Distributed Statistical Inference under Heterogeneity

Jia Gu GUJIA@PKU.EDU.CN
Center for Statistical Science
Peking University
Bejing, China

Song Xi Chen SONGXICHEN@PKU.EDU.CN
School of Mathematical Science and Guanghua School of Management
Peking University
Beijing, China

Editor:

Abstract

We consider distributed statistical optimization and inference in the presence of heterogeneity among distributed data blocks. A weighted distributed estimator is proposed to improve the statistical efficiency of the standard "split-and-conquer" estimator for the common parameter shared by all the data blocks. The weighted distributed estimator is at least as efficient as the would-be full sample and the generalized method of moment estimators with the latter two estimators requiring full data access. A bias reduction is formulated for the weighted distributed estimator to accommodate much larger numbers of data blocks (relax the constraint from $K = o(N^{1/2})$ to $K = o(N^{2/3})$, where K is the number of blocks and N is the total sample size) than the existing methods without sacrificing the statistical efficiency at the same time. The mean square error bounds, the asymptotic distributions, and the corresponding statistical inference procedures of the weighted distributed and the debiased estimators are derived, which shows an advantageous performance of the debiased weighted estimators when the number of data blocks is large.

Keywords: Bias Correction; Distributed Inference; Heterogeneity; Split and Conquer Method; Weighted Estimation.

1. Introduction

Modern big data have brought new challenges to statistical inference. One such challenge is that despite the sheer volume of the data, full communication among the data points may not be possible due to either the cost of data communication or the privacy concern. The distributed or the "split-and-conquer" method has been proposed to divide the full data sample into smaller size data blocks to avoid data communication. The split and conquer estimator is also suited to situations where the data are naturally divided into data blocks and data communication among the data blocks are prohibited due to privacy concern. The "split and conquer" estimation had been considered in Lin and Xi (2010) for the U-statistics, Zhang et al. (2013) for the statistical optimization, Chen and Xie (2014) for the generalized linear models, Volgushev et al. (2017) and Chen et al. (2019) for the quantile regression, Battey et al. (2018) for high dimensional testing and estimation, and Chen and Peng (2021) for asymptotic symmetric statistics (Lai and Wang, 1993). Bootstrap resampling-based methods had been introduced to facilitate statistical inference. Kleiner et al. (2011) proposed the bag-of-little bootstrap (BLB) method for the plug-in estimators by making up economically the full sample for the distributed inference. Sengupta et al. (2015) suggested a sub-sampled double bootstrap method designed to improve the computational efficiency of the BLB. Chen and Peng (2021) proposed the distributed and the pseudo-distributed bootstrap methods with the former conducting the resampling within each data block while the latter directly resampling the distributed statistics.

Privacy has been a major concern in big data applications where people are naturally reluctant to share the raw data to form a pool of big data as practised in the traditional full sample estimation. However, the data holders may like to contribute summary statistics without having to give away the full data information. Federated Learning or the distributed inference with a central host has been proposed to accommodate such reality (McMahan et al., 2017; Yang et al., 2019; Li et al., 2020; Kairouz et al., 2021), where summary statistics of the data blocks or the gradients of the objective functions associated with the private data blocks are submitted to a central host for forming aggregated estimation or computation.

Homogeneous distribution among the data blocks is assumed in the majority of the statistical distributed inference studies with a few exceptions (Zhao et al., 2014; Duan et al., 2021). Federated Learning, on the other hand, was introduced to mitigate challenges arising from classical distributed optimization. In particular, heterogeneous or non-IID distributed data across different data blocks is one of the defining characteristics in the Federated Learning (Li et al., 2020; Kairouz et al., 2021). Indeed, it is natural to expect the existence of heterogeneity, especially for data stored in different locations or generated by different stochastic mechanisms, for instance, mobile phones of different users. But few works have focused on the asymptotic statistical properties of the estimator, especially in a heterogeneous setting.

Main Contributions. This paper considers distributed statistical inference under heterogeneous distributions among the data blocks, where there is a common parameter shared by the distributions of the data blocks and data-block-specific heterogeneous parameters. It is noted that Duan et al. (2021) also considered a heterogeneous setting but under a fully parametric framework. Specifically, the main contributions of this paper are as follows:

- Our study reveals that in the presence of heterogeneity the full sample estimator of the common parameter obtained by requiring full data access, can be less efficient than the split and conquer estimator. It is found that this phenomenon disappears if the objective function of the statistical optimization satisfies a generalized second-order Bartlett's identity.

- We propose a weighted distributed (WD) estimator, which is asymptotically at least as efficient as the full sample and the split and conquer estimator when the number of data blocks $K = o(N^{1/2})$ where N is the local sample size. The mean square error bound and the asymptotic distribution of the proposed weighted distributed estimator are derived, as well as the asymptotic equivalence between the weighted distributed and the generalized method of moment estimator Hansen (1982).

- We also propose a debiased weighted distributed estimator with a data splitting mechanism on each data block to remove the dependency between the bias correction and the weights used to tackle the heterogeneity. The debiased weighted distributed estimator is asymptotically as efficient as the WD estimator but allows quicker growth for the number of blocks $K = o(N^{2/3})$. The bias correction is also applied to the split and conquer formulation leading to a more communication-efficient debiased split and conquer estimator.

2. Preliminaries

Suppose that there is a large data sample of size N, which is divided into K data blocks of sizes $\{n_k\}_{k=1}^K$ such that $N = \sum_{k=1}^K n_k$ and let $n = NK^{-1}$ be the average sample size of the data blocks. For the relative sample sizes among data blocks, we assume the following.

Assumption 1 *There exist $c, C > 0$ such that $c \leq n_{k_1}/n_{k_2} \leq C$ for all pairs of (k_1, k_2).*

The k-th data block consists of a sub-sample $\{X_{k,i}\}_{i=1}^{n_k}$ which are independent and identically distributed (IID) random vectors from a probability space $(\Omega, \mathcal{F}, P)$ to $(\mathbb{R}^d, \mathcal{R}^d)$ with F_k as the distribution. The K distributions $\{F_k\}$ share a common parameter $\phi \in \mathbb{R}^{p_1}$, while each F_k has another parameter $\lambda_k \in \mathbb{R}^{p_2}$ specific to F_k. The parameters of interests in the k-th block are $\theta_k = (\phi^T, \lambda_k^T)^T$, and the overall parameters of interests are $\theta = (\phi^T, \lambda_1^T, \lambda_2^T, ..., \lambda_K^T)^T \in \mathbb{R}^{p_1+Kp_2}$.

Suppose there is a common objective function $M(X; \phi, \lambda_k)$ that is convex with respect to the parameter (ϕ, λ_k) and facilitates the statistical optimization in each data block. In general, the criteria function can be made block specific, say M_k function. Indeed, the presence of the heterogeneous local parameters $\{\lambda_k\}_{k=1}^K$ leads to different $M_k(x, \phi) = M(x, \phi, \lambda_k)$ for the inference on ϕ, which connects to the multi-task learning.

In the k-th data block the true parameter $\theta_k^* = (\phi^{*T}, \lambda_k^{*T})^T$ is defined as the unique minimum of the expected objective function, namely $\theta_k^* = argmin_{\theta_k \in \Theta_k} E_{F_k}(M(X_{k,1}; \phi, \lambda_k))$. The true common parameter ϕ^* appears in all θ_k^*, and the block-specific $\{\lambda_k^*\}_{k=1}^K$ may differ from each other. The entire true parameters $\theta^* = (\phi^{*T}, \lambda_1^{*T}, \cdots, \lambda_K^{*T})^T$, can be also identified as $\theta^* = argmin_{\theta \in \Theta} \sum_{k=1}^K \gamma_k E_{F_k}(M(X_{k,1}; \phi, \lambda_k))$. If the data could be shared across the data blocks, we would attain the conventional full sample estimator $\hat{\theta}_{full} = argmin_{\theta \in \Theta} \sum_{k=1}^K \sum_{i=1}^{n_k} M(X_{k,i}; \phi, \lambda_k)$, which serves as a benchmark for distributed estimators. The estimating equations for the full sample estimators are

$$\begin{cases} \sum_{k=1}^K \sum_{i=1}^{n_k} \psi_\phi(X_{k,i}; \phi, \lambda_k) = 0, \\ \sum_{i=1}^{n_k} \psi_\lambda(X_{k,i}; \phi, \lambda_k) = 0 \quad k = 1, ..., K, \end{cases}$$

where $\psi_\phi(X_{k,i}; \phi, \lambda_k) = \partial M(X_{k,i}; \phi, \lambda_k)/\partial \phi$ and $\psi_\lambda(X_{k,i}; \phi, \lambda_k) = \partial M(X_{k,i}; \phi, \lambda_k)/\partial \lambda_k$ are the score functions. The above full sample estimation is not attainable for the distributed situations due to privacy or the costs associated with the data communications. The distributed estimation first conducts local estimation on each data block, namely the local estimator $\hat{\theta}_k = (\hat{\phi}_k, \hat{\lambda}_k) = argmin_{\theta_k \in \Theta_k} \sum_{i=1}^{n_k} M(X_{k,i}; \theta_k)$ with the corresponding estimating equations

$$\begin{cases} \sum_{i=1}^{n_k} \psi_\phi(X_{k,i}; \phi_k, \lambda_k) = 0, \\ \sum_{i=1}^{n_k} \psi_\lambda(X_{k,i}; \phi_k, \lambda_k) = 0. \end{cases} \tag{1}$$

The split and conquer estimator for the common parameter ϕ is

$$\hat{\phi}^{SaC} = \frac{1}{N} \sum_{k=1}^K n_k \hat{\phi}_k. \tag{2}$$

The heterogeneity among the distributions of the data blocks call for study the relative efficiency and the estimation errors, which are the focus of this paper. We are to show that the split and conquer estimator (2) may not be the best formulation for estimating ϕ. Throughout this paper, unless otherwise stated, $\|\cdot\|_2$ represents the L_2 norm of a vector and a matrix. We will use C and C_i to denote absolute positive constants independent of (n_k, K, N).

An important question is the efficiency and the estimation errors of the split and conquer estimator $\hat{\phi}^{SaC}$ relative to the full sample estimator $\hat{\phi}_{full}$. For the homogeneous case, Chen and Peng (2021) found that for the asymptotic symmetric statistics, the split and conquer estimator (2) attains the same efficiency as the full sample estimator in the non-degenerate case but encounters an efficiency loss in the degenerate case due to a lack of communications among different data blocks. Zhang et al. (2013) derived the mean square error bound for the split and conquer estimator in the homogeneous case and showed that whenever $K \le N^{1/2}$, the split and conquer estimator achieves the best possible rate of convergence when all N data are accessible.

Consider the estimating equations of the full sample statistical optimization

$$\Psi_N(X;\theta) = \begin{pmatrix} \sum_{k=1}^{K}\sum_{i=1}^{n_k}\psi_\phi(X_{k,i};\phi,\lambda_k) \\ \sum_{i=1}^{n_1}\psi_\lambda(X_{1,i};\phi,\lambda_1) \\ \vdots \\ \sum_{i=1}^{n_K}\psi_\lambda(X_{K,i};\phi,\lambda_K) \end{pmatrix}. \tag{3}$$

Let $\Psi_\theta(\theta_k) = E\left(\nabla_{\theta_k} M(X_{k,1};\theta_k)\right)$ and $\Psi_\theta^\theta(\theta_k) = E(\nabla_{\theta_k}^2 M(X_{k,1};\theta_k))$ be the first and second order gradients of the k-th population objective function, whose matrix forms are:

$$\Psi_\theta(\theta_k) = (\Psi_\phi(\theta_k)^T, \Psi_\lambda(\theta_k)^T)^T, \quad \Psi_\theta^\theta(\theta_k) = \begin{pmatrix} \Psi_\phi^\phi(\theta_k) & \Psi_\phi^\lambda(\theta_k) \\ \Psi_\lambda^\phi(\theta_k) & \Psi_\lambda^\lambda(\theta_k) \end{pmatrix}.$$

Let $J_{\phi|\lambda}(\theta_k) = \Psi_\phi^\phi(\theta_k) - \Psi_\phi^\lambda(\theta_k)\Psi_\lambda^\lambda(\theta_k)^{-1}\Psi_\lambda^\phi(\theta_k)$, $J_{\lambda|\phi}(\theta_k) = \Psi_\lambda^\lambda(\theta_k) - \Psi_\lambda^\phi(\theta_k)\Psi_\phi^\phi(\theta_k)^{-1}\Psi_\phi^\lambda(\theta_k)$, $S_\phi(X_{k,i};\theta_k) = \psi_\phi(X_{k,i};\theta_k) - \Psi_\phi^\lambda(\theta_k)\Psi_\lambda^\lambda(\theta_k)^{-1}\psi_\lambda(X_{k,i};\theta_k)$ and $S_\lambda(X_{k,i};\theta_k) = \psi_\lambda(X_{k,i};\theta_k) - \Psi_\lambda^\phi(\theta_k)\Psi_\phi^\phi(\theta_k)^{-1}\psi_\phi(X_{k,i};\theta_k)$. Then, apply Taylor's expansion to obtain (see Section A.1)

$$\hat{\phi}_{full} - \phi^* = -\{\sum_{k=1}^{K}(n_k/N)J_{\phi|\lambda}(\theta_k^*)\}^{-1}N^{-1}\big\{\sum_{k=1}^{K}\sum_{i=1}^{n_k}S_\phi(X_{k,i};\theta_k^*)\big\} + o_p(N^{-1/2}). \tag{4}$$

For the local estimator $(\hat{\phi}_k, \hat{\lambda}_k)$ that solves (1), the same derivation leads to

$$\begin{cases} \hat{\phi}_k - \phi^* &= -n_k^{-1}J_{\phi|\lambda}(\theta_k^*)^{-1}\sum_{i=1}^{n_k}S_\phi(X_{k,i};\theta_k^*) + o_p(n_k^{-1/2}), \\ \hat{\lambda}_k - \lambda_k^* &= -n_k^{-1}J_{\lambda|\phi}(\theta_k^*)^{-1}\sum_{i=1}^{n_k}S_\lambda(X_{k,i};\theta_k^*) + o_p(n_k^{-1/2}), \end{cases}$$

Our analysis requires the following conditions.

Assumption 2 (Identifiability) *The parameters $\theta_k^* = (\phi^*, \lambda_k^*)$ is the unique minimizer of $M_k(\theta_k) = E(M(X_{k,1};\theta_k))$ for $\theta_k \in \Theta_k$.*

Assumption 3 (Compactness) *The true parameter θ_k^* is an interior point of the parameter space Θ_k which is a compact and convex set in $\mathbb{R}^p$, and $sup_{\theta_k\in\Theta_k}\|\theta_k - \theta_k^*\|_2 \le r$ for all $k \ge 1$ and some $r > 0$. The true common parameter ϕ^* is an interior point of a subset $\Phi \subset \Theta_k$.*

Assumption 4 (Local strong convexity) *The population objective function on the k-th data block $M_k(\theta_k) = E(M(X_{k,1};\theta_k))$ is twice differentiable, and there exists a constant $\rho_- > 0$ such that $\nabla_{\theta_k}^2 M_k(\theta_k^*) \succeq \rho_- I_{p\times p}$. Here $A \succeq B$ means $A - B$ is a positive semi-definite matrix.*

Assumption 5 (Smoothness, 1) *The objective function on the k-th data is twice differentiable with respect to θ_k and there are positive constants R, L, v and v_1 such that $E(\|\nabla_{\theta_k} M(X_{k,1};\theta_k^*)\|_2^{2v_1}) \leq R^{2v_1}$ and $E(\|\nabla^2_{\theta_k} M(X_{k,1};\theta_k^*) - \nabla^2_{\theta_k} M_k(\theta_k^*)\|_2^{2v}) \leq L^{2v}$ for all $k \geq 1$. There are also positive constants ρ and G such that $\|\nabla^2_{\theta_k} M(x;\theta_k) - \nabla^2_{\theta_k} M(x;\theta_k')\|_2 \leq G(x)\|\theta_k - \theta_k'\|_2$ for all $\theta_k, \theta_k' \in U_k = \{\theta_k \mid \|\theta_k - \theta_k^*\|_2 \leq \rho\}$ and $x \in \mathbb{R}^d$, and $E(G(X_{k,1})^{2v}) \leq G^{2v}$.*

Assumptions 2-4 are standard ones on the parameter space and population objective functions for the homogeneous case (Jordan et al., 2019). In the heterogeneous case, Duan et al. (2021) requires the parameter space for the common parameter to be bounded, i.e. $\|\phi - \phi^*\| \leq r$ under a fully parametric setting, while we need the overall parameter space to be bounded. The stronger condition is needed since we do not fully specify the distributions $\{F_k\}_{k=1}^K$ and it will be used when we derive the mean squared error bound for the proposed weighted distributed estimator in Section 4. Assumption 5 specifies the Lipschitz continuity of the outer product $Z(x;\theta_k)$ with respect to θ_k, which is to control the estimation error when we estimate the asymptotic covariance matrix of the local estimator $\hat{\theta}_k$. Section A.2 shows it is valid for the logistic regression model.

3. Full Sample versus split and conquer Estimation

It is expected that the full sample estimator $\hat{\phi}_{full}$ should be at least as efficient as $\hat{\phi}^{SaC}$ since the former utilizes the full sample information allowing communications among data blocks. However, we show that this is not necessarily the case under heterogeneity.

It is worth mentioning that we assume K being fixed in the following Proposition 1 and Theorem 2 for simplicity of formulating the asymptotic variance of the estimators, which helps us to motivate the weighted distributed estimator. We allow diverging K along with N in the subsequent theoretical results. In particular, we will discuss how to improve the divergence rate of K in Section 5.

Proposition 1 *Under Assumptions 1 - 4 and Assumption 5 with $v, v_1 \geq 1$, and if K is fixed, then $\hat{\theta}_k \to \theta_k^*$ and $\hat{\theta}_{full} \to \theta^*$ in probability; $\hat{\phi}^{SaC} = (1/N)\sum_{k=1}^K n_k \hat{\phi}_k$ and $\hat{\phi}_{full}$ are consistent to ϕ^*.*

Theorem 2 *Under Assumptions 1 - 4 and Assumption 5 with $v, v_1 \geq 2$, if K is fixed and $n_k/N \to \gamma_k \in (0,1)$ for a set of constants $\{\gamma_k\}_{k=1}^K$, then*

$$\sqrt{N}(\hat{\phi}^{SaC} - \phi^*) \to \mathcal{N}(0_{p_1}, \sum_{k=1}^K \gamma_k J_{\phi|\lambda}(\theta_k^*)^{-1}\Sigma_k(\theta_k^*)J_{\phi|\lambda}(\theta_k^*)^{-1}) \quad \text{and}$$

$$\sqrt{N}(\hat{\phi}_{full} - \phi^*) \to \mathcal{N}(0_{p_1}, (\sum_{k=1}^K \gamma_k J_{\phi|\lambda}(\theta_k^*))^{-1}(\sum_{k=1}^K \gamma_k \Sigma_k(\theta_k^*))(\sum_{k=1}^K \gamma_k J_{\phi|\lambda}(\theta_k^*))^{-1}),$$

where $\Sigma_k = Var\{S_\phi(X_{k,1};\theta_k^*)\}$.

