \left\|\nabla \mathcal{L}_N(\theta^*) - \nabla \widetilde{\mathcal{L}}_N(\theta^*)\right\|_2,
\end{aligned}$$

where C_1 is independent of (K, n, N, p). ■

References

Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10(2):441–474, 2009.

Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 1756–1764. MIT Press: Cambridge, U.S.A., 2015.

Peter L Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3):1352, 2018.

Peter J Bickel. One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434, 1975.

Arup Bose. Bahadur representation of m_m estimates. *The Annals of Statistics*, 26:771–777, 1998.

Arup Bose and Snigdhansu Chatterjee. Resampling U-statistics and M-estimators. In *U-Statistics, M_m-Estimators and Resampling*, pages 103–125. Springer: Singapore, 2018a.

Arup Bose and Snigdhansu Chatterjee. *U-Statistics, M_m-Estimators and Resampling*. Springer: Singapore, 2018b.

Bruce M Brown and You Gan Wang. Induced smoothing for rank regression with censored survival times. *Statistics in Medicine*, 26(4):828–836, 2007.

Charles George Broyden. The convergence of a class of double-rank minimization algorithms. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

Sébastien Bubeck. Convex optimization: algorithms and complexity. In *Foundations and Trends in Machine Learning, Vol.8, No. 3-4*, pages 231–357. 2015.

Tianxi Cai, Molei Liu, and Yin Xia. Privacy-preserving integrative regression analysis of high-dimensional heterogeneous data. *arXiv preprint arXiv:1902.06115*, 2019.

Lanjue Chen and Yong Zhou. Quantile regression in big data: a divide and conquer based strategy. *Computational Statistics & Data Analysis*, 144:106892, 2019.

Song Xi Chen and Liuhua Peng. Distributed statistical inference for massive data. *The Annals of Statistics*, 49(5):2851–2869, 2021.

Xi Chen, Weidong Liu, and Yichen Zhang. Quantile regression under memory constraint. *The Annals of Statistics*, 47:3244–3273, 2019.

Xi Chen, Weidong Liu, and Yichen Zhang. First-order Newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, pages 1–17, 2021.

Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24:1655–1684, 2014.

Matthias Chung, Qi Long, and Brent A Johnson. A tutorial on rank-based coefficient estimation for censored data in small-and large-scale problems. *Statistics & Computing*, 23(5):601–614, 2013.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and scoring using empirical risk minimization. In *Lecture Notes in Computer Science, Vol. 3559*, pages 1–15. Springer, Heidelberg, 2005.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Stephan Clémençon, Igor Colin, and Aurélien Bellet. Scaling-up empirical risk minimization: optimization of incomplete u-statistics. *Journal of Machine Learning Research*, 17(1): 2682–2717, 2016.

Jianqing Fan, Dong Wang, Kaizheng Wang, Ziwei Zhu, et al. Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009–3031, 2019.

Jianqing Fan, Yongyi Guo, and Kaizheng Wang. Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*, pages 1–11, 2021.

Roger Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13 (3):317–322, 1970.

Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4(6):933–969, 2003.

Mendel Fygenson and Ya'acov Ritov. Monotone estimating equations for censored data. *The Annals of Statistics*, pages 732–746, 1994.

Wei Gao, Rong Jin, Shenghuo Zhu, and Zhi-Hua Zhou. One-pass AUC optimization. In *Proceedings of the 30th International Conference on Machine Learning, PMLR 28*, pages 906–914. 2013.

Dan Garber, Ohad Shamir, and Nathan Srebro. Communication-efficient algorithms for distributed stochastic principal component analysis. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1203–1212. 2017.

Donald Goldfarb. A family of variable metric updates derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

Bo E Honoré and James L Powell. Pairwise difference estimators of censored and truncated regression models. *Journal of Econometrics*, 64(1-2):241–278, 1994.

Bo E Honoré and James L. Powell. Pairwise difference estimators for nonlinear models. In D. W. K. Andrews and J. H. Stock, editors, *Identification and Inference for Econometric Models, Essays in Honour of Thomas Rothenberg*, pages 520–553. Cambridge: New York, 1997.