Define $V(\Sigma, A) = (A^T)^{-1}\Sigma A^{-1}$ as a mapping from $\mathbb{S}_{++}^{p_1\times p_1} \times GL(\mathbb{R}^{p_1})$ to $\mathbb{S}_{++}^{p_1\times p_1}$, where $\mathbb{S}_{++}^{p_1\times p_1}$ and $GL(\mathbb{R}^{p_1})$ denote the symmetric positive definite matrices and invertible real matrices of order p_1, respectively. In fact, $V(\cdot,\cdot)$ is a non-convex function, which means that

$$(\sum_{k=1}^K \gamma_k J_{\phi|\lambda}(\theta_k^*))^{-1}(\sum_{k=1}^K \gamma_k \Sigma_k(\theta_k^*))(\sum_{k=1}^K \gamma_k J_{\phi|\lambda}(\theta_k^*))^{-1} \not\preceq \sum_{k=1}^K \gamma_k J_{\phi|\lambda}(\theta_k^*)^{-1}\Sigma_k(\theta_k^*)J_{\phi|\lambda}(\theta_k^*)^{-1}.$$

In other words, $\hat{\phi}_{full}$ is not necessarily more efficient than $\hat{\phi}^{SaC}$.

To gain understanding of Theorem 2, we consider the errors-in-variables model. Suppose there are K independent data blocks $\{(X_{k,i}, Y_{k,i})\}_{i=1}^{n}$ for $k = 1, 2..., K$, where $(X_{k,i}, Y_{k,i})$ are IID and generated from

$$X_k = Z_k + e_k, \quad Y_k = \phi^* + \lambda_k^* Z_k + f_k, \tag{6}$$

where $\{Z_k\}_{k=1}^{K}$ are random variables whose measurements $\{(X_k, Y_k)\}_{k=1}^{K}$ are subject to errors $\{(e_k, f_k)\}_{k=1}^{K}$, and (e_k, f_k) are bivariate normally distributed with zero mean and covariance matrix $\sigma^2 I_2$ and is independent of Z. Obviously, ϕ^* is the common parameter across all data blocks while $\lambda_k^*(\lambda_k^* > 0)$ represents the block specific parameter. The condition $Var(e) = Var(f)$ is assumed to avoid any identification issue arising when Z is also normally distributed (Reiersol, 1950). We consider the approach in Example 5.26 of van der Vaart (1999) as detailed in Section A.3, which leads to the M-function

$$M(X_k, Y_k; \theta_k) = \frac{1}{2\sigma^2(1+\lambda_k^2)}(\lambda_k X_k - (Y_k - \phi))^2. \tag{7}$$

For simplicity we assume $K = 2$, then from Theorem 2 we have

$$\begin{cases} \text{var}(\hat{\phi}_{full}) \approx \left(\frac{\sigma^2 E(Z^2)}{\text{var}(Z)} \frac{2}{\frac{1}{1+\lambda_1^{*2}} + \frac{1}{1+\lambda_2^{*2}}} + \frac{\sigma^4 (E(Z))^2}{\text{var}^2(Z)} \frac{\frac{2}{(1+\lambda_1^{*2})^2} + \frac{2}{(1+\lambda_2^{*2})^2}}{(\frac{1}{1+\lambda_1^{*2}} + \frac{1}{1+\lambda_2^{*2}})^2} \right) \frac{1}{N}, \\ \text{var}(\hat{\phi}^{SaC}) \approx \left(\frac{\sigma^2 E(Z^2)}{\text{var}(Z)} \frac{(1+\lambda_1^{*2}) + (1+\lambda_2^{*2})}{2} + \frac{\sigma^4 (E(Z))^2}{\text{var}^2(Z)} \right) \frac{1}{N}. \end{cases} \tag{8}$$

In the heterogeneous setting ($\lambda_1^* \neq \lambda_2^*$), cases are presented in Section C.2 where $\hat{\phi}_{full}$ has larger variance than the split and conquer estimator $\hat{\phi}^{SaC}$.

4. Weighted Distributed Estimator

That the full sample estimator $\hat{\phi}_{full}$ under heterogeneity may be less efficient than the simple averaged $\hat{\phi}^{SaC}$ suggests that the wisdom formulated in the homogeneous context may not be applicable to the heterogeneous case. How to better aggregate the local estimators $\{\hat{\phi}_k\}$ for more efficient estimation is the focus of this section.

4.1 Formulation and Results

Consider a class of estimators formed by linear combinations of the local estimators $\{\hat{\phi}_k\}$:

$$\{\hat{\phi}_w^{SaC} \mid \hat{\phi}_w^{SaC} = \sum_{k=1}^{K} W_k \hat{\phi}_k, W_k \in \mathbb{R}^{p_1 \times p_1}, \sum_{k=1}^{K} W_k = I_{p_1}\}.$$

We want to minimize the asymptotic variance of $\hat{\phi}_w^{SaC}$ with respect to weighting matrices $\{W_k\}_{k=1}^{K}$. It may be shown from Theorem 2 that var$(\hat{\phi}_w^{SaC}) \approx \sum_{k=1}^{K} n_k^{-1} W_k A_k^{-1} \Sigma_k (A_k^T)^{-1} W_k^T$, where $A_k = J_{\phi|\lambda}(\theta_k^*)$ and $\Sigma_k = \text{var}\left(S_\phi(X_{k,i}; \theta_k^*)\right)$. It is noted that the asymptotic variance is defined via the asymptotic normality of the statistical optimization. For the time being, A_k and Σ_k are assumed known and denote $H_k = A_k^{-1} \Sigma_k (A_k^T)^{-1}$. We choose the trace operator as a measure of the size of the covariance matrix, which leads to a minimization problem:

$$\underset{W_k}{Minimize} \quad tr\left(\sum_{k=1}^{K} n_k^{-1} W_k H_k W_k^T\right) \quad s.t. \quad \sum_{k=1}^{K} W_k = I_{p_1}. \tag{9}$$

It is a convex optimization problem and can be solved via the Lagrangian multiplier method, which gives the optimal weighting matrices $W_k^* = (\sum_{s=1}^K n_s H_s^{-1})^{-1} n_k H_k^{-1}$. If we replace the trace with the Frobenius norm in (9), the same solution is attained as shown in Section A.4. The split and conquer estimator with the optimal weights W_k^* is called the weighted distributed estimator and denoted as $\hat{\phi}^{WD}$, which is at least as efficient as $\hat{\phi}^{SaC}$ by construction.

To compare the efficiency between $\hat{\phi}_{full}$ and $\hat{\phi}^{WD}$, we note that their covariances

$$\begin{aligned} \text{var}(\hat{\phi}_{full}) &\approx \{(\sum_{k=1}^{K} n_k A_k)^T (\sum_{k=1}^{K} n_k \Sigma_k)^{-1} (\sum_{k=1}^{K} n_k A_k)\}^{-1} \quad \text{and} \\ \text{var}(\hat{\phi}^{WD}) &\approx \left(\sum_{k=1}^{K} n_k A_k^T \Sigma_k^{-1} A_k\right)^{-1}, \quad \text{respectively.} \end{aligned} \tag{10}$$

Define $F(\Sigma, A) = A^T \Sigma^{-1} A$, which is a generalized convex function with respect to the matrix inequality shown in Lemma S1. Applying Jensen's inequality leads to the that the weighed distributed estimator is at least as efficient as the full sample estimator $\hat{\phi}_{full}$. Thus, the estimating equations (3) obtained from the first-order derivative of the simple summation of local objectives $\sum_{i=1}^{n_k} M(X_{k,i}; \theta_k)$ may not be the best formulation. In contrast, the weighted distributed estimator exploits the potential efficiency gain from the heterogeneity by re-weighting of the local estimators, which is why the full sample estimator may not be as efficient as the weighted distributed estimator.

4.2 Likelihood and Quasi-likelihood

The above results lead us to wonder whether the weighted distributed estimator can also be more efficient than the full sample estimator under the heterogeneity in a fully parametric setting. The answer is negative as shown below.

When the distribution of $X_{k,i}$ is fully parametric with density function $f(\cdot; \phi, \lambda_k)$, the Fisher information matrix in the k-th data block is

$$I(\theta_k) = I(\phi, \lambda_k) = \begin{pmatrix} I_{\phi\phi} & I_{\phi\lambda_k} \\ I_{\lambda_k\phi} & I_{\lambda_k\lambda_k} \end{pmatrix} = -E\begin{pmatrix} \frac{\partial^2}{\partial\phi^2} logf(X_{k,1}; \theta_k) & \frac{\partial^2}{\partial\phi\partial\lambda^T} logf(X_{k,1}; \theta_k) \\ \frac{\partial^2}{\partial\lambda\partial\phi^T} logf(X_{k,1}; \theta_k) & \frac{\partial^2}{\partial\lambda^2} logf(X_{k,1}; \theta_k) \end{pmatrix},$$

and the partial information matrix $I_{\phi|\lambda_k} = I_{\phi\phi} - I_{\phi\lambda_k} I_{\lambda_k\lambda_k}^{-1} I_{\lambda_k\phi}$.

Now, the objective function for the statistical optimization is $M(X_{k,i}; \phi, \lambda_k) = -\log f(X_{k,i}; \phi, \lambda_k)$. Routine derivations show that $\Sigma_k = \text{var}\left(S_\phi(X_{k,1}; \theta_k^*)\right) = I_{\phi|\lambda_k}$ and $A_k = J_{\phi|\lambda}(\theta_k^*) = I_{\phi|\lambda_k}$. Hence, $\text{var}(\hat{\phi}_{full}) \approx \text{var}(\hat{\phi}^{WD}) \approx \left(\sum_{k=1}^K n_k I_{\phi|\lambda_k}\right)^{-1}$ and $\text{var}(\hat{\phi}^{SaC}) \approx (1/N^2) \sum_{k=1}^K n_k I_{\phi|\lambda_k}^{-1}$. A direct application of Lemma S1 shows that $\left(\sum_{k=1}^K n_k I_{\phi|\lambda_k}\right)^{-1} \preceq (1/N^2) \sum_{k=1}^K n_k I_{\phi|\lambda_k}^{-1}$. Thus, the full sample maximum likelihood estimator automatically adjusts for the heterogeneity and has the same asymptotic efficiency as that of the weighted distributed estimator. Both estimators are at least as efficient as the split and conquer estimator $\hat{\phi}^{SaC}$. The same is true for the quasi-likelihood estimation with independent observations (see Section A.5).

A close examination reveals that the underlying reason for the asymptotic equivalence between the weighted distributed estimator and the likelihood-based full sample estimators is that the two statistical optimization functions satisfy the second order Bartlett's identity (Bartlett, 1953; McCullagh, 1983): $E(\nabla M(X_k, \theta_k^*)\nabla M(X_k, \theta_k^*)^T) = E(\nabla^2 M(X_k, \theta_k^*))$. By the asymptotic variance formula of the estimator and Lemma S1, it is apparent that Bartlett's identity can be relaxed by allowing a factor $\gamma \neq 0$ such that

$$E(\nabla M(X_k, \theta_k^*)\nabla^T M(X_k, \theta_k^*)) = \gamma E(\nabla^2 M(X_k, \theta_k^*)). \tag{11}$$

An example of such cases is the least square estimation in the parametric regression with homoscedastic and non-autocorrelated residuals in Section A.6. Otherwise, the full sample least square estimator may not be efficient and there is an opportunity for the weighted least square estimation. Thus, if $M(x_k, \theta_k)$ satisfies (11), $\hat{\phi}_{full}$ attains the same efficiency as $\hat{\phi}^{WD}$.

4.3 Relation to Generalized Method of Moment Estimation

To further justify the efficiency of the weighed distributed estimation, we consider the generalized method of moment (GMM) estimator (Hansen, 1982), which has certain optimal property for the semiparametric inference that the weighted distributed estimation can compare with, despite it requires full data sharing.

The score functions of the statistical optimization on each data block are aggregated to form the moment equations

$$\begin{cases} \sum_{i=1}^{n_k} \psi_\phi(X_{k,i}; \phi, \lambda_k) = 0, \\ \sum_{i=1}^{n_k} \psi_\lambda(X_{k,i}; \phi, \lambda_k) = 0, \quad k = 1, ..., K, \end{cases} \tag{12}$$

which have pK estimating equations, where the dimension of θ^* is $pK - (K-1)p_1$. Thus, the parameter is over-identified which offers potential efficiency gain for the generalized method of moment. The GMM estimation based on the moment restrictions (12) solves the minimization problem $\hat{\theta}_{GMM} = argmin\ \tilde{\psi}_N^T(\theta) W_0 \tilde{\psi}_N(\theta)$, where $W_0 = \{\text{var}(\tilde{\psi}_N(\theta^*))\}^{-1}$ is the optimal weighting matrix (Hansen, 1982; Yaron et al., 1996) and $\tilde{\psi}_N(\theta) = (\sum_{i=1}^{n_1} \psi_\phi(X_{1,i}; \theta_1)^T, \sum_{i=1}^{n_1} \psi_\lambda(X_{1,i}; \theta_1)^T, \cdots, \sum_{i=1}^{n_K} \psi_\phi(X_{K,i}; \theta_K)^T, \sum_{i=1}^{n_K} \psi_\lambda(X_{K,i}; \theta_K)^T)^T$.

Let the first p_1 elements of $\hat{\theta}_{GMM}$ be $\hat{\phi}_{GMM}$ as an estimator of the common parameter. A derivation in Section A.7 shows that $\text{var}(\hat{\phi}_{GMM}) \approx \{\sum_{k=1}^K n_k J_{\phi|\lambda} \Sigma_k^{-1} J_{\phi|\lambda}\}^{-1}$. Thus, the weighted distributed estimator's efficiency is the same as that of $\hat{\phi}_{GMM}$. This is encouraging as the weighted distributed estimator does it without requiring data sharing among the blocks.

4.4 Estimation of weights in one round communication

To formulate the weighed distributed estimator, we have to estimate the optimal weights $W_k^* = (\sum_{s=1}^K n_s H_s^{-1})^{-1} n_k H_k^{-1}$. As we will show in Theorem 4, the estimation of the weights will not affect the estimation efficiency of the weighted distributed estimator attained in (10). By the structure of W_k^*, we only need to estimate H_k, the leading principal submatrix of order p_1 of the asymptotic covariance matrix $\widetilde{H}_k$ of $\hat{\theta}_k$. Note that

$$\widetilde{H}_k = (\nabla \Psi_\theta(\theta_k^*))^{-1} E\{\psi_{\theta_k}(X_{k,1}; \theta_k^*) \psi_{\theta_k}(X_{k,1}; \theta_k^*)^T\} (\nabla \Psi_\theta(\theta_k^*))^{-1} = \begin{pmatrix} H_k & * \\ * & * \end{pmatrix},$$

where $\Psi_\theta(\theta_k) = E\psi_{\theta_k}(X_{k,1};\theta_k)$. We can construct a sandwich type estimator (Stefanski and Boos, 2002) to estimate $\widetilde{H}_k$ and then H_k. The procedure to obtain the weighted distributed estimator is summarized in Algorithm 1.

Input: Distributed datasets: $\{X_{k,i}, k = 1, ..., K; i = 1, ..., n_k\}$
Output: Weighted distributed estimator: $\hat{\phi}^{WD}$
1 In each data block k ($k = 1, 2, \cdots, K$):
2 Solve (1) and obtain $\hat{\theta}_k = (\hat{\phi}_k, \hat{\lambda}_k)$;
3 Calculate $\widehat{H}_k(\hat{\theta}_k)$, which is the leading principal sub-matrix of order p_1 of $(\nabla_{\theta_k}\widehat{\Psi}_{\theta_k})^{-1}(n_k^{-1}\sum_{i=1}^{n_k} Z(X_{k,i};\hat{\theta}_k))(\nabla_{\theta_k}\widehat{\Psi}_{\theta_k})^{-T}$, where $Z(x, \theta_k)$ is defined in Assumption 5 and $\widehat{\Psi}_{\theta_k} = n_k^{-1}\sum_{i=1}^{n_k}\psi_{\theta_k}(X_{k,i};\hat{\theta}_k)$;
4 In a central server:
5 Collect $(\hat{\phi}_k, \widehat{H}_k(\hat{\theta}_k)^{-1})$ from all the K data blocks;
6 Calculate $\hat{\phi} = (\sum_{k=1}^K n_k\widehat{H}_k(\hat{\theta}_k)^{-1})^{-1}\sum_{k=1}^K n_k(\widehat{H}_k(\hat{\theta}_k))^{-1}\hat{\phi}_k$;
7 $\hat{\phi}^{WD} = \hat{\phi}I(\hat{\phi} \in \Phi) + \hat{\phi}^{SaC}I(\hat{\phi} \notin \Phi)$, where $\hat{\phi}^{SaC} = N^{-1}\sum_{k=1}^K n_k\hat{\phi}_k$.

Algorithm 1: Weighted Distributed estimator

Step 7 of the algorithm is necessary since there is no guarantee that after weighting the estimator $\tilde{\hat{\phi}}^{WD}$ belongs to the set Φ as required in Assumption 3. However, the event $\{\tilde{\hat{\phi}}^{WD} \in \Phi\}$ should happen with probability approaching one. Hence, the $\hat{\phi}^{SaC}I(\tilde{\hat{\phi}}^{WD} \notin \Phi)$ term is negligible. To establish the theoretical properties of the weighted distributed estimator, we impose the following assumptions.

Assumption 6 (Smoothness, 2) *Denote $Z(x, \theta_k) = \nabla_{\theta_k}M(x;\theta_k)\nabla_{\theta_k}M(x;\theta_k)^T$, then there are positive constants ρ and B such that $Z(x, \theta_k)$ is $B(x)-$Lipschitz continuous with respect to θ_k, in the sense that $\|Z(x, \theta_k) - Z(x, \theta_k')\|_2 \le B(x)\|\theta_k - \theta_k'\|_2$ for all $\theta_k, \theta_k' \in U_k = \{\theta_k \mid \|\theta_k - \theta_k^*\|_2 \le \rho\}$ and $x \in \mathbb{R}^d$, and $E(B(X_{k,1})^{2v}) \le B^{2v}$.*

Assumption 7 (Boundedness) *Denote $\Sigma_{S,k}(\theta_k) = E_{F_k}(\psi_{\theta_k}(X_{k,1};\theta_k)\psi_{\theta_k}(X_{k,1};\theta_k)^T)$, then there exists constants $\rho_\sigma, c > 0$ such that $\|\Sigma_{S,k}(\theta_k^*)\|_2 \le \rho_\sigma$ and $H_k \succeq cI_{p_1\times p_1}$ for $k \ge 1$, where θ_k^* is the minimizer of the k-th population objective function and $H_k = J_{\phi|\lambda}(\theta_k^*)^{-1}var(S_\phi(X_{k,1};\theta_K^*))J_{\phi|\lambda}(\theta_k^*)^{-1}$.*

By H_k's definition, $\|H_k\|_2 \le \|\Psi_\theta^\theta(\theta_k^*)^{-1}\|_2^2\|\Sigma_{S,k}(\theta_k^*)\|_2 \le \rho_\sigma\rho_-^{-2}$, implying $H_k^{-1} \succeq (\rho_-^2/\rho_\sigma)I_{p_1\times p_1}$. On the other hand, the above inequality leads to $\|\Psi_\theta^\theta(\theta_k^*)^{-1}\|_2 \ge (c/\rho_\sigma)^{1/2}$, which indicates a finite upper bound for the norm of the Hessian, as assumed in Jordan et al. (2019) and Duan et al. (2021).

Theorem 3 *Under Assumptions 1 - 4 and 7, and Assumption 5 - 6 with $v, v_1 \ge 2$, the mean-squared error of the weighed distributed estimator $\hat{\phi}^{WD}$ satisfies*

$$E\left(\|\hat{\phi}^{WD} - \phi^*\|_2^2\right) \le \frac{C_1}{nK} + \frac{C_2}{n^2} + \frac{C_3}{n^2K} + \frac{C_4}{n^3} + \frac{C_5K}{n^{\bar{v}}},$$

for $n = NK^{-1}$ and $\bar{v} = min\{v, v_1/2\}$.

The v and v_1 appeared in Assumptions 5 - 6 quantify the moments of the first two orders of the gradients of the M-function and their corresponding Lipschitz functions. When the number of data blocks $K = \mathcal{O}(n^{min\{1,(\bar{v}-1)/2\}})$, the convergence rate of mean squared error of $\hat{\phi}^{WD}$ is $\mathcal{O}((nK)^{-1})$, which is the same as the standard full sample estimator. However, when there are too many data blocks such that $K \gg n$, the convergence rate is reduced to $\mathcal{O}(n^{-2})$. Furthermore, if the derivatives of the M function and their corresponding Lipschitz functions are heavy-tailed so that $\bar{v} < 3$, the convergence rate is further reduced to $\mathcal{O}(Kn^{-\bar{v}})$.

Theorem 4 *Under Assumptions 1 - 4 and 7, and Assumptions 5 - 6 with $v, v_1 \geq 2$, if $K = o(N^{1/2})$, then $(\hat{\phi}^{WD} - \phi^*)^T \left(\sum_{k=1}^K n_k H_k^{-1}\right) (\hat{\phi}^{WD} - \phi^*) \to \chi^2_{p_1}$.*

As mentioned before, K is allowed to diverge with the full sample size at the rate $o(N^{1/2})$. Although $\{H_k\}_{k=1}^K$ have bounded spectral norms, $\sum_{k=1}^K (n_k/N) H_k^{-1}$ may not converge to a fixed matrix in the presence of heterogeneity. Thus, we can only obtain the asymptotic normality of the standardized $N^{-1/2}\{\sum_{k=1}^K (n_k/N) H_k^{-1}\}^{1/2}(\hat{\phi}^{DW} - \phi^*)$. This is why Theorem 4 is presented in the limiting chi-squared form, which implies that we can construct confidence regions for ϕ with confidence level $1 - \alpha$ as

$$\{\phi \mid (\hat{\phi}^{WD} - \phi)^T \left(\sum_{k=1}^K n_k \widehat{H}_k(\hat{\theta}_k)^{-1}\right) (\hat{\phi}^{WD} - \phi) \leq \chi^2_{p_1,\alpha}\}$$

after replacing $\sum_{k=1}^K n_k H_k^{-1}$ with its sample counterpart $\sum_{k=1}^K n_k \widehat{H}_k(\hat{\theta}_k)^{-1}$, where $\chi^2_{p_1,\alpha}$ is the upper α quantile of the $\chi^2_{p_1}$ distribution. Given the weighted distributed estimator of the common parameter ϕ^*, a natural question is that whether a more efficient estimator of the block-specific λ_k^* can be obtained, if we plug in the weighed distributed estimator to each data block and re-estimate λ_k. Let $\hat{\lambda}_k^{(2)}$ be the updated estimator. Results in Section A.8 show that $\hat{\lambda}_k^{(2)}$ is not necessarily more efficient than the local estimator $\hat{\lambda}_k$.

5. Debiased Estimator for diverging K

It is noted that $K = o(N^{1/2})$ is required in Theorems 3 and 4 to attain the $\mathcal{O}(N^{-1})$ leading order mean square error and the limiting chi-squared distribution of the weighed distributed estimator $\hat{\phi}^{WD}$. A reason for this requirement is that the bias of the local estimator $\hat{\theta}_k$ is at order $O_p(n_k^{-1})$, which can not be reduced by the weighted averaging. This leads to the bias of $N^{1/2}(\hat{\phi}^{WD} - \phi^*)$ being at the order $O_p(KN^{-1/2})$, which is not necessarily diminishing to zero unless $K = o(N^{1/2})$. It is worth mentioning that Duan et al. (2021) needed the same $K = o(N^{1/2})$ order in their maximum likelihood estimation framework to obtain the $N^{1/2}$-convergence since Li et al. (2003) showed that the maximum likelihood estimator is asymptotically biased when $K/n \to C \in (0, +\infty)$. This calls for a bias reduction step for the local estimators before aggregation to allow for larger K.

To facilitate the bias correction, we have to simplify the notation. Suppose $F(\theta)$ is a $p \times 1$ vector function, $\nabla F(\theta)$ is the usual Jacobian whose l-th row contains the partial derivatives of the l-th element of $F(\theta)$. Then, the matrices of higher derivatives are defined recursively so that the j-th element of the l-th row of $\nabla^s L(\theta)$ (a $p \times p^s$ matrix) is the $1 \times p$ vector $f_{lj}^v(\theta) = \partial f_{lj}^{v-1}(\theta)/\partial\theta^T$, where f_{lj}^{v-1} is the $l-$th row and j-th element of $\nabla^{v-1} F(\theta)$. Let $\otimes$ denote the Kronecker product. Using Kronecker product we can express $\nabla^v F(\theta) = \partial^v F(\theta)/(\partial\theta^T \otimes \partial\theta^T \otimes \cdots \otimes \partial\theta^T)$. Besides, define $M_{n,k}(\theta_k) = n_k^{-1} \sum_{i=1}^{n_k} M(X_{k,i}; \theta_k)$, $H_{3,k}(\theta_k) = E(\nabla^2_{\theta_k} \psi_{\theta_k}(X_{k,1}; \theta_k))$, $Q_k(\theta_k) =$

$\{-E(\nabla_{\theta_k}\psi_{\theta_k}(X_{k,1};\theta_k))\}^{-1}$, $d_{i,k}(\theta_k) = Q_k(\theta_k)\psi_{\theta_k}(X_{k,i};\theta_k)$ and $v_{i,k}(\theta_k) = \nabla_{\theta_k}\psi_{\theta_k}(X_{k,i},\theta_k) - \nabla_{\theta_k}\Psi_\theta(\theta_k)$. Then, the leading order bias of $\hat{\theta}_k$ (Rilstone et al., 1996) is

$$Bias(\hat{\theta}_k) = n_k^{-1}Q_k(\theta_k^*)\{E\left(v_{i,k}(\theta_k^*)d_{i,k}(\theta_k^*)\right) + \frac{1}{2}H_{3,k}(\theta_k^*)E\left(d_{i,k}(\theta_k^*)\otimes d_{i,k}(\theta_k^*)\right)\}.$$

Let $B_k(\theta_k) = Q_k(\theta_k)\{E\left(v_{i,k}(\theta_k)d_{i,k}(\theta_k)\right) + \frac{1}{2}H_{3,k}(\theta_k)E\left(d_{i,k}(\theta_k)\otimes d_{i,k}(\theta_k)\right)\}$, whose first p_1 dimension associated with ϕ are denoted as $B_k^1(\theta_k)$. An estimator of $B_k(\theta_k)$ is

$$\hat{B}_k(\theta_k) = \hat{Q}_k(\theta_k)\big(n_k^{-1}\sum_{i=1}^{n_k}\hat{v}_{i,k}(\theta_k)\hat{d}_{i,k}(\theta_k) + \frac{1}{2}\hat{H}_{3,k}(\theta_k)n_k^{-1}\sum_{i=1}^{n_k}(\hat{d}_{i,k}(\theta_k)\otimes\hat{d}_{i,k}(\theta_k))\big), \tag{13}$$

where $\hat{H}_{3,k}(\theta_k) = n_k^{-1}\sum_{i=1}^{n_k}\nabla^2_{\theta_k}\psi_{\theta_k}(X_{k,i};\theta_k)$, $\hat{Q}_k(\theta_k) = \{-n_k^{-1}\sum_{i=1}^{n_k}\nabla_{\theta_k}\psi_{\theta_k}(X_{k,i};\theta_k)\}^{-1}$, $\hat{d}_{i,k}(\theta_k) = \hat{Q}_k(\theta_k)\psi_{\theta_k}(X_{k,i};\theta_k)$ and $\hat{v}_{i,k}(\theta_k) = \nabla_{\theta_k}\psi_{\theta_k}(X_{k,i};\theta_k)$. Applying it to each data block, we have the bias-corrected local estimator

$$\hat{\theta}_{k,bc} = \hat{\theta}_k - n_k^{-1}\hat{B}_k(\hat{\theta}_k)1_{\mathcal{E}_{k,bc}}$$

where $\mathcal{E}_{k,bc} = \{\hat{\theta}_k - n_k^{-1}\hat{B}_k(\hat{\theta}_k)\in\Theta_k\}$, and the indicator function is to ensure that $\hat{\theta}_{k,bc}\in\Theta_k$.

After the local debiased estimators are obtained, we need to aggregate them via the estimated weights. A direct aggregation will invalidate the bias correction due to the dependence between the estimated weights and the local debiased estimator if they are constructed with the same dataset. The accumulation of dependence over a large number of data blocks can make the bias correction fail. To remove the dependence between the local estimator $\hat{\theta}_{k,bc}$ and the estimated local weights $\hat{W}_k = \{\sum_{s=1}^K \hat{H}_s(\hat{\theta}_s)^{-1}\}^{-1}\widehat{H}_k(\hat{\theta}_k)^{-1}$, we divide each local dataset $\{X_{k,i}\}_{i=1}^{n_k}$ to two basically equal-sized splits $D_k^s = \{X_{k,i}^{(s)}\}_{i=1}^{n_k/2}$, $s=1,2$. For $s=1,2$, we calculate the local estimators $\hat{\theta}_{k,s}$ and obtain $\hat{H}_{k,s}(\hat{\theta}_{k,s})$, which is the first p_1 principal sub-matrix of

$$(\nabla_{\theta_k}\widehat{\Psi}_{\theta_k})^{-1}((2/n_k)\sum_{i=1}^{n_k/2}\psi_{\theta_k}(X_{k,i}^{(s)};\hat{\theta}_{k,s})\psi_{\theta_k}(X_{k,i}^{(s)};\hat{\theta}_{k,s})^T)(\nabla_{\theta_k}\widehat{\Psi}_{\theta_k})^{-T},$$

where $\widehat{\Psi}_{\theta_k} = (2/n_k)\sum_{i=1}^{n_k/2}\psi_{\theta_k}(X_{k,i}^{(s)};\hat{\theta}_{k,s})$. We perform the local bias correction to $\hat{\theta}_{k,s}$ based on a split with the weight obtained by the other, leading to two debiased estimators of the form

$$\{\sum_{k=1}^K n_k\hat{H}_{k,s}(\hat{\theta}_{k,s})^{-1}\}^{-1}\sum_{k=1}^K n_k(\hat{H}_{k,s}(\hat{\theta}_{k,s}))^{-1}\hat{\phi}^{bc}_{k,2-|s-1|} \quad \text{for} \quad s=1,2.$$

The two debiased local estimators are averaged to obtain the final debiased weighed distributed estimator, whose procedure is summarized in Algorithm 2. To provide a theoretical guarantee on the bias correction, we need an assumption on the third-order gradient of the M-function (see Zhang et al. (2013)), which strengthens a part of Assumption 5.

Assumption 8 (Strong smoothness) *For each $x\in\mathbb{R}^p$, the third order derivatives of $M(x;\theta_k)$ with respect to θ_k exist and are $A(x)-$ Lipschitz continuous in the sense that*

$$\|(\nabla^2_{\theta_k}\psi_{\theta_k}(x;\theta_k) - \nabla^2_{\theta_k}\psi_{\theta_k}(x;\theta_k^{'}))(u\otimes u)\|_2 \le A(x)\|\theta_k-\theta_k^{'}\|_2\|u\|_2^2,$$

for all $\theta_k,\theta_k^{'}\in U_k$ defined in Assumption 5 and $u\in\mathbb{R}^p$, where $E(A(X_{k,i})^{2v})\le A^{2v}$ for some $v>0$ and $A<\infty$.

Input: Distributed datasets: $\{X_{k,i}, k = 1, ..., K; i = 1, ..., n_k\}$
Output: debiased weighted distributed estimator: $\hat{\phi}^{dWD}$

In each data block k ($k = 1, 2, \cdots, K$):
  Split the local dataset into two equal sized subsets: $D_k^s = \{X_{k,i}^{(s)}\}_{i=1}^{n_k/2}, s = 1, 2$;
  Solve (1) based on D_k^s and obtain $\hat{\theta}_{k,s} = (\hat{\phi}_{k,s}, \hat{\lambda}_{k,s})$ for $s = 1, 2$;
  Calculate $\widehat{H}_{k,s}(\hat{\theta}_{k,s})$ based on D_k^s and $\hat{\theta}_k^s$ for $s = 1, 2$;
  Calculate $\hat{\theta}_{k,s}^{bc} = \hat{\theta}_{k,s} - 2n_k^{-1}\hat{B}_{k,s}(\hat{\theta}_{k,s})1_{\mathcal{E}_{k,bc,s}}$ using formula (13) for $s = 1, 2$, where $\mathcal{E}_{k,bc,s} = \{\hat{\theta}_{k,s} - 2n_k^{-1}\hat{B}_{k,s}(\hat{\theta}_{k,s}) \in \Theta_k\}$;
In a central server:
  Collect $\{\hat{\phi}_{k,s}^{bc}, \widehat{H}_{k,s}(\hat{\theta}_{k,s})^{-1}, s = 1, 2\}$ from all the K data blocks;
  Construct $\hat{\phi}^s = (\sum_{k=1}^K n_k \widehat{H}_{k,s}(\hat{\theta}_{k,s})^{-1})^{-1} \sum_{k=1}^K n_k \widehat{H}_{k,s}(\hat{\theta}_{k,s})^{-1} \hat{\phi}_{k,2-|s-1|}^{bc}$;
  Calculate $\hat{\phi}_s^{dWD} = \hat{\phi}^s I(\hat{\phi}^s \in \Phi) + K^{-1}\sum_{k=1}^K n_k \hat{\phi}_{k,2-|s-1|}^{bc} I(\hat{\phi}^s \notin \Phi)$ for $s = 1, 2$;
  $\hat{\phi}^{dWD} = 2^{-1}\sum_{s=1}^2 \hat{\phi}_s^{dWD}$.

Algorithm 2: debiased Weighted Distributed (dWD) Estimator

Theorem 5 *Under Assumptions 1 - 4 and 7 - 8, and Assumptions 5 - 6 with $v, v_1 \geq 4$,*

$$E(\|\hat{\phi}^{dWD} - \phi^*\|_2^2) \leq \frac{C_1}{nK} + \frac{C_2}{n^2K} + \frac{C_3}{n^3} + \frac{C_4K}{n^{min\{v, v_1/2\}}}.$$

The main difference between the upper bounds in Theorem 5 from that in Theorem 3 for the weighed distributed estimator is the disappearance of the $\mathcal{O}(n^{-2})$ term for the weighed distributed estimator, which has been absorbed into the $\mathcal{O}((n^2K)^{-1} + n^{-3})$ terms for the debiased weighed distributed estimator. As shown next, this translates to a more relaxed $K = o(N^{2/3})$ condition as compared with the $K = o(N^{1/2})$ condition for the weighed distributed estimator in Theorem 4.

Theorem 6 *Under the conditions required by Theorem 5, if $K = o(N^{2/3})$,*

$$(\hat{\phi}^{dWD} - \phi^*)^T \left(\sum_{k=1}^K n_k H_k(\theta_k^*)^{-1}\right)(\hat{\phi}^{dWD} - \phi^*) \xrightarrow{d} \chi_{p_1}^2.$$

Theorem 6 is also formulated in the chi-squared distribution form for the same reason when we formulate Theorem 4, and similar confidence region with confidence level $1 - \alpha$ can be constructed as $\{\phi \mid (\hat{\phi}^{dWD} - \phi)^T\{\sum_{k=1}^K n_k H_k(\hat{\theta}_k)^{-1}\}(\hat{\phi}^{dWD} - \phi) \leq \chi_{p_1,\alpha}^2\}$.

The fact that the confidence regions of debiased weighted distributed and weighted distributed estimators use the same standardizing matrix $\sum_{k=1}^K n_k \widehat{H}_k(\hat{\theta}_k)^{-1}$ reflects that both estimators have the same estimation efficiency. However, the debiased version has more relaxed constraint on $K = o(N^{2/3})$ than that of the WD estimator at $K = o(N^{1/2})$.

Both the debiased and non-debiased weighted distributed estimators are communication efficient as they only require one round of communication. When the communication budget is strictly limited, people may only share the debiased estimators without transmitting the weights. In this case, one may consider the following debiased split and conquer estimator

$$\hat{\phi}^{dSaC} = N^{-1}\sum_{k=1}^K n_k(\hat{\phi}_k - n_k^{-1}\hat{B}_k^1(\hat{\theta}_k)1_{\mathcal{E}_{k,bc}}),$$

which only performs bias correction and may be preferable when the heterogeneity is not severe. The asymptotic property of $\hat{\phi}^{dSaC}$ is summarized in the following theorem.

Theorem 7 *Under the conditions required by Theorem 5, if $K = o(N^{2/3})$, the debiased split and conquer estimator $\hat{\phi}^{dSaC}$ satisfies that (i) $E\left(\|\hat{\phi}^{dSaC} - \phi^*\|_2^2\right) \leq C_1/(nK) + C_2/(n^2K) + C_3/n^3$ and (ii) $N^2(\hat{\phi}^{dSaC} - \phi^*)^T \left(\sum_{k=1}^K n_k H_k(\theta_k^*)\right)^{-1} (\hat{\phi}^{dSaC} - \phi^*) \to \chi^2_{p_1}$.*

The corresponding confidence region with confidence level $1-\alpha$ can be constructed as $\{\phi \mid N^2\big(\hat{\phi}^{dSaC} - \phi\big)^T \left(\sum_{k=1}^K n_k \widehat{H}_k(\hat{\theta}_k)\right)^{-1} \big(\hat{\phi}^{dSaC} - \phi\big) \leq \chi^2_{p_1,\alpha}\}$. It is noted that the debiased version of the split and conquer estimator $\hat{\phi}^{dSaC}$ has the same asymptotic distribution as that of $\hat{\phi}^{SaC}$, but under a much more relaxed constraint on the divergence rate of K. Hence, the confidence regions based on the split and conquer estimator can be constructed in the same way as that based on the weighted distributed estimator with $\hat{\phi}^{dSaC}$ replaced by $\hat{\phi}^{SaC}$.

To compare with the subsampled average mixture method (SAVGM) estimator proposed in Zhang et al. (2013), which also performs local bias correction but under the homogeneous setting, we have the following corollary to Theorem 7.