Cheng Huang and Xiaoming Huo. A distributed one-step estimator. *Mathematical Programming*, 174:41–76, 2015.

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*, 76(4): 795–816, 2014.

Vladimir S Korolyuk and Yu V Borovskich. *Theory of U-Statistics*. Springer: Netherlands, 2013.

Léa Laporte, Rémi Flamary, Stéphane Canu, Sébastien Déjean, and Josiane Mothe. Nonconvex regularizations for feature selection in ranking with sparse SVM. *IEEE Transactions on Neural Networks and Learning Systems*, 25(6):1118–1130, 2013.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

Nan Lin and Ruibin Xi. Fast surrogates of U-statistics. *Computational Statistics & Data Analysis*, 54(1):16–24, 2010.

Guillaume Papa, Stéphan Clémençon, and Aurélien Bellet. Sgd algorithms based on incomplete u-statistics: large-scale minimization of empirical risk. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28 (NIPAS 2015)*, pages 1027–1035. MIT Press: Cambridge, U.S.A., 2015.

Steven L Scott, Alexander W Blocker, Fernando V Bonassi, Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayes and big data: The consensus monte carlo algorithm. *International Journal of Management Science & Engineering Management*, 11(2):78–88, 2016.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *Proceedings of the 31st International conference on Machine Learning, PMLR 32*, pages 1000–1008. 2014.

David F Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

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.

Virginia Smith, Simone Forte, Chenxin Ma, Martin Takáč, Michael I Jordan, and Martin Jaggi. CoCoA: A general framework for communication-efficient distributed optimization. *Journal of Machine Learning Research*, 18(1):8590–8638, 2017.

Xiao Song and Shuangge Ma. Penalised variable selection with U-estimates. *Journal of Nonparametric Statistics*, 22(4):499–515, 2010.

Marc A Suchard, Quanli Wang, Cliburn Chan, Jacob Frelinger, Andrew Cron, and Mike West. Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of Computational and Graphical Statistics*, 19(2):419–438, 2010.

Alexander Terenin, Daniel Simpson, and David Draper. Asynchronous gibbs sampling. In *Proceedings of the 23rd International Conference on Artifical Intelligence and Statistics (AISTATS), PMLR 108*. 2020.

Jiyuan Tu, Weidong Liu, Xiaojun Mao, and Xi Chen. Variance reduced median-of-means estimator for Byzantine-robust distributed inference. *Journal of Machine Learning Research*, 22(1):3780–3846, 2021.

Sara van de Geer. *Empirical Processes in M-Estimation*. Cambridge: New York, 2000.

Aad W van der Vaart. *Asymptotic Statistics*. Cambridge: New York, 1998.

Robin Vogel, Aurélien Bellet, Stephan Clémençon, Ons Jelassi, and Guillaume Papa. Tradeoffs in large-scale distributed tuplewise estimation and learning. In U. Brefeld, E. Fromont E., A. Hotho, A. Knobbe, M. Maathuis, and C. Robardet, editors, *Machine Learning*

and Knowledge Discovery in Databases. ECML PKDD 2019. Lecture Notes in Computer Science, Vol. 11907, pages 229–245. 2019.

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

Xiangyu Wang and David B Dunson. Parallelizing MCMC via weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

Xiaozhou Wang, Zhuoyi Yang, Xi Chen, and Weidong Liu. Distributed inference for linear support vector machine. *Journal of Machine Learning Research*, 20:1–41, 2019.

Xueqin Wang, Xin Dang, Hanxiang Peng, and Heping Zhang. The Theil-Sen estimators in multiple linear regression models. *http://home.olemiss.edu/ xdang/papers/MTSE.pdf*, 2009.

Yan Wang, Nathan Palmer, Qian Di, Joel Schwartz, Isaac Kohane, and Tianxi Cai. A fast divide-and-conquer sparse cox regression. *Biostatistics*, 22:381–401, 2021.

Ruibin Xi and Nan Lin. Direct regression modelling of high-order moments in big data. *Statistics & Its Interface*, 9(4):445–452, 2016.

Yiming Ying and Ding-Xuan Zhou. Online pairwise learning algorithms. *Neural Computation*, 28(4):743–777, 2016.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(1):3321–3363, 2013.