Corollary 8 *Under the homogeneous case such that $\{X_{k,i}, k = 1, ..., K, i = 1, ..., n;\}$ are IID distributed, and the assumptions required by Theorem 5,*

$$E\left(\|\hat{\theta}^{dSaC} - \theta_1^*\|_2^2\right) \leq \frac{2E\left(\|\nabla_{\theta_1}\Psi_\theta(\theta_1^*)^{-1}\psi_{\theta_1}(X_{1,1};\theta_1^*)\|_2^2\right)}{nK} + \frac{C_1}{n^2K} + \frac{C_2}{n^3},$$

where θ_1^* is the true parameter for all the K data blocks.

The SAVGM estimator resamples $\lfloor rn_k \rfloor$ data points from each data block k for a $r \in (0,1)$ to obtain a local estimator $\hat{\theta}_{k,r}^{SaC}$ based on the sub-samples, and has the form

$$\bar{\theta}_{SAVGM} = \frac{\hat{\theta}_k^{SaC} - r\hat{\theta}_{k,r}^{SaC}}{1-r}. \tag{14}$$

Its mean squred error bound as given in Theorem 4 of Zhang et al. (2013) is

$$E\left(\|\bar{\theta}_{SAVGM} - \theta_1^*\|_2^2\right) \leq \frac{2+3r}{(1-r)^2}\frac{E\left(\|\nabla_{\theta_1}\Psi_\theta(\theta_1^*)^{-1}\psi_{\theta_1}(X_{1,1};\theta_1^*)\|_2^2\right)}{nK} + \frac{C_1}{n^2K} + \frac{C_2}{n^3}. \tag{15}$$

Thus, the mean squared error bound (15) of the SAVGM estimator has an inflated factor $(2+3r)(1-r)^{-2}/2 > 1$ for $r \in (0,1)$ when compared with that of the dSaC estimator, although it is computationally more efficient than the debiased split and conquer and debiased weighted distributed estimators as it only draws one subsample in its resampling. For more comparisons between the debiased split and conquer estimator and one-step estimators proposed by Huang and Huo (2019), see Section A.10.

6. Numerical Results

6.1 Simulation study

We report results from simulation experiments designed to verify the theoretical findings made in the previous sections, which was to evaluate the numerical performance of the proposed weighted distributed (WD), debaised split and conquer (dSaC) and debiased weighted distributed (dWD) estimators of the common parameter and compare them with the existing split and conquer (SaC) and subsampled average mixture method (SAVGM) (with subsampling rate $r = 0.05$) estimators. Although Zhang's SAVGM estimator (Zhang et al., 2013) was proposed under the homogeneous setting, but since its main bias correction is performed locally on each data block k as shown in (14), similar theoretical bounds as (15) can be derived without much modifications on the original proof. Throughout the simulation experiments, the results of each simulation setting were based on $B = 500$ number of replications and were conducted in R with a 10-core Intel(R) Core(TM) i9-10900K @3.7 GHz processor. We evaluated the numerical performance of the five estimators for the common parameter ϕ under a logistic regression model. For each of K data block with $K \in \{10, 50, 100, 250, 500, 1000, 2000\}$, $\{(X_{k,i}; Y_{k,i})\}_{i=1}^{n} \subset \mathbb{R}^p \times \{0, 1\}$ were independently sampled from the following model:

$$X_{k,i} \sim \mathcal{N}(0_{p\times 1}, 0.75^2 I_{p\times p}) \quad \text{and} \quad P(Y_{k,i} = 1 \mid X_{k,i}) = \frac{exp(X_{k,i}^T \theta_k^*)}{1 + exp(X_{k,i}^T \theta_k^*)},$$

where $\theta_k^* = (\phi^{*T}, \lambda_k^{*T})^T$, $\phi^* = 1$, $\lambda_k^* = (\lambda_{k,1}^*, \lambda_{k,2}^*, \cdots, \lambda_{k,p_2}^*)^T$ and $\lambda_{k,j}^* = (-1)^j 10(1 - 2(k - 1)/(K - 1))$. The sample sizes of the data blocks were equal at $n = NK^{-1}$ with $N = 2 \times 10^6$. Two levels of the dimension $p_2 = 4$ and 10 of the nuisance parameter λ_k were considered. Due to space limit, we only report the set of result with $p_2 = 10$ in the main paper. See Section A.9 for the result with $p_2 = 4$ and a derivation of the bias correction formula for the logistic model.

Figure 1 reports the root mean square errors and absolute bias of the estimators when $p_2 = 10$. It is observed that the weighted distributed estimator and the two debiased estimators had smaller root mean square errors than those of the SaC and SAVGM for almost all the simulation settings. The classical split and conquer estimator fared better than Zhang's SAVGM estimator as K became larger, which is due to the extra variation introduced by the subsampling method as indicated in (15), especially when K is large (the local sample size n is small). It was evident that the WD estimator had much smaller root mean square errors than the SaC and SAVGM estimators for all the block number K, realizing its theoretical promises. In most cases, the WD estimator had smaller bias than the SaC estimator although it was not debiased. The WD estimator was advantageous for $K \leq 250$. In comparison, both bias corrected dWD and dSaC were very effective in reducing the bias of the WD and SaC estimators, respectively, especially for larger K when the bias was more severe. The dWD attained the smallest root mean square error and the bias in all settings, suggesting the need for conducting both weighting and the bias correction in the distributed inference especially for large K. These empirical results were consistent with Theorems 3 and 5, namely the leading root mean square error term of the WD estimator changes from $\mathcal{O}((Kn)^{-1})$ to $\mathcal{O}(n^{-2})$ when K surpasses the local sample size n, while the leading term of the dWD is still $\mathcal{O}((nK)^{-1})$ until K is much larger than n^2.

We also evaluated the coverage probabilities and widths of the $1 - \alpha$ ($\alpha = 0.01, 0.05, 0.1$) confidence intervals (CIs) of the common parameter based on the asymptotic normality as given

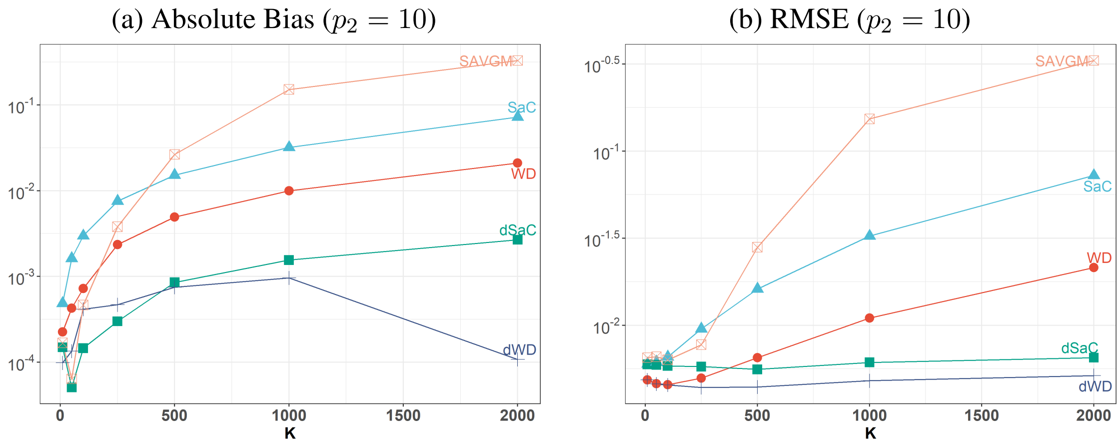


Figure 1: Average simulated bias (a) and the root mean square errors (RMSE) (b) of the weighted distributed (WD) (red circle), the split and conquer(SaC) (blue triangle), the debiased split and conquer (dSaC) (green square), the debiased weighted distributed (dWD) (purple cross), the subsampled average mixture SAVGM (pink square cross) estimators, with respect to the number of data block K for the logistic regression model with the dimension p_2 of the nuisance parameter λ_k being 10, and the full sample size $N = 2 \times 10^6$.

after Theorems 4 and 6. The SAVGM estimator was not included as its asymptotic distribution was not made available in Zhang et al. (2013). Table 1 reports the empirical coverage and the average width of the CIs. It is observed that the four types of the CIs all had quite adequate coverage levels when $K \leq 100$. However, for $K \geq 250$, the SaC CIs first started to lose coverage, followed by those of the WD, while the CIs of the dSaC and dWD estimators can hold up to the promised coverage for all cases of K. Although the dSaC CIs had comparable coverages with the dWD CIs, their widths were much wider than those of the dWD. This was largely due to the fact that the weighted averaging conducted in the weighted distributed estimation reduced the variation and hence the width of the CIs. The widths of the WD CIs were largely the same with those of the dWD, and yet the coverage levels of the dWD CIs were much more accurate indicating the importance of the bias correction as it shifted the CIs without inflating the width.

In addition to the simulation experiments on the statistical properties of the estimators, the computation efficiency of the estimators was also evaluated. Table 2 reports the average CPU time per simulation run based on 500 replications of the five estimators for a range of K of the nuisance parameter for the logistic regression model with the total sample size $N = 2 \times 10^6$ and $p_2 = 10$. The computation speed of the dSaC and dWD estimators were relatively slower than those of the SaC, WD and Zhang's SAVGM estimators. The WD estimator was quite fast, which means that the re-weighting used less computing time than the bias-reduction. In comparison, the dWD estimator was the slowest as a cost for attaining the best root mean square error among the five estimators in all settings. It is observed in Table 2 that the overall computation time for each estimator first decreased and then increased as K became larger. The decrease in time was because of the benefit of the distributed computation, while the increase was due to the increase in the number of optimization associated with the statistical optimization performed as K got larger. However, it

Table 1: Coverage probabilities and widths (in parentheses, multiplied by 100) of the $1-\alpha$ confidence intervals for the common parameter ϕ in the logistic regression model based on the asymptotic normality of the split and conquer (SaC), the weighted distributed (WD), the debiased split and conquer (dSaC) and the debiased weighted distributed (dWD) estimators with respect to the number of data blocks K. The dimension p_2 of the nuisance parameter λ_k is 10 and total sample size $N = 2 \times 10^6$

K		SaC			WD			dSaC			dWD	
	$1-\alpha$	0.99	0.95	0.90	0.99	0.95	0.90	0.99	0.95	0.90	0.99	0.95	0.90
10		0.99	0.94	0.88	1.00	0.96	0.92	1.00	0.94	0.88	1.00	0.96	0.92
		(3.05)	(2.32)	(1.95)	(2.41)	(1.84)	(1.54)	(3.05)	(2.32)	(1.95)	(2.42)	(1.84)	(1.54)
50		0.99	0.93	0.87	0.99	0.95	0.88	0.98	0.94	0.88	0.99	0.96	0.88
		(2.94)	(2.24)	(1.88)	(2.29)	(1.74)	(1.46)	(2.94)	(2.24)	(1.88)	(2.29)	(1.74)	(1.46)
100		0.97	0.89	0.84	0.97	0.93	0.87	0.98	0.95	0.90	0.98	0.94	0.89
		(2.93)	(2.23)	(1.87)	(2.28)	(1.74)	(1.46)	(2.93)	(2.23)	(1.87)	(2.29)	(1.74)	(1.46)
250		0.89	0.72	0.63	0.98	0.92	0.87	1.00	0.97	0.90	1.00	0.96	0.90
		(2.94)	(2.24)	(1.88)	(2.28)	(1.74)	(1.46)	(2.94)	(2.24)	(1.88)	(2.29)	(1.74)	(1.46)
500		0.51	0.28	0.18	0.93	0.81	0.70	0.99	0.94	0.90	0.98	0.94	0.88
		(2.97)	(2.26)	(1.90)	(2.29)	(1.74)	(1.46)	(2.97)	(2.26)	(1.90)	(2.30)	(1.75)	(1.47)
1000		0.00	0.00	0.00	0.66	0.37	0.28	0.99	0.95	0.90	0.99	0.96	0.89
		(3.04)	(2.31)	(1.94)	(2.30)	(1.75)	(1.47)	(3.04)	(2.31)	(1.94)	(2.34)	(1.78)	(1.49)
2000		0.00	0.00	0.00	0.02	0.00	0.00	0.99	0.96	0.90	0.99	0.93	0.87
		(3.22)	(2.45)	(2.06)	(2.34)	(1.78)	(1.49)	(3.22)	(2.45)	(2.06)	(2.40)	(1.82)	(1.53)

is worth mentioning that these results did not account for the potential time expenditure in data communication among different data blocks.

Table 2: Average CPU time for each replication based on $B = 500$ replications for the split and conquer (SaC), the Zhang's SAVGM, the weighted distributed (WD), the debiased split and conquer (dSaC) and the debiased weighted distributed (dWD) estimators for the logistic regression model with respect to K. The dimension p_2 of the nuisance parameter λ_k is 10 and total sample size $N = 2 \times 10^6$

K	SaC	SAVGM	WD	dSaC	dWD
10	34.60	35.19	43.84	50.47	55.35
50	20.13	20.18	24.16	29.99	33.69
100	15.60	16.20	17.74	23.63	24.47
250	10.77	12.61	11.88	18.22	20.39
500	11.55	14.50	12.56	18.80	23.73
1000	15.23	18.27	16.28	22.38	32.24
2000	23.42	27.99	24.62	30.43	48.05

6.2 Real data analysis

In this sub-section, we report results from an empirical analysis on an airline's on-time performance data to demonstrate the proposed weighted distributed estimation for massive data. We aim at quantifying the association between flight departure delay and a set of covariates, the arrival delay of the previous flight of the same plane, the seasonal effects, and the weather conditions with a logistic regression model, based on data from the top 10 busiest airports in the United States in 2007. The flight data are available from https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2009 and the weather data are obtained from https://cds.climate.copernicus.eu/. We segmented the full data of $N = 2412782$ according to the airports of departing flights and obtained 10 data segments. For each segment, we split it to data blocks of size $n = 5000$, while the residual data blocks were discarded, such that the total number of blocks $K = 479$.

We included seven covariates in the logistic regression: the arrival delay of the previous flight, the season (encoded by three dummy variables: spring (March-May), summer (June-August), autumn (September-November) with winter as the baseline, the near-surface air temperature and pressure, and the rain rate before the scheduled departure time. The coefficients of the three weather variables were treated as the common parameters while the remaining coefficients including the intercept were regarded as heterogeneous; see Section C.3 in the supplementary material for the justification. The estimated common parameters of the near-surface air pressure, temperature, and convective rain rate with 95% confidence intervals using the weighted distributed estimator and the split and conquer estimator are shown in Figure 2. Both methods successfully identified a significant association between the three weather variables and the departure delay of a flight. Besides, the weighted distributed estimator reduced the lengths of the confidence intervals of the estimated common parameters compared with the split and conquer method. In particular, the confidence interval of the rain parameter was shortened by 19.1%, while those of the other two common parameters were shorted by 2.2% (pressure) and 2.9% (temperature), which justified the statistical efficiency of the weighted distributed estimator.

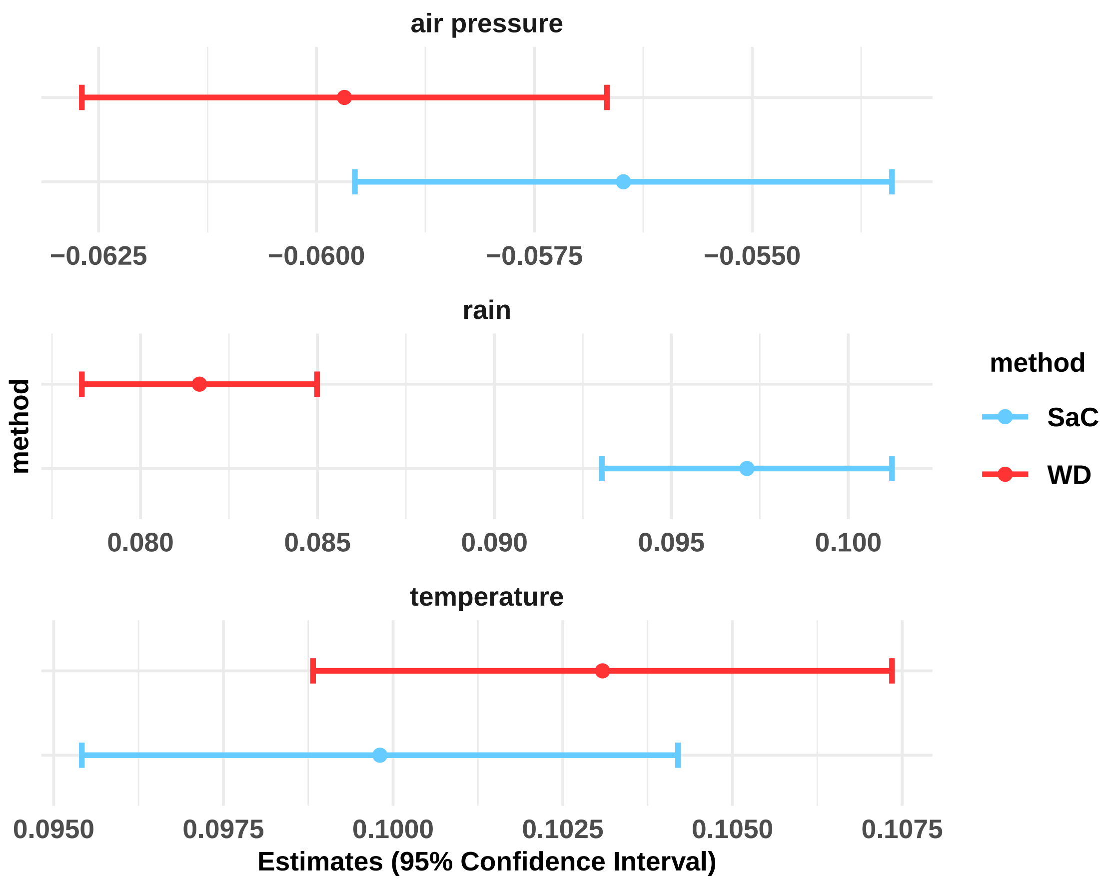
The data analysis demonstrated the feasibility of implementing the proposed weighted distributed estimation method for real-world distributed inference problems. With only one round of weighting to tackle the heterogeneity among the nuisance parameters, more efficient estimation can be obtained.

7. Discussion

This paper investigates distributed statistical optimization in the presence of heterogeneity in the data blocks. The weighted distributed estimator is able to improve the estimation efficiency of the split and conquer estimator for the common parameter. Two debiased estimators are proposed to allow for larger numbers of data blocks K. The statistical properties of the proposed estimators are shown to be advantageous over the split and conquer and SAVGM estimators. In particular, the weighted distributed estimator has good performance for smaller K relative to n, and the debiased weighted distributed estimator that conducted both bias correction and weighting offers good estimation accuracy for large K.

An important issue for the distributed estimation is the size of K relative to the full sample size N. This is especially true in a Federated Learning setting where the number of data blocks is usually very large. Both the split and conquer and weighted distributed estimators require $K = o(N^{1/2})$

Figure 2: Estimated common parameters of the near surface air pressure, temperature and convective rain rate with 95% confidence intervals using the weighted distributed estimator and the split and conquer estimator



to preserve the the $N^{1/2}$ rate for the asymptotic variance. The debiased weighted distributed and debiased split and conquer estimators relax the restriction to $K = o(N^{2/3})$ without sacrificing the convergence rate.

In machine learning, the multi-task learning (Smith et al., 2017) framework is a strategy to tackle the statistical heterogeneity in a distributed network, which fits separate local parameters $\{\phi_k\} \in \mathbb{R}^p$ to different data blocks (tasks) through convex loss functions $\{\ell_k(\cdot,\cdot)\}$ and is formulated as

$$\min_{\Phi,\Omega} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{n_k} \ell_k(\phi_k^T X_{k,i}, Y_{k,i}) + \mathcal{R}(\Phi, \Omega) \right\}, \tag{16}$$

where Φ is the matrix with $\{\phi_k\}_{k=1}^K$ as column vectors, $\Omega \in \mathbb{R}^{K\times K}$ and $\mathcal{R}(\cdot,\cdot)$ measures the extent of the heterogeneity among different data blocks. Choices of $\mathcal{R}(\cdot,\cdot)$ include the bi-convex function $\mathcal{R}(\Phi,\Omega) = \delta_1 tr(\Phi\Omega\Phi^T) + \delta_2\|\Phi\|_F^2$ for $\delta_1, \delta_2 > 0$ and $\Omega = I_{K\times K} - (1/K)1_K 1_K^T$ such that $tr(\Phi\Omega\Phi^T) = \sum_{k=1}^K \|\phi_k - \bar{\phi}_K\|_2^2$ where $\bar{\phi}_K = (1/K)\sum_{k=1}^K \phi_k$, which leads to the mean-regularized multi-task learning (Evgeniou and Pontil, 2004) with $\mathcal{R}$ conducting regularization on each local model.

The distributed framework in this paper is well connected to multi-task learning in two aspects. One is that despite we use the same objective (loss) function M over the data blocks, the heterogeneity induced by local parameters $\{\lambda_k\}$ and the distributions effectively define $M_k(\phi, x) = M(x, \phi, \lambda_k)$ is equivalent to the block specific loss functions ℓ_k used in (16). Another aspect is that although multi-task learning assumes different parameters $\{\phi_k\}$ over the data blocks, it regularizes them toward a common one. In contrast, we assume there is a common parameter ϕ shared by the distributions. By doing so, we are able to clarify the source of heterogeneity $\{\lambda_k\}$ and homogeneity ϕ instead of putting an equal treatment on all the dimensions of the parameter and focusing on the statistical inference of the common parameter.

Appendix

The Appendix is organized as follows. Section A provides derivations of the formulas given in the main text. Section B contains detailed proofs of the theoretical results. More simulation results and details about the real data analysis are reported in Section C.

A. Derivation of formulas

A.1 Expansion of the full sample estimator $\hat{\phi}_{full}$

By integral form of Taylor's expansion around the true value θ^*, we have

$$\begin{aligned} 0_{p\times 1} &= \Psi_N(X; \hat{\phi}_{full}, \hat{\lambda}_{1,full}, ..., \hat{\lambda}_{K,full}) \\ &= \Psi_N(X;\theta^*) + J(\theta^*)(\hat{\theta}_{full} - \theta^*) + (\nabla\Psi_N(X;\theta^*) - J(\theta^*))(\hat{\theta}_{full} - \theta^*) \\ &\quad + \{\int_0^1 \nabla\Psi_N(X;\theta^* + t(\hat{\theta}_{full} - \theta^*))(\hat{\theta}_{full} - \theta^*)dt - \nabla\Psi_N(X;\theta^*)\}(\hat{\theta}_{full} - \theta^*), \end{aligned}$$

where $J(\theta) = E\left(\nabla\Psi_N(X;\theta)\right)$. Then, inverting the above leads to

$$\hat{\theta}_{full} - \theta^* \quad = -J(\theta^*)^{-1}\Psi_N(X;\theta^*) + R_{N1} + R_{N2}, \tag{17}$$

where $R_{N1} = -J(\theta^*)^{-1}\{\nabla\Psi_N(X;\theta^*) - J(\theta^*)\}(\hat{\theta}_{full} - \theta^*)$ and $R_{N2} = -J(\theta^*)^{-1}\{\int_0^1 \nabla\Psi_N(X;\theta^* + t(\hat{\theta}_{full} - \theta^*))(\hat{\theta}_{full} - \theta^*)dt - \nabla\Psi_N(X;\theta^*)\}(\hat{\theta}_{full} - \theta^*)$ are both higher-order remainder terms. Since $J(\theta)$ has the following form

$$J(\theta) = \begin{pmatrix} \sum_{k=1}^K n_k\Psi_\phi^\phi(\theta_k) & n_1\Psi_\phi^\lambda(\theta_1) & \cdots & n_K\Psi_\phi^\lambda(\theta_K) \\ n_1\Psi_\lambda^\phi(\theta_1) & n_1\Psi_\lambda^\lambda(\theta_1) & \mathbf{0} & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ n_K\Psi_\lambda^\phi(\theta_K) & \mathbf{0} & \mathbf{0} & n_K\Psi_\lambda^\lambda(\theta_K), \end{pmatrix}, \tag{18}$$

then the right bottom part of $J(\theta)$ is a block diagonal matrix, whose inverse is at hand. Thus we can see $J(\theta)$ as a 2×2 block matrix and directly apply the block matrix inverse formula (Lu and Shiou, 2002). Thus from (17) we have $\hat{\phi}_{full} - \phi^* = -\{\sum_{k=1}^K (n_k/N)J_{\phi|\lambda}(\theta_k^*)\}^{-1}(1/N)\{\sum_{k=1}^K\sum_{i=1}^{n_k} S_\phi(X_{k,i};\theta_k^*)\} + o_p(N^{-1/2})$.

A.2 Lipschitz continuity of the outer product of the gradient in logistic regression model

First we define the logit function $logit(a) = exp(a)/(1 + exp(a))$ for $a \in \mathbb{R}$. Then the logistic regression model can be defined as $P(Y = 1|X) = logit(X^T\beta^*)$, where $X, \beta^* \in \mathbb{R}^p$. If we define the objective M as $M(z, \beta) = -ylog(logit(x^T\beta)) + (y-1)log(1-logit(x^T\beta))$, where $z = (y, x)$, then the outer product of gradient, denoted as $f(z, \beta)$, is $f(z, \beta) = (y - logit(x^T\beta))^2xx^T$. Now we have

$$\begin{aligned}
& \|f(z,\beta_1) - f(z,\beta_2)\|_2 \\
= & \|xx^T(2y - logit(x^T\beta_1) - logit(x^T\beta_2))(logit(x^T\beta_1) - logit(x^T\beta_2))\|_2 \\
= & \|xx^T(2y - logit(x^T\beta_1) - logit(x^T\beta_2))(1 - logit(\xi))logit(\xi)x(\beta_1 - \beta_2)\|_2 \\
\leq & \|x\|_2^3\|\beta_1 - \beta_2\|_2,
\end{aligned}$$

where the second equality comes from an application of the mean value theorem.

A.3 Errors-in-variables model

We first give a derivation of the objective function from the perspective of statistical optimization. As we will see, the derived objective is exactly the same as that when we do orthogonal regression or "Deming's regression" (Carroll and Ruppert, 1996). Consider the conditional likelihood of $(X_{k,i}, Y_{k,i})$ given $Z_{k,i}$ in block k

$$\begin{aligned}
& f(\{X_{k,i}\},\{Y_{k,i}\}|\{Z_{k,i}\},\theta_k) = \prod_{i=1}^n f_1(X_{k,i}|Z_{k,i})f_2(Y_{k,i}|Z_{k,i}) \\
=& (\frac{1}{2\pi\sigma^2})^n \prod_{i=1}^n exp\{-\frac{1}{2\sigma^2}\left[(X_{k,i}^2 + (Y_{k,i}-\phi)^2) - 2Z_{k,i}(X_{k,i} + \lambda_k(Y_{k,i}-\phi)) + (1+\lambda_k^2)Z_{k,i}^2\right]\}.
\end{aligned}$$

By the factorization theorem, $X_{k,i} + \lambda_k(Y_{k,i} - \phi)$ is a sufficient statistic for $Z_{k,i}$ if $\theta_k = (\phi, \lambda_k)$ is assumed to be known. And $X_{k,i} + 2\lambda_k(Y_{k,i} - \phi)|Z_{k,i} \sim \mathcal{N}((1 + \lambda_k^2)Z_{k,i}, (1 + \lambda_k^2)\sigma^2)$. Then, the above conditional likelihood can be factorized as

$$\begin{aligned}
& f(\{X_{k,i}\},\{Y_{k,i}\}|\{Z_{k,i}\},\theta_k) \\
= & (\frac{\sqrt{1+\lambda_k^2}}{\sqrt{2\pi}\sigma})^n \prod_{i=1}^n exp\{-\frac{1}{2\sigma^2(1+\lambda_k^2)}(\lambda_k X_{k,i} - (Y_{k,i}-\phi))^2\}h(X_{k,i} + \lambda_k(Y_{k,i}-\phi)|Z_{k,i}),
\end{aligned}$$

where $h(s_i|z_i)$ is the conditional density of $\mathcal{N}((1 + \lambda_k^2)z_i, (1 + \lambda_k^2)\sigma^2)$. Since $\{Z_{k,i}\}_{i=1}^n$ are not observable, we discard the factor h and construct the estimator based on the first part of the factorization, which is denoted as $\tilde{f}(\{X_{k,i}\}, \{Y_{k,i}\}|\{Z_{k,i}\}, \theta_k)$. Differentiate $\log \tilde{f}(\{X_{k,i}\}, \{Y_{k,i}\}|\{Z_{k,i}\}, \theta_k)$ with respect to $\theta_k = (\phi, \lambda_k)^T$, we obtain

$$\begin{cases}
\frac{\partial}{\partial\phi}log\tilde{f}(\{X_{k,i}\},\{Y_{k,i}\}|\{Z_{k,i}\},\theta_k) = -\frac{1}{\sigma^2(1+\lambda_k^2)}\sum_{i=1}^n(\lambda_k X_{k,i} - (Y_{k,i}-\phi)), \\
\frac{\partial}{\partial\lambda_k}log\tilde{f}(\{X_{k,i}\},\{Y_{k,i}\}|\{Z_{k,i}\},\theta_k) = n\frac{\lambda_k}{1+\lambda_k^2} + \sum_{i=1}^n \frac{\lambda_k}{\sigma^2(1+\lambda_k^2)^2}(\lambda_k X_{k,i} - (Y_{k,i}-\phi))^2 \\
-\sum_{i=1}^n \frac{X_{k,i}}{\sigma^2(1+\lambda_k^2)}(\lambda_k X_{k,i} - (Y_{k,i}-\phi)).
\end{cases}$$

However, $E\left(\nabla\tilde{f}(\{X_{k,i}\},\{Y_{k,i}\}|\{Z_{k,i}\},\theta_k^*)\right) = (0, n\lambda_k^*/(1+\lambda_k^{*2}))^T \neq 0_{2\times 1}$, thus a correction term should be added to construct an appropriate objective function which satisfies the standard first-order condition in statistical optimization framework:

$$\begin{aligned} M_{n,k}(\{X_{k,i}\},\{Y_{k,i}\}|\{Z_{k,i}\},\theta_k) &= -log\tilde{f}(\{X_{k,i}\},\{Y_{k,i}\}|\{Z_{k,i}\},\theta_k) + \frac{n}{2}log(1+\lambda_k^2) \\ &= \frac{1}{2\sigma^2(1+\lambda_k^2)}\sum_{i=1}^{n}(\lambda_k X_{k,i} - (Y_{k,i}-\phi))^2 + C(\sigma), \end{aligned}$$

where $C(\sigma) = nlog(\sqrt{2\pi}\sigma)$ is an absolute constant so we also discard it. The corresponding M-function is

$$M(X_k,\theta_k) = \frac{1}{2\sigma^2(1+\lambda_k^2)}(\lambda_k X_k - (Y_k-\phi))^2. \tag{19}$$

Below we check the identification of the true parameter under this objective function. We can directly solve the population level first-order conditions (FOC) using $E\left(\nabla M(X_k,Y_k|Z_k,\theta_k)\right) = 0_{2\times 1}$, which are given as

$$0_{2\times 1} = \begin{pmatrix} (1+\lambda_k^2)\big((\lambda_k-\lambda_k^*)E\left(Z_k\right) - (\phi^*-\phi)\big) \\ (\lambda_k\lambda_k^*+1)(\lambda_k-\lambda_k^*)E\left(Z_k^2\right) - \lambda_k(\phi-\phi^*)^2 + (\phi-\phi^*)(1+2\lambda_k\lambda_k^*-\lambda_k^2)E\left(Z_k\right) \end{pmatrix}. \tag{20}$$

To solve the above set of equations, we consider the two scenarios. When $EZ_k = 0$, from the first equation we obtain $\phi = \phi^*$, then the second equation reduces to $C(\lambda_k\lambda_k^*+1)(\lambda_k-\lambda_k^*)EZ_k^2 = 0$. Since we have assumed $\lambda_k, \lambda_k^* > 0$, we must have $\lambda_k = \lambda_k^*$. When $E\left(Z_k\right) \neq 0$, if $\lambda_k \neq \lambda_k^*$ we would obtain $E\left(Z_k\right) = (\phi^*-\phi)/(\lambda_k-\lambda_k^*)$. Plugging it into the second equation of (20) and we can obtain

$$\frac{(1+\lambda_k\lambda_k^*)}{\sigma^2(1+\lambda_k^2)^2(\lambda_k-\lambda_k^*)}\left((\lambda_k-\lambda_k^*)^2 EZ_k^2 - (\phi-\phi^*)^2\right) = 0,$$

which is impossible unless Z_k is degenerate, namely $Z_k = (\phi^*-\phi)(\lambda_k-\lambda_k^*)$ with probability one. This leads to a contradiction. Thus we must have $\lambda_k = \lambda_k^*$. Again from the first equation of (20) we will obtain that $\phi = \phi^*$. In summary, $E\nabla M(X_k,Y_k|Z_k,\theta_k) = 0_{2\times 1}$ if and only if $\theta_k = \theta_k^*$.

To give an explicit form of asymptotic variance of the estimator obtained from the M-function (19), we can directly calculate the following two terms:

$$\begin{aligned} &E\left(\nabla^2 M(X_k,Y_k|Z_k;\theta_k^*)\right) \\ = \; &E\begin{pmatrix} \frac{1}{\sigma^2(1+\lambda_k^{*2})} & \frac{X_k}{\sigma^2(1+\lambda_k^{*2})} - \frac{2\lambda_k^*(\lambda_k^* X_k-(Y_k-\phi^*))}{\sigma^2(1+\lambda_k^{*2})^2} \\ \frac{X_k}{\sigma^2(1+\lambda_k^{*2})} - \frac{2\lambda_k^*(\lambda_k^* X_k-(Y_k-\phi^*))}{\sigma^2(1+\lambda_k^{*2})^2} & \frac{(3\lambda_k^{*2}-1)(\lambda_k^* X_k-(Y_k-\phi^*))^2}{\sigma^2(1+\lambda_k^{*2})^3} - \frac{4\lambda_k^* X_k(\lambda_k^* X_k-(Y_k-\phi^*))}{\sigma^2(1+\lambda_k^{*2})^2} + \frac{X_k^2}{\sigma^2(1+\lambda_k^{*2})} \end{pmatrix} \\ = \; &\frac{1}{\sigma^2(1+\lambda_k^{*2})}\begin{pmatrix} 1 & EZ_k \\ EZ_k & EZ_k^2 \end{pmatrix} \quad \text{and} \\ &E\left(\nabla M(X_k,Y_k|Z_k,\theta_k^*)(\nabla M(X_k,Y_k|Z_k,\theta_k^*))^T\right) \\ = \; &\begin{pmatrix} \frac{1}{\sigma^2(1+\lambda_k^{*2})} & \frac{EZ_k}{\sigma^2(1+\lambda_k^{*2})} \\ \frac{EZ_k}{\sigma^2(1+\lambda_k^{*2})} & \frac{EZ_k^2}{\sigma^2(1+\lambda_k^{*2})} + \frac{1}{(1+\lambda_k^{*2})^2} \end{pmatrix} = \frac{1}{\sigma^2(1+\lambda_k^{*2})}\begin{pmatrix} 1 & EZ_k \\ EZ_k & EZ_k^2 + \frac{\sigma^2}{1+\lambda_k^{*2}} \end{pmatrix}. \end{aligned}$$

Thus we have

$$\begin{aligned} J_{\phi|\lambda}(\theta_k^*) &= \frac{1}{\sigma^2(1+\lambda_k^{*2})}(1-\frac{(EZ_k)^2}{EZ_k^2}) = \frac{1}{\sigma^2(1+\lambda_k^{*2})}\frac{Var(Z_k)}{EZ_k^2} \quad \text{and} \\ Var(S_\phi) &= \begin{pmatrix} 1 & -\frac{EZ_k}{EZ_k^2} \end{pmatrix} \frac{1}{\sigma^2(1+\lambda_k^{*2})} \begin{pmatrix} 1 & EZ_k \\ EZ_k & EZ_k^2 + \frac{\sigma^2}{1+\lambda_k^{*2}} \end{pmatrix} \begin{pmatrix} 1 \\ -\frac{EZ_k}{EZ_k^2} \end{pmatrix} \\ &= \frac{1}{\sigma^2(1+\lambda_k^{*2})}\left\{\frac{Var(Z_k)}{EZ_k^2} + \frac{\sigma^2}{1+\lambda_k^{*2}}\frac{(EZ_k)^2}{(EZ_k^2)^2}\right\}, \end{aligned}$$

which leads to the Equation (8) in the main text.

A.4 Equivalent variance minimization formulations of the weighted estimators

For simplicity, we assume that $n_1 = n_2 = \cdots = n_K = n$. We claim that the following two formulations of the variance minimization problem have identical solution.

Formulation 1: Trace Operator

$$\underset{W_k}{Minimize} \quad tr\left(\sum_{k=1}^{K} W_k H_k W_k^T\right), \quad s.t. \quad \sum_{k=1}^{K} W_k = I_{p_1}. \tag{21}$$

Formulation 2: Frobenius Norm

$$\underset{W_k}{Minimize} \quad \|\sum_{k=1}^{K} W_k H_k W_k^T\|_F, \quad s.t. \quad \sum_{k=1}^{K} W_k = I_{p_1}. \tag{22}$$

Proof We solve problem (21) first. The Lagrangian of this problem is $\tilde{L}_1 = tr\left(\sum_{k=1}^K W_k H_k W_k^T\right) + < \Lambda_1, \sum_{k=1}^K W_k - I_{p_1} >$, where $\Lambda_1 \in \mathbb{R}^{p_1 \times p_1}$ is the corresponding Lagrangian multiplier. If we take derivative of $\tilde{L}_1$ w.r.t. W_k we can obtain $2W_k H_k + \Lambda_1 = \mathbf{0}, k = 1, 2, \cdots, K$. Then $W_k = -\frac{1}{2}\Lambda_1 H_k^{-1}$. Using the constraint $\sum_{k=1}^K W_k = I_{p_1}$, we can obtain $\Lambda_1^* = -2(\sum_{s=1}^K A_s^{-1})^{-1}$ and $W_k^* = (\sum_{s=1}^K A_s^{-1})^{-1} A_k^{-1}$. Now we turn to solve the problem (22). Equivalently we can minimize the square of the Frobenius norm, and the corresponding Lagrangian is $\tilde{L}_2 = \| \sum_{k=1}^K W_k H_k W_k^T \|_F^2 + < \Lambda_2, \sum_{k=1}^K W_k - I_{p_1} >$. Taking derivative w.r.t. W_k we can obtain $4(\sum_{s=1}^K W_s A_s W_s^T) W_k A_k + \Lambda_2 = \mathbf{0}$. Now we can use the constraint $\sum_{k=1}^K W_k = I_{p_1}$ and get $\Lambda_2^* = -4(\sum_{s=1}^K W_s A_s W_s^T)(\sum_{s=1}^K A_s^{-1})^{-1}$ and $W_k^* = (\sum_{s=1}^K A_s^{-1})^{-1} A_k^{-1}$. ■

A.5 Second-order Bartlett's indentity under QMLE

For the quasi maximum likelihood estimation (QMLE), we only check that the second order Bartlett's identity holds for independent observarions. Suppose that the components of the response vector Y are independent with mean vector μ and covariance matrix $\sigma^2 V(\mu)$, where σ^2 maybe unknown and $V(\mu)$ is a matrix of known functions. It is assumed that the parameters of interest, θ, is a function of μ. By independence of the components of Y and the physical mechanism plausibility, it is reasonable to assume further that $V_i(\mu)$ depends on μ only through μ_i, which implies that

$$V(\mu) = diag\{V_1(\mu_1), V_2(\mu_2), \cdots, V_n(\mu_n)\}.$$

For a single observation Y, we can construct the score function as $U = u(\mu; Y) = (Y - \mu)/(\sigma^2 V(\mu))$. Then the corresponding objective function can be defined as

$$Q(\mu; y) = -\int_y^\mu \frac{y - t}{\sigma^2 V(t)} dt, \tag{23}$$

which behaves like a negative log-likelihood: $E(\nabla_\mu Q) = 0$, $Var(\nabla_\mu Q) = E(\nabla_\mu^2 Q) = 1/\{\sigma^2 V(\mu)\}$. We refer to $Q(\mu; y)$ as the negative quasi-likelihood (McCullagh, 1983), or more precisely the negative log quasi-likelihood for μ based on data y. By independence, the negative quasi-likelihood for the complete data is the sum of the individual contributions: $Q(\mu; y) = \sum_{i=1}^n Q(\mu_i; y_i)$. The quasi-likelihood estimating equations for the regression parameters θ, obtained by differentiating $Q(\mu; y)$, can be written in the form $U(\hat{\theta}) = 0$, where $U(\theta) = -DV^{-1}(Y - \mu)/\sigma^2$ is called the quasi-score function. The components of D, of order $n \times p$, are $D_{ir} = \partial \mu_i / \partial \theta_r$, the derivatives of $\mu(\theta)$ with respect to the parameters. Suppose the true parameters are θ^* and μ^*, then by the zero-mean of $U(\theta^*)$, we have

$$\begin{aligned} \text{CoV}\{U(\theta^*)\} &= E\left(U(\theta^*)U(\theta^*)^T\right) = D^T V^{-1} D/\sigma^2 \quad \text{and} \\ E\left(\frac{\partial U}{\partial \theta^T}(\theta^*)\right) &= E\{D^T V^{-1} \frac{\partial \mu}{\partial \theta^T}/\sigma^2 + \frac{\partial D^T V^{-1}}{\partial \theta^T} \frac{Y - \mu^*}{\sigma^2}\} = D^T V^{-1} D/\sigma^2. \end{aligned}$$

A.6 Generalized second-order Bartlett's identity for parametric regression

Suppose that we observe a random sample $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, which follows

$$Y = f_{\theta^*}(X) + e, E(e|X) = 0, Var(e|X) = \sigma^2(X), X \sim p(x).$$

Then the objective function for the least square estimation is $M(Z, \theta) = (Y - f_\theta(X))^2$ with $Z = (X, Y)$. Note that

$$E\left(M(Z, \theta)\right) = E(f_\theta(X) - f_{\theta^*}(X))^2 + Ee^2 \approx E\left(M(Z, \theta^*)\right) + E((\theta - \theta^*)^T \nabla f_{\theta^*}(X))^2, \tag{24}$$

which suggests that $\nabla_\theta^2 M(\theta^*) = 2E\nabla f_{\theta^*}(X) \nabla f_{\theta^*}(X)^T$ where $M(\theta) = EM(Z, \theta)$. For the approximation (24), see van der Vaart (1999). If we assume the independence between e and X, which implies $Var(e) = \sigma^2$, then $E\left(\nabla M(Z, \theta^*) \nabla M(Z, \theta^*)^T = 4\sigma^2 E\nabla f_{\theta^*}(X) \nabla f_{\theta^*}(X)^T\right)$ with the multiplicative factor γ for the generalized second-order Bartlett's identity being $4\sigma^2$.

A.7 GMM formulation of the full sample statistical optimization under heterogeneity

It is noted that W_0 admits the following form

$$W_0 = \begin{pmatrix} Var\{\psi_{\theta_1}(X_{1,1}; \phi^*, \lambda_1^*)\}^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & Var\{\psi_{\theta_2}(X_{2,1}; \phi^*, \lambda_2^*)\}^{-1} & & \mathbf{0} \\ \vdots & & \ddots & \\ \mathbf{0} & \mathbf{0} & & Var\{\psi_{\theta_K}(X_{K,1}; \phi^*, \lambda_K^*)\}^{-1} \end{pmatrix}.$$

Thus, W_0 is a block diagonal matrix. Also note that

$$
\begin{aligned}
G_0^T &= E\{\frac{\partial \tilde{\psi}_N^T(\theta^*)}{\partial \theta}\} \\
&= E\begin{pmatrix}
\psi_\phi^\phi(X_{1,i};\phi^*,\lambda_1^*) & \psi_\phi^\lambda(X_{1,i};\phi^*,\lambda_1^*) & \cdots & \cdots & \psi_\phi^\phi(X_{K,i};\phi^*,\lambda_K^*) & \psi_\phi^\lambda(X_{K,i};\phi^*,\lambda_K^*) \\
\psi_\lambda^\phi(X_{1,i};\phi^*,\lambda_1^*) & \psi_\lambda^\lambda(X_{1,i};\phi^*,\lambda_1^*) & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \psi_\lambda^\phi(X_{2,i};\phi^*,\lambda_2^*) & \psi_\lambda^\lambda(X_{2,i};\phi^*,\lambda_2^*) & \cdots & \mathbf{0} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \psi_\lambda^\phi(X_{K,i};\phi^*,\lambda_K^*) & \psi_\lambda^\lambda(X_{K,i};\phi^*,\lambda_K^*)
\end{pmatrix}
\end{aligned}
$$

then the asymptotic variance of the GMM estimtator (Hansen, 1982) is $AsyVAr(\hat{\theta}_{GMM}) = (G_0^T W_0 G_0)^{-1}$ and has the following form:

$$
\begin{pmatrix}
\sum_{k=1}^K n_k D\Psi_\phi(\theta_k^*)^T \Sigma_{S,k}^{-1} D\Psi_\phi(\theta_k) & n_1 D\Psi_\phi(\theta_1^*)^T \Sigma_{S,1}^{-1} D\Psi_\lambda(\theta_1^*) & \cdots & \cdots & n_K D\Psi_\phi(\theta_K^*)^T \Sigma_{S,K}^{-1} D\Psi_\lambda(\theta_K^*) \\
n_1 D\Psi_\lambda(\theta_1^*)^T \Sigma_{S,1}^{-1} D\Psi_\phi(\theta_1^*) & n_1 D\Psi_\lambda(\theta_1^*)^T \Sigma_{S,1}^{-1} D\Psi_\lambda(\theta_1^*) & \mathbf{0} & \cdots & \mathbf{0} \\
\vdots & \mathbf{0} & \ddots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \mathbf{0} \\
n_K D\Psi_\lambda(\theta_K^*)^T \Sigma_{S,K}^{-1} D\Psi_\phi(\theta_K^*) & \mathbf{0} & \cdots & \mathbf{0} & n_K D\Psi_\lambda(\theta_K^*)^T \Sigma_{S,K}^{-1} D\Psi_\lambda(\theta_K^*)
\end{pmatrix}^{-1},
$$

where

$$
D\Psi_\phi(\theta_k)^T = \begin{pmatrix} \Psi_\phi^\phi(\theta_k) & \Psi_\phi^\lambda(\theta_k) \end{pmatrix}, D\Psi_\lambda(\theta_k)^T = \begin{pmatrix} \Psi_\lambda^\phi(\theta_k) & \Psi_\lambda^\lambda(\theta_k) \end{pmatrix} \text{ and } \Sigma_{S,k} = Var\{\psi_{\theta_k}(X_{k,1};\phi^*,\lambda_k^*)\}.
$$

By the inversion of block matrix, $AsyVar(\hat{\phi}_{GMM})^{-1}$ has the following form:

$$
\sum_{k=1}^K n_k \Big\{ D\Psi_\phi(\theta_k^*)^T \Sigma_{S,k}^{-1} D\Psi_\phi(\theta_k^*) - D\Psi_\phi(\theta_k^*)^T \Sigma_{S,k}^{-1} D\Psi_\lambda(\theta_k^*) \Big(D\Psi_\lambda(\theta_k^*)^T \Sigma_{S,k}^{-1} D\Psi_\lambda(\theta_k^*) \Big)^{-1} D\Psi_\lambda(\theta_k^*)^T \Sigma_{S,k}^{-1} D\Psi_\phi(\theta_k^*) \Big\}.
$$

If we denote the elements in the above summation as $n_k U_k$, then it is straightforward to verify that

$$
\begin{pmatrix} U_k^{-1} & * \\ * & * \end{pmatrix} = \left\{ \begin{pmatrix} D\Psi_\phi(\theta_k^*)^T \\ D\Psi_\lambda(\theta_k^*)^T \end{pmatrix} \Sigma_{S,k} \begin{pmatrix} D\Psi_\phi(\theta_k^*) & D\Psi_\lambda(\theta_k^*) \end{pmatrix} \right\}^{-1},
$$

namely, the inverse of U_k is the left top part of the inverse of a bigger matrix in the RHS of the above equation, from which we are able to obtain the simplified expression of U_k:

$$
\begin{aligned}
U_k &= \left\{ J_{\phi|\lambda}^{-1} \begin{pmatrix} I_{p_1\times p_1} & -\Psi_\phi^\lambda(\theta_k^*)\Psi_\lambda^\lambda(\theta_k^*)^{-1} \end{pmatrix} \Sigma_{S,k} \begin{pmatrix} I_{p_1\times p_1} \\ -\Psi_\lambda^\lambda(\theta_k^*)^{-1}\Psi_\lambda^\phi(\theta_k^*) \end{pmatrix} J_{\phi|\lambda_k^*}^{-1} \right\}^{-1} \\
&= J_{\phi|\lambda} \Sigma_k^{-1} J_{\phi|\lambda}.
\end{aligned}
$$

Now we conclude that $AsyVar(\hat{\phi}_{GMM}) = \left(\sum_{k=1}^K J_{\phi|\lambda} \Sigma_k^{-1} J_{\phi|\lambda} \right)^{-1}$, which is the same as that of the WD estimator $\hat{\phi}^{WD}$.

A.8 Asymptotic efficiency comparison of $\hat{\lambda}_k$ and $\hat{\lambda}_k^{(2)}$

Theorem 9 *Under the conditions required in Theorem 4, if $K \to \infty$, then for the updated estimator $\hat{\lambda}_k^{(2)}$, we have that*

$$\sqrt{n_k}(\hat{\lambda}_k^{(2)} - \lambda_k^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Psi_\lambda^\lambda(\theta_k^*)^{-1} E\psi_\lambda(X_{k,1};\theta_k^*)\psi_\lambda(X_{k,1};\theta_k^*)^T \Psi_\lambda^\lambda(\theta_k^*)^{-1}). \tag{25}$$

Hence, the asymptotic distribution of $\hat{\lambda}_k^{(2)}$ is the same as that of the estimator of λ_k^* obtained when the common parameter ϕ^* is known. It is noted that the joint asymptotic distribution for the estimator $\hat{\theta}_k = (\hat{\phi}_k^T, \hat{\lambda}_k^T)^T$ is

$$\sqrt{n_k}(\hat{\theta}_k - \theta_k^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Psi_\theta^\theta(\theta_k^*)^{-1} E\psi_{\theta_k}(X_{k,1};\theta_k^*)\psi_{\theta_k}(X_{k,1};\theta_k^*)^T \Psi_\theta^\theta(\theta_k^*)^{-1}),$$

which leads to

$$\sqrt{n_k}(\hat{\lambda}_k - \lambda_k^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, J_{\lambda|\phi}(\theta_k^*)^{-1} Var(S_\phi(X_{k,1};\theta_k^*)) J_{\lambda|\phi}(\theta_k^*)^{-1}). \tag{26}$$

There is not a definite order on the relative efficiency between $\hat{\lambda}_k$ and $\hat{\lambda}_k^{(2)}$ by comparing the two asymptotic variances in (25) and (26), suggesting it would depend on the specific M function and the model setting. For general statistical optimization, a known nuisance parameter (here ϕ^*) does not necessarily improve the efficiency of a parameter of interest Yuan and Jennrich (2000); Henmi and Eguchi (2004), which is the case for the current setting. Consider again the errors-in-variables model where it can be shown that

$$Var(\hat{\lambda}_k^{(2)}) \approx \frac{\sigma^4}{(Var(Z_k))^2}\frac{1}{n_k} \quad \text{and} \quad Var(\hat{\lambda}_k) \approx \left(\frac{\sigma^4}{(E(Z_k^2))^2} + \frac{\sigma^2(1+\lambda_k^2)}{E(Z_k^2)}\right)\frac{1}{n_k}.$$

When $E(Z_k) = 0$, i.e. $Var(Z_k) = E(Z_k^2)$, the updated estimator $\hat{\lambda}_k^{(2)}$ is more efficient, and the efficiency gain gets large as λ_k^2 increases. However, if $E(Z_k)$ has a large absolute magnitude, $\hat{\lambda}_k$ can be more efficient than $\hat{\lambda}_k^{(2)}$. Moreover, the requirement in Theorem 9 that $K \to \infty$ is to obtain a succinct asymptotic variance of $\hat{\lambda}_k^{(2)}$. The above conclusion does not change for the fixed K case. Consider block 1, we assume $\hat{\lambda}_1^{(2)} \xrightarrow{p} \lambda_1^*$ and $\hat{\phi}^{WD}$ is $\sqrt{n_1}-$ consistent (detailed proofs of both claims are available in the next section). Then by Theorem 1 in Yuan and Jennrich (2000), if $\sqrt{n_1}\big(\frac{1}{n_1}\sum_{i=1}^{n_1}\psi_\lambda(X_{1,i};\theta_1^*) + \Psi_\lambda^\phi(\theta_1^*)(\hat{\phi}^{WD} - \phi^*)\big) \xrightarrow{d} \mathcal{N}(0, Q)$, we will have $\sqrt{n_1}(\hat{\lambda}_1^{(2)} - \lambda_1^*) \xrightarrow{d} \mathcal{N}(0, \Omega)$ where $\Omega = \Psi_\lambda^\lambda(\theta_1^*)^{-1} Q \Psi_\lambda^\lambda(\theta_1^*)^{-1}$. Denote $T_{n,K} = \sqrt{n}\Psi_\lambda^\lambda(\theta_1^*)^{-1}\big(\frac{1}{n}\sum_{i=1}^{n}\psi_\lambda(X_{1,i};\theta_1^*) + \Psi_\lambda^\phi(\theta_1^*)(\hat{\phi}^{WD} - \phi^*)\big)$, then $T_{n,K}$ should have the same asymptotic distribution as $\sqrt{n}(\hat{\lambda}_1^{(2)} - \lambda_1^*)$. So, we study the limiting behavior of $T_{n,K}$ for simplicity. Consider the homogeneous scenario as a special case when $\theta_1^* = \theta_2^* = \cdots = \theta_K^*$, $n_1 = n_2 = \cdots = n_K = n$, then the optimal weights are

$W_1^* = W_2^* = \cdots = W_K^* = \frac{1}{K} I_{p_1 \times p_1}$. Now we have

$$\begin{aligned} T_{n,K} &= \frac{1}{\sqrt{n}} \Psi_\lambda^\lambda(\theta_1^*)^{-1} \big(\sum_{i=1}^n \psi_\lambda(X_{1,i}; \theta_1^*) - \Psi_\lambda^\phi(\theta_1^*) \sum_{k=1}^K \frac{1}{K} (\hat{\phi}_k - \phi^*) \big) \\ &= \frac{1}{\sqrt{n}} \Psi_\lambda^\lambda(\theta_1^*)^{-1} \big(\sum_{i=1}^n \psi_\lambda(X_{1,i}; \theta_1^*) - \Psi_\lambda^\phi(\theta_1^*) \sum_{k=1}^K \frac{1}{K} \sum_{i=1}^n J_{\phi|\lambda}^{-1} S_\phi(X_{k,i}; \theta_k^*) \big) + o_p(1) \\ &= \Big((1 - \tfrac{1}{K}) \Psi_\lambda^\lambda(\theta_1^*)^{-1} \Psi_\lambda^\phi(\theta_1^*) \quad I_{p_2 \times p_2} \Big) \nabla^2 M_1(\theta_1^*)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} \psi_\phi(X_{1,i}; \theta_1^*) \\ \psi_\lambda(X_{1,i}; \theta_1^*) \end{pmatrix} \\ &\quad - \Psi_\lambda^\lambda(\theta_1^*)^{-1} \Psi_\lambda^\phi(\theta_1^*) \frac{1}{\sqrt{n}} \frac{1}{K} \sum_{k=2}^K \sum_{i=1}^n J_{\phi|\lambda}^{-1} S_\phi(X_{k,i}; \theta_k^*) + o_p(1) \triangleq T_{n,k}^{(1)} + o_p(1). \end{aligned}$$

We can verify that $Var(T_{n,k}^{(1)}) = (1 - \frac{1}{K}) \Psi_\lambda^\lambda(\theta_1^*)^{-1} Var(\psi_\lambda(X_{1,1}; \theta_1^*)) \Psi_\lambda^\lambda(\theta_1^*)^{-1} + \frac{n}{K} AsyVar(\hat{\lambda}_1)$, or equivalently, $AsyVar(\hat{\lambda}_1^{(2)}) \approx (1 - \frac{1}{K}) \Psi_\lambda^\lambda(\theta_1^*)^{-1} Var(\psi_\lambda(X_{1,1}; \theta_1^*)) \Psi_\lambda^\lambda(\theta_1^*)^{-1} \frac{1}{n} + \frac{1}{K} AsyVar(\hat{\lambda}_1)$. Thus $AsyVar(\hat{\lambda}_1^{(2)}) \preceq AsyVar(\hat{\lambda}_1)$ if and only if

$$\Psi_\lambda^\lambda(\theta_1^*)^{-1} Var(\psi_\lambda(X_{1,1}, \theta_1^*)) \Psi_\lambda^\lambda(\theta_1^*)^{-1} / n \preceq AsyVar(\hat{\lambda}_1). \tag{27}$$

The LHS of inequality (27) is the asymptotic variance of the estimator of λ_1^* if ϕ_1^* is known and RHS is the asymptotic variance of estimator of λ_1^* when we jointly estimate $(\phi_1^{*T}, \lambda_1^{*T})^T$. Henmi and Eguchi (2004) showed that the inequality does not always hold for general statistical optimization problem and derived a sufficient condition under which a known nuisance parameter (ϕ^*) will lead to a bigger asymptotic variance of the estimator of the parameter of interest (λ_1^*).

A.9 Bias correction for statistical optimization under logistic regression model

Given observations $\{(y_i, X_i)\}_{i=1}^n$, we now construct $\hat{B}(\beta)$. Denote $y = (y_1, y_2, \cdots, y_n)^T, X = (X_1, X_2, \cdots, X_n)^T$ and $\hat{y} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)$ with $\hat{y}_i = logit(x_i^T \beta)$. Since $\frac{d^j}{da^j} logit(a) = logit(a) \prod_{s=1}^j (1 - logit(a) s)$, then we have $\nabla M_n(\beta) = \frac{1}{n} X^T (\hat{y} - y)$, $\nabla^2 M_n(\beta) = \frac{1}{n} X^T diag\{\hat{y} \cdot (1 - \hat{y})\} X$ and $\nabla^3 M_n(\beta) = \frac{1}{n} \sum_{i=1}^n \hat{y}_i (1 - \hat{y}_i)(1 - 2\hat{y}_i) x_i vec(x_i \otimes x_i)^T$, where $\cdot$ denotes the element-wise product of two vectors and vec is the vectorization operator. Then, the bias-correction formula is a combination of the gradients up to the third order.

A.10 Comparison with a one-step estimator

Huang and Huo (2019), also under the same homogeneous setting, considered to utilize the second order information of the M-function to allow for a larger K. They proposed a one-step estimator which aggregates the local Hessian matrices and gradients and performs a single Newton-Raphson updating. The estimator, denoted as $\hat{\theta}^{(1)}$, has a MSE upper bound

$$E\left(\|\hat{\theta}^{(1)} - \theta_1^*\|_2^2 \right) \le \frac{2E\left(\|\nabla_{\theta_1} \Psi_\theta(\theta_1^*)^{-1} \psi_{\theta_1}(X_{1,1}; \theta_1^*)\|_2^2 \right)}{nK} + \frac{C_1}{N^2} + \frac{C_2}{n^4}. \tag{28}$$

Thus, this method allows for $K = o(n^3)$, while still preserves the $\mathcal{O}(N^{-1})$ convergence rate. The price of this procedure is one extra round of transmission of the local Hessians and gradients. To mitigate the communication burden, they considered to use only one local Hessian matrix instead of the averaged one. Let $\hat{\theta}_{LH}^{(1)}$ be the estimator. They showed that

$$E\left(\|\hat{\theta}_{LH}^{(1)}-\theta_1^*\|_2^2\right) \leq \frac{2E\left(\|\nabla_{\theta_1}\Psi_\theta(\theta_1^*)^{-1}\psi_{\theta_1}(X_{1,1};\theta_1^*)\|_2^2\right)}{nK} + \frac{C_1}{n^2K} + \frac{C_2}{n^3}, \tag{29}$$

which is similar to the MSE bound of the dSaC estimator in Corollary 8. However, both $\hat{\theta}^{(1)}$ and $\hat{\theta}_{LH}^{(1)}$ are not readily extended to the heterogeneous setting, as the one-step update procedure relies crucially on the $N^{1/2}-$consistency of the initial estimators of all the unknown parameters (van der Vaart, 1999), but the convergence rate of the block-specific estimators $\hat{\lambda}_k$ are only of order $\mathcal{O}_p(n_k^{1/2})$.

B. Proofs

Without loss of generality, we assume equal sample size n in each data block. Besides, unless otherwise stated, we will use C, c_i and C_i to denote positive constants independent of (n_k, K, N), and the same C_i can have different values from one context to another.

B.1 Lemmas

Before presenting the proofs of the theoretical results established in the main paper, we first establish some technical lemmas in the following sub-section.

Lemma B.1 *Suppose H and K are positive definite matrices of order p, and X and Y are arbitrary $p \times m$ matrices. Then, $Q = X^TH^{-1}X + Y^TK^{-1}Y - (X+Y)^T(H+K)^{-1}(X+Y) \succeq \mathbf{0}$.*

Proof Let A and B be defined as follows

$$A = \begin{pmatrix} H & X \\ X^T & X^TH^{-1}X \end{pmatrix}, \quad B = \begin{pmatrix} K & Y \\ Y^T & Y^TK^{-1}Y \end{pmatrix}$$

Since H, K are positive definite, we can directly check that A, B are positive semi-definite. Thus $A + B$ is also positive semi-definite, and the conclusion follows. See Ando (1979); Haynsworth (1970) for more similar types of matrix inequalities. ■

Lemma B.2 *Under Assumptions 1 - 4 and Assumptions 5 - 6 with $v_2 = min\{v, v_1\} \geq 1$, if $K = o(n^{v_2})$, then*

$$\sup_{1\leq k\leq K} \|\hat{\theta}_k - \theta_k^*\|_2 \xrightarrow{P} 0.$$

Proof Let $G_{n,k} = \frac{1}{n}\sum_{i=1}^n G_k(X_{k,i})$ and $\delta_\rho = min\{\rho, \rho\rho_-/4G\}$. For $k = 1, ..., K$, define the following "good events":

$$\mathcal{E}_k = \{G_{n,k} \leq 2G, \|\nabla^2_{\theta_k}M_{n,k}(\theta_k^*) - \nabla^2_{\theta_k}M_k(\theta_k^*)\|_2 \leq \frac{\rho\rho_-}{2}, \|\nabla_{\theta_k}M_{n,k}(\theta_k^*)\|_2 \leq \frac{(1-\rho)\rho_-\delta_\rho}{2}\}.$$

Then by Lemma 6 in Zhang et al. (2013), we obtain that under the event $\cap_{k=1}^K \mathcal{E}_k$,

$$\|\hat{\theta}_k - \theta_k^*\|_2 \leq \frac{2\|\nabla_{\theta_k}M_{n,k}(\theta_k^*)\|_2}{(1-\rho)\rho_-}.$$

Similar to the proof of Lemma C.1 in Jordan et al. (2019), there exist constants c_1, c_2, c_3 independent of (n, K, d, G, L) such that $P(\cup_{k=1}^K \mathcal{E}_k^c) \leq (c_1 + c_2(log2d)^{2v}L^{2v} + c_3R^{2v})\frac{K}{n^v}$. For any $\epsilon > 0$ and $k \leq K$, we define events $\mathcal{E}_k^{'} = \{\|\nabla_{\theta_k} M_{n,k}(\theta_k^*)\|_2 \leq (1-\rho)\rho_-\epsilon/2\}$. Then by Markov's inequality and the union bound, there exist constants c_4 such that $P(\cup_{k=1}^K \mathcal{E}_k^{'c}) \leq c_4 K L^{2v}/n^{v_2}$. Thus, $P(\sup_{1\leq k\leq K} \|\hat{\theta}_k - \theta_k^*\| > \epsilon) = O(\frac{K}{n^{v_2}})$, implying that $\sup_{1\leq k\leq n} \|\hat{\theta}_k - \theta_k^*\|_2 \xrightarrow{P} 0$ for $K = o(n^{v_2})$. ■

Lemma B.3 *$Inv(A) : GL(\mathbb{R}^p) \to GL(\mathbb{R}^p) : A \mapsto A^{-1}$ is Lipschitz continuous at any $A \in GL(\mathbb{R}^p)$, where $GL(\mathbb{R}^p)$ consists of all $p \times p$ invertible matrices of real numbers.*

Proof Let $A_0 \in GL(\mathbb{R}^p)$ be given. Denote $1/\|A_0^{-1}\|_2 = \delta > 0$. It follows that for all $x \in \mathbb{R}^p$ we have $\|x\|_2 = \|A_0^{-1}A_0x\|_2 \leq (1/\delta)\|A_0x\|_2$, namely $\|A_0x\|_2 \geq \delta\|x\|_2$. Assume that $\|A - A_0\|_2 < \delta/2$, then $\|Ax\|_2 \geq \|A_0x\|_2 - \|(A - A_0)x\|_2 \geq \frac{\delta}{2}\|x\|_2$, which means A^{-1} exists and $\|A^{-1}\|_2 \leq 2/\delta$. Since $A^{-1} - A_0^{-1} = A^{-1}(A_0 - A)A_0^{-1}$, $\|A^{-1} - A_0^{-1}\|_2 \leq \|A^{-1}\|_2\|A_0 - A\|_2\|A_0^{-1}\|_2 \leq (2/\delta^2)\|A - A_0\|_2$, which completes the proof. ■

Lemma B.4 *Under Assumptions 1 - 4 and 7, and Assumptions 5 - 6 for $v, v_1 \geq 2$, if $K = o(n)$,*

$$\{n_k \sum_{k=1}^K H_k(\theta_k^*)^{-1}(\hat{\phi}_k - \phi^*)\}^T \{\sum_{k=1}^K n_k H_k(\theta_k^*)^{-1}\}^{-1} \{\sum_{k=1}^K n_k H_k(\theta_k^*)^{-1}(\hat{\phi}_k - \phi^*)\} \xrightarrow{d} \chi_{p_1}^2.$$

Proof We prove for the case when $K \to \infty$, and the proof for the fixed K case is straightforward to derive. Denote $T_1 = \{\frac{1}{K}\sum_{s=1}^K H_s(\theta_s^*)^{-1}\}^{-1}\frac{1}{K}\sum_{k=1}^K H_k(\theta_k^*)^{-1}(\hat{\phi}_k - \phi^*)$, $J_k(\theta_k) = E\nabla_{\theta_k}^2 M(X_{k,1};\theta_k)$. Lemma B.2 has shown that $P(\cap_{k=1}^K \mathcal{E}_k) = 1 - \mathcal{O}(K/n^{v_2})$, where $v_2 = min\{v, v_1\}$. And since all the smoothness conditions in Assumptions 5 - 6 only holds locally, namely in the U_ρ ball, so all the expansions hold only under the event $\cap_{k=1}^K \mathcal{E}_k$. When $K = o(1/n^{v_2})$, $P(\cap_{k=1}^K \mathcal{E}_k) \to 1$ and thus $T_1 = T_1 I(\cap_{k=1}^K \mathcal{E}_k) + o_p(1)$. Then by Slutsky's lemma it is equivalent to obtain the asymptotic distribution of $T_1 I(\cap_{k=1}^K \mathcal{E}_k)$. In the following proof, we assume the event $\cap_{k=1}^K \mathcal{E}_k$ holds. By the integral form of Taylor's expansion of $\nabla_{\theta_k} M_{n,k}(\theta_k)$ around the true parameter θ_k^*, we have

$$\hat{\theta}_k - \theta_k^* = -J_k(\theta_k^*)^{-1}\nabla_{\theta_k} M_{n,k}(\theta_k^*) + R_n^{(k)}, \tag{30}$$

where $R_n^{(k)} = R_{n,1}^{(k)} + R_{n,2}^{(k)}$,

$$\begin{aligned} R_{n,1}^{(k)} &= -J_k(\theta_k^*)^{-1}\{\nabla_{\theta_k}^2 M_{n,k}(\theta_k^*) - J_k(\theta_k^*)\}(\hat{\theta}_k - \theta_k^*) \quad \text{and} \\ R_{n,2}^{(k)} &= -J_k(\theta_k^*)^{-1}\{\int_0^1 \nabla_{\theta_k}^2 M_{n,k}(\theta_k^* + t(\hat{\theta}_k - \theta_k^*))dt - \nabla_{\theta_k}^2 M_{n,k}(\theta_k^*)\}(\hat{\theta}_k - \theta_k^*) \end{aligned}$$

for each k. Recall the definition of $J_{\phi|\lambda}$ and $S_\phi(X_{k,i};\theta_k)$, if we denote

$$T_1^{'} = -\{\frac{1}{K}\sum_{s=1}^K H_s(\theta_s^*)^{-1}\}^{-1}\frac{1}{\sqrt{N}}\sum_{k=1}^K\sum_{i=1}^n H_k(\theta_k^*)^{-1}J_{\phi|\lambda}(\theta_k^*)^{-1}S_\phi(X_{k,i};\theta_k^*) + R_1 + R_2, \tag{31}$$

then $T_1 = T_1^{'} I(\cap_{k=1}^K \mathcal{E}_k) + T_1(1 - I(\cap_{k=1}^K \mathcal{E}_k)) = T_1^{'} + (T_1 + T_1^{'})(1 - I(\cap_{k=1}^K \mathcal{E}_k))$. When $K = o(n^{v_2})$, we can directly show that $(T_1 + T_1^{'})(1 - I(\cap_{k=1}^K \mathcal{E}_k)) = o_p(1)$, so as long as we can show the asymptotic normality of $T_1^{'}$, by applying Slutsky's lemma we can also show the asymptotic normality of T_1. The explicit expressions of R_1, R_2 and their asymptotic properties will be investigated after we establish the asymptotic normality of $T_{1,0} = -\{\frac{1}{K}\sum_{s=1}^K H_k(\theta_s^*)^{-1}\}^{-1}\frac{1}{\sqrt{N}}\sum_{k=1}^K\sum_{i=1}^n H_k(\theta_k^*)^{-1}J_{\phi|\lambda}(\theta_k^*)^{-1}S_\phi(X_{k,i};\theta_k^*)$. Here we apply the Cramer-Wold device to reduce the problem into a scalar case. Since $(1/K)\sum_{s=1}^K H_s(\theta_s^*)^{-1}$ may not converge as $K \to \infty$ in the presence of heterogeneity, we turn to establish the asymptotic normality of the standardized version of $T_{1,0}$, which is $T_{1,1} = -\{\frac{1}{K}\sum_{s=1}^K H_s(\theta_s^*)^{-1}\}^{-1/2}\frac{1}{\sqrt{N}}\sum_{k=1}^K\sum_{i=1}^n H_k(\theta_k^*)^{-1}J_{\phi|\lambda}(\theta_k^*)^{-1}S_\phi(X_{k,i};\theta_k^*)$. For any non-zero $l \in \mathbb{R}^{p_1}$, let $l_k^T = -l^T\{\frac{1}{K}\sum_{s=1}^K H_k(\theta_s^*)^{-1}\}^{-1/2}P_k$, where

$$P_k = H_k(\theta_k^*)^{-1}J_{\phi|\lambda}(\theta_k^*)^{-1}\left(I_{p_1\times p_1} \quad -\Psi_\phi^\lambda(\theta_k^*)\Psi_\lambda^\lambda(\theta_k^*)^{-1}\right). \tag{32}$$

Then $l^T T_{1,1} = N^{-1/2}\sum_{k=1}^K\sum_{i=1}^n l_k^T\psi_{\theta_k}(X_{k,i};\theta_k^*)$. If we denote $Z_{K,k} = N^{-1/2}\sum_{i=1}^n l_k^T\psi_{\theta_k}(X_{k,i};\theta_k^*)$, then $l^T T_{1,1} = \sum_{k=1}^K Z_{K,k}$ and $E\left(Z_{K,k}\right) = 0$. Below we check the Lindeberg conditions. First, $\sum_{k=1}^K EZ_{K,k}^2 = l^T l = \sigma_l^2 > 0$. Second, for any $\epsilon > 0$,

$$\begin{aligned}
&\sum_{k=1}^K E(|Z_{K,k}|^2; |Z_{K,k}| > \epsilon) = \sum_{k=1}^K E(|Z_{K,k}|^2\mathbb{I}_{\{|Z_{K,k}|>\epsilon\}}) \\
= \;& 2\sum_{k=1}^K\left(\int_0^\epsilon + \int_\epsilon^\infty\right)tP(|Z_{K,k}I_{\{|Z_{K,k}|>\epsilon\}}| > t)dt \\
= \;& \epsilon^2\sum_{k=1}^K P(|Z_{K,k}| > \epsilon) + 2\sum_{k=1}^K\int_\epsilon^\infty tP(|Z_{K,k}| > t)dt,
\end{aligned}$$

where the second equality comes from the tail-sum formula for expectations of absolute moments. Using Chebyshev's inequality and Marcinkiewicz-Zygmund inequality with b_3 being the corresponding constant, we can show that

$$\sum_{k=1}^K P(|Z_{K,k}| > \epsilon) \;=\; \sum_{k=1}^K P\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n l_k^T\psi_{\theta_k}(X_{k,i};\theta_k^*)\right| > \epsilon\sqrt{K}\right) \le \frac{b_3}{\epsilon^3K^{3/2}}\sum_{k=1}^K\|l_k\|_2^3E\left(\|\psi_{\theta_k}(X_{k,1};\theta_k^*)\|_2^3\right),$$

Recall the definition of l_k, then we can use the boundedness of $H_k(\theta_k^*)$ and $\nabla_{\theta_k}^2 M_k(\theta_k^*)$ to show that $\|l_k\|_2 \le C\|l\|_2$. Thus we have that

$$\sum_{k=1}^K P(|Z_{K,k}| > \epsilon) \le \frac{b_3C}{\epsilon^3K^{3/2}}\|l\|_2K\max_{1\le k\le K}E\left(\|\psi_{\theta_k}(X_{k,1};\theta_k^*)\|_2^3\right) = c_\epsilon\frac{\max_{1\le k\le K}E\left(\|\psi_{\theta_k}(X_{k,1};\theta_k^*)\|_2^3\right)}{\sqrt{K}} \to 0.$$

Now we consider the second part, namely $\sum_{k=1}^K\int_\epsilon^\infty tP(|Z_{K,k}| > t)dt$. Note

$$\begin{aligned}
&\sum_{k=1}^K\int_\epsilon^\infty tP(|Z_{K,k}| > t)dt = \sum_{k=1}^K\int_\epsilon^\infty tP\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n l_k^T\psi_{\theta_k}(X_{k,i};\theta_k^*)\right| > t\sqrt{K}\right)dt \\
\overset{u=t\sqrt{K}}{\le}\;& c\max_{1\le k\le K}E\|\psi_{\theta_k}(X_{k,1};\theta_k^*)\|_2^3\sum_{k=1}^K\frac{1}{K}\int_{\epsilon\sqrt{K}}^\infty\frac{1}{u^2}du \le c'\,\frac{\max_{1\le k\le K}E\|\psi_{\theta_k}(X_{k,1};\theta_k^*)\|_2^3}{\sqrt{K}} \to 0.
\end{aligned}$$

Thus we conclude that $T_{1,1} \xrightarrow{d} \mathcal{N}(0, I_{p_1\times p_1})$. Now we consider the remainder term R_2. We first give the explicit expression of R_2: $R_2 = -(\frac{1}{K}\sum_{k=1}^K H_k(\theta_k^*)^{-1})^{-1}\sqrt{\frac{n}{K}}\sum_{k=1}^K P_k\{\nabla^2_{\theta_k} M_{n,k}(\theta_k^*) - J_k(\theta_k^*)\}(\hat{\theta}_k - \theta_k^*)$. Since $\|\frac{1}{K}\sum_{k=1}^K H_k(\theta_k^*)^{-1}\|_2$ is bounded, we only need to show $R_{2,1} \triangleq \{\frac{1}{K}\sum_{k=1}^K H_k(\theta_k^*)^{-1}\}R_2 = o_p(1)$. Since $\|R_{2,1}\|_2 \leq \sqrt{\frac{K}{n}}\frac{1}{K}\sum_{k=1}^K \|P_k\|_2\|\sqrt{n}\{\nabla^2_{\theta_k} M_{n,k}(\theta_k^*) - J_k(\theta_k^*)\}\|_2\|\sqrt{n}(\hat{\theta}_k - \theta_k^*)\|_2$, by Markov's inequality and Hölder's inequality, we will have

$$P(\|R_{2,1}\| \geq \epsilon) \quad \leq \quad C_1\sqrt{\frac{K}{n}}\frac{1}{K}\sum_{k=1}^K \sqrt{E\left(\|\sqrt{n}\{\nabla^2_{\theta_k} M_{n,k}(\theta_k^*) - J_k(\theta_k^*)\}\|_2^2\right) E\left(\|\sqrt{n}(\hat{\theta}_k - \theta_k^*)\|_2^2\right)}.$$

From Lemma 7 of Zhang et al. (2013) and Assumption 5 with $v \geq 1$, we know that $E\left(\|\sqrt{n}\{\nabla^2_{\theta_k} M_{n,k}(\theta_k^*) - J_k(\theta_k^*)\}\|_2^2\right) \leq C$. On the other hand, by Lemma 6 of Zhang et al. (2013) and using the event $\mathcal{E}_k$ we can show that $E\left(\|\sqrt{n}(\hat{\theta}_k - \theta_k^*)\|_2^2\right) \leq C_1$. Now since $K = o(N^{1/2}) = o(n)$, we conclude that $R_2 = o_p(1)$. Then we control R_1, which is

$$-(\frac{1}{K}\sum_{k=1}^K H_k(\theta_k^*)^{-1})^{-1}\sqrt{\frac{n}{K}}\sum_{k=1}^K P_k\{\int_0^1 \nabla^2_{\theta_k} M_{n,k}(\theta_k^* + t(\hat{\theta}_k - \theta_k^*))dt - \nabla^2_{\theta_k} M_{n,k}(\theta_k^*)\}(\hat{\theta}_k - \theta_k^*),$$

where P_k is defined in (32). Now with $K = o(n)$, we can similarly prove that $\|R_{1,1}\|_2 = o_p(1)$. ■

Lemma B.5 *Under the same conditions required by Lemma B.4, the following term is asymptotically negligible (i.e. $o_p(1)$):*

$$\sqrt{N}\Big(\sum_{k=1}^K\{\sum_{s=1}^K n_s\hat{H}_s(\hat{\theta}_s)^{-1}\}^{-1}n_k\hat{H}_k(\hat{\theta}_k)^{-1}(\hat{\phi}_k - \phi^*) - \sum_{k=1}^K\{\sum_{s=1}^K n_s H_s(\theta_s^*)^{-1}\}^{-1}n_k H_k(\theta_k^*)^{-1}(\hat{\phi}_k - \phi^*)\Big).$$

Proof Denote the LHS of the above equation as T_2, then we have

$$\begin{aligned}
& \|T_2\|_2 \\
\leq \; & \sqrt{\frac{K}{n}}\Big(\|\{\frac{1}{K}\sum_{s=1}^K \hat{H}_s(\hat{\theta}_s)^{-1}\}^{-1}\|_2\frac{1}{K}\sum_{k=1}^K \|\sqrt{n}(\hat{H}_k(\hat{\theta}_k)^{-1} - H_k(\theta_k^*)^{-1})\|_2\|\sqrt{n}(\hat{\phi}_k - \phi^*)\|_2 \\
& + \frac{1}{K}\sum_{k=1}^K \|H_k(\theta_k^*)^{-1}\|_2\|\sqrt{n}\big(\{\frac{1}{K}\sum_{s=1}^K \hat{H}_s(\hat{\theta}_s)^{-1}\}^{-1} - \{\frac{1}{K}\sum_{s=1}^K H_s(\theta_s^*)^{-1}\}^{-1}\big)\|_2\|\sqrt{n}(\hat{\phi}_k - \phi^*)\|_2\Big) \\
:= \; & \sqrt{\frac{K}{n}}(T_{2,1}^{(1)} + T_{2,1}^{(2)}).
\end{aligned}$$

Since $K = o(n)$, it suffices to show $T_{2,1}^{(1)}$ and $T_{2,1}^{(2)}$ are both $O_p(1)$. Under the event $\mathcal{A}_K$ defined in Equation (51), we have $T_{2,1}^{(2)} I(\mathcal{A}_K) \leq \frac{C}{K}\sum_{k=1}^K\Big(\sqrt{n}\|\hat{\Sigma}_{S,k}(\theta_k^*) - \Sigma_{S,k}(\theta_k^*)\|_2 + \sqrt{n}\|\hat{L}_k(\theta_k^*) - L_k(\theta_k^*)\|_2 + \|\sqrt{n}(\hat{\theta}_k - \theta_k^*)\|_2\Big)\|\sqrt{n}(\hat{\theta}_k - \theta_k^*)\|_2$. Thus for $v \geq 1, v_1 \geq 2$, by Markov's inequality and Cauchy's inequality we have

$$\begin{aligned} P(T_{2,1}^{(2)} > 1, \mathcal{A}_K) &\leq n \max_{1\leq k\leq K}\Bigg(C_1\sqrt{E\left(|\hat{\Sigma}_{S,k}(\theta_k^*) - \Sigma_{S,k}(\theta_k^*)\|_2^2\right) E\left(\|\hat{\theta}_k - \theta_k^*\|_2^2\right)} \\ &\quad + C_2\sqrt{E\left(\|\hat{L}_k(\theta_k^*) - L_k(\theta_k^*)\|_2^2\right) E\left(\|\hat{\theta}_k - \theta_k^*\|_2^2\right)} + C_3^{'}E\left(\|\hat{\theta}_k - \theta_k^*\|_2^2\right)\Bigg) \\ &= \mathcal{O}(1). \end{aligned}$$

Since we have shown $P(\mathcal{A}_K) \to 1$ if $K = o(n^{\bar{v}})$ with $\bar{v} = min\{v, \frac{v_1}{2}\}$, and we have assumed that $K = o(n)$, we can conclude that $T_{2,1}^{(2)} = O_p(1)$. We can similarly show that $T_{2,1}^{(1)} = O_p(1)$. Now we complete the proof. ■

Lemma B.6 *Under Assumptions 1 - 4 and 7 - 8, and Assumption5 with* $v, v_1 \geq 2$ *,*

$$E\|\hat{\theta}_k - \theta_k^* - n_k^{-1}B_k(\theta_k^*)\|_2^2 \leq \frac{2E\|\{\nabla_{\theta_k}\Psi_\theta(\theta_k^*)\}^{-1}\psi_{\theta_k}(X_{k,1};\theta_k^*)\|_2^2}{n_k} + \frac{C_1}{n_k^2}.$$

Proof By the expansion (30) of $\hat{\theta}_k - \theta_k^*$, we have that

$$\begin{aligned} &E\|\hat{\theta}_k - \theta_k^* - \frac{1}{n}B_k(\theta_k^*)\|_2^2 = E\|(-J_k(\theta_k^*)^{-1}\nabla_{\theta_k}M_{n,k}(\theta_k^*) + R_n^{(k)})I(\mathcal{E}_k) - \frac{1}{n}B_k(\theta_k^*) + (\hat{\theta}_k - \theta_k^*)I(\mathcal{E}_k^C)\|_2^2 \\ \leq\ & 2E\|J_k(\theta_k^*)^{-1}\nabla_{\theta_k}M_{n,k}(\theta_k^*)\|_2^2 + 2E\|R_n^{(k)}I(\mathcal{E}_k) - \frac{1}{n}B_k(\theta_k^*) + (\hat{\theta}_k - \theta_k^*)I(\mathcal{E}_k^C)\|_2^2 \\ \leq\ & \frac{2E\|J_k(\theta_k^*)^{-1}\nabla_{\theta_k}M(X_{k,1};\theta_k^*)\|_2^2}{n} + C_1E\|R_n^{(k)}\|_2^2 + \frac{C_2}{n^2}\|B_k(\theta_k^*)\|_2^2 + C_3E\|(\hat{\theta}_k - \theta_k^*)I(\mathcal{E}_k^C)\|_2^2. \end{aligned}$$

We have shown that $E\|R_n^{(k)}\|_2^2 = \mathcal{O}(\frac{1}{n^2})$. For the boundedness of $\|B_k(\theta_k^*)\|_2$, see the proof of Lemma B.9. Besides, we have from Lemma B.2 that

$$\begin{aligned} E\|(\hat{\theta}_k - \theta_k^*)I(\mathcal{E}_k^C)\|_2^2 &\leq \sqrt{E\|\hat{\theta}_k - \theta_k^*\|_2^4 P(\mathcal{E}_k^C)} \\ &\leq \frac{1}{n^{v_2/2}}\sqrt{2E\|(\hat{\theta}_k - \theta_k^*)I(\mathcal{E}_k)\|_2^4 + 2E\|(\hat{\theta}_k - \theta_k^*)I(\mathcal{E}_k^C)\|_2^4} \\ &\leq \frac{1}{n^{v_2/2}}\sqrt{\frac{C_1}{n^2} + \frac{C_2}{n^{v_2}}} = \mathcal{O}(\frac{1}{n^2}) \end{aligned}$$

■

Lemma B.7 *Let* $A_1, A_2, \cdots, A_n \in \mathbb{S}^{p\times p}$*, if* $\forall\Delta \in \mathbb{R}^p$*, we have*

$$\|\begin{pmatrix} vec(A_1)^T \\ vec(A_2)^T \\ \vdots \\ vec(A_n)^T \end{pmatrix}(\Delta \otimes \Delta)\|_2 \leq A\|\Delta\|_2^2.$$

Then $\|\tilde{A}\|_2 \leq \sqrt{pn}A$*, where* $\tilde{A} = (vec(A_1), vec(A_2), \cdots, vec(A_n))^T$.

Proof Since $\tilde{A}(\Delta \otimes \Delta) = (\Delta^T A_1 \Delta, \Delta^T A_2 \Delta, \cdots, \Delta^T A_n \Delta)^T$, $A^2 \|\Delta\|_2^4 \geq \sum_{i=1}^n (\Delta^T A_i \Delta)^2$ which implies $\underset{i \leq n}{max} \|A_i\|_2 \leq A$. On the other hand, for $B = (A_1, A_2, \cdots, A_n) \in \mathbb{R}^{p \times np}$, we have $\|B\|_2^2 = \lambda_{max}(\sum_{i=1}^n A_i A_i^T) \leq \sum_{i=1}^n \lambda_{max}(A_i A_i^T) = \sum_{i=1}^n \|A_i\|^2 \leq nA^2$, which gives $\|\tilde{A}\|_2 = \|\tilde{A}^T\|_2 \leq \sqrt{\sum_{i=1}^n \|vec(A_i)\|_2^2} = \sqrt{\sum_{i=1}^n \|A_i\|_F^2} \leq \sqrt{pn} A$. ■

Let the infeasible debiased weighted distributed estimator be $\hat{\phi}^{IdWD} = \sum_{k=1}^K W_k^* \hat{\phi}_{k,bc}$, where $\hat{\phi}_{k,bc}$ is the first p_1 dimension of $\hat{\theta}_{k,bc}$. We first give a lemma on the MSE bound of this estimator.

Lemma B.8 *Under Assumptions 1 - 4 and 7 - 8, and Assumption5 with* $v, v_1 \geq 4$ *,*

$$E\left(\|\hat{\phi}^{IdWD} - \phi^*\|_2^2\right) \leq \frac{C_1}{nK} + \frac{C_2}{n^2 K} + \frac{C_3}{n^3}. \tag{33}$$

Proof Under the event $\mathcal{E}_k$ defined in the Lemma B.2, we have that

$$\begin{aligned} \mathbf{0} &= \nabla_{\theta_k} M_{n,k}(\theta_k^*) + \nabla_{\theta_k}^2 M_{n,k}(\theta_k^*)\Delta_k + \frac{1}{2}\{\int_0^1 \nabla_{\theta_k}^3 M_{n,k}(\theta_k^* + t\Delta_k)dt\}(\Delta_k \otimes \Delta_k) \\ &= \nabla_{\theta_k} M_{n,k}(\theta_k^*) + \nabla_{\theta_k}^2 M_k(\theta_k^*)\Delta_k + \frac{1}{2}\nabla_{\theta_k}^3 M_k(\theta_k^*)(\Delta_k \otimes \Delta_k) \\ &\quad + (\nabla_{\theta_k}^2 M_{n,k}(\theta_k^*) - \nabla_{\theta_k}^2 M_k(\theta_k^*))\Delta_k + \frac{1}{2}\{\int_0^1 \nabla_{\theta_k}^3 M_{n,k}(\theta_k^* + t\Delta_k)dt - \nabla_{\theta_k}^3 M_k(\theta_k^*)\}(\Delta_k \otimes \Delta_k). \end{aligned}$$

Recall that we have denoted $J_k(\theta_k) = \nabla_{\theta_k}^2 M_k(\theta_k)$, solve for the above equation and we will have

$$\begin{aligned} \Delta_k &= -J_k(\theta_k^*)^{-1}\nabla_{\theta_k} M_{n,k}(\theta_k^*) - J_k(\theta_k^*)^{-1}(\nabla_{\theta_k}^2 M_{n,k}(\theta_k^*) \\ &\quad - \nabla_{\theta_k}^2 M_k(\theta_k^*))\Delta_k - \frac{1}{2}J_k(\theta_k^*)^{-1}\nabla_{\theta_k}^3 M_k(\theta_k^*)(\Delta_k \otimes \Delta_k) \\ &\quad - \frac{1}{2}J_k(\theta_k^*)^{-1}\{\int_0^1 \nabla_{\theta_k}^3 M_{n,k}(\theta_k^* + t\Delta_k)dt - \nabla_{\theta_k}^3 M_k(\theta_k^*)\}(\Delta_k \otimes \Delta_k). \end{aligned} \tag{34}$$

Now we first derive the MSE bound of the pseudo debiased weighted distributed estimator (with known weights and bias correction term) $\hat{\phi}^{pdWD}$: $\hat{\phi}^{pdWD} = \sum_{k=1}^K W_k(\theta_k^*)(\hat{\phi}_k - \frac{1}{n}B_k^1(\theta_k^*))$. Recall the definition of $W_k(\theta_k^*)$, we have that $\|\hat{\phi}^{pdWD} - \phi^*\|_2^2 \leq C\|\frac{1}{K}\sum_{k=1}^K H_k(\theta_k^*)^{-1}(\hat{\phi}_k - \phi^* - \frac{1}{n}B_k^1(\theta_k^*))\|_2^2 = C\|\frac{1}{K}\sum_{k=1}^K \tilde{H}_k(\theta_k^*)(\Delta_k - \frac{1}{n}B_k(\theta_k^*))\|_2^2$, where $\tilde{H}_k(\theta_k^*) = \begin{pmatrix} H_k(\theta_k^*)^{-1} & \mathbf{0} \end{pmatrix}$ and thus $\|\tilde{H}_k(\theta_k^*)\|_2 = \|H_k(\theta_k^*)\|_2$. Denote

$$\begin{aligned} \Omega_{k,1} &= (\nabla_{\theta_k}^2 M_{n,k}(\theta_k^*) - \nabla_{\theta_k}^2 M_k(\theta_k^*))\Delta_k - \frac{1}{n}Ev_{1,k}(\theta_k^*)d_{1,k}(\theta_k^*), \\ \Omega_{k,2} &= (\Delta_k \otimes \Delta_k) - \frac{1}{n}Ed_{1,k}(\theta_k^*) \otimes d_{1,k}(\theta_k^*) \quad \text{and} \\ \Omega_{k,3} &= \{\int_0^1 \nabla_{\theta_k}^3 M_{n,k}(\theta_k^* + t\Delta_k)dt - \nabla_{\theta_k}^3 M_k(\theta_k^*)\}(\Delta_k \otimes \Delta_k), \end{aligned} \tag{35}$$

then

$$\Delta_k - \frac{1}{n}B_k(\theta_k^*) = \{\frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*) + Q_k(\theta_k^*)(\Omega_{k,1} + \frac{1}{2}H_{3,k}(\theta_k^*)\Omega_{k,2} + \Omega_{k,3})\}I(\mathcal{E}_k) + \Delta_k I(\mathcal{E}_k^C). \tag{36}$$

Furthermore, we denote

$$\begin{aligned}\Omega_{k,1} &= \{(\nabla^2_{\theta_k} M_{n,k}(\theta_k^*) - \nabla^2_{\theta_k} M_k(\theta_k^*))(\Delta_k - \frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*))\} \\ &\quad + \{(\nabla^2_{\theta_k} M_{n,k}(\theta_k^*) - \nabla^2_{\theta_k} M_k(\theta_k^*))\frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*) - \frac{1}{n} E v_{1,k}(\theta_k^*) d_{1,k}(\theta_k^*)\} \\ &:= \Omega_{k,1}^{(1)} + \Omega_{k,1}^{(2)}.\end{aligned}$$

For $\Omega_{k,1}^{(1)}$, under the event $\mathcal{E}_k$ and by Taylor's expansion we have that $\Delta_k - \frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*) = Q_k(\theta_k^*)(\int_0^1 \nabla^2_{\theta_k} M_{n,k}(\theta_k^* + t\Delta_k)dt - \nabla^2_{\theta_k} M_k(\theta_k^*))\Delta_k$, thus $\|\Delta_k - \frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*)\|_2^2 I_{\mathcal{E}_k} \leq C_1(\frac{1}{n}\sum_{i=1}^n G(X_{k,i}))^2\|\Delta_k\|_2^4 I_{\mathcal{E}_k} + C_2\|\nabla^2_{\theta_k} M_{n,k}(\theta_k^*) - \nabla^2_{\theta_k} M_k(\theta_k^*)\|_2^2\|\Delta_k\|_2^2 \leq C_3\|\Delta_k\|_2^4 + C_2\|\nabla^2_{\theta_k} M_{n,k}(\theta_k^*) - \nabla^2_{\theta_k} M_k(\theta_k^*)\|_2^2\|\Delta_k\|_2^2$. Now using Hölder's inequality we can show that

$$\begin{aligned}E\left(\|\Omega_{k,1}^{(1)}\|_2^2 I_{\mathcal{E}_k}\right) &\leq C_3\{E\left((\frac{1}{n}\sum_{i=1}^n G(X_{k,i}))^6\right)\}^{1/3}\{E\left(\|\nabla^2_{\theta_k} M_{n,k}(\theta_k^*) - \nabla^2_{\theta_k} M_k(\theta_k^*)\|_2^6\right)\}^{2/3}\{E\left(\|\Delta_k\|_2^8\right)\}^{1/4} \\ &\quad + C_2\{E\left(\|\nabla^2_{\theta_k} M_{n,k}(\theta_k^*) - \nabla^2_{\theta_k} M_k(\theta_k^*)\|_2^6\right)\}^{2/3}\{E\left(\|\Delta_k\|_2^6\right)\}^{1/3} \leq \frac{C}{n^3}.\end{aligned}$$

For $\Omega_{k,1}^{(2)}$, by the independence of the data samples, it is easy to show that $E\left(\Omega_{k,1}^{(2)}\right) = 0$, now we show that $E\|\Omega_{k,1}^{(2)}\|_2^2 \leq \frac{C}{n^2}$. Denote $e_{ij} = Ev_{i,k}(\theta_k^*)d_{j,k}(\theta_k^*)$, then we have

$$E\left(\|\Omega_{k,1}^{(2)}\|_2^2\right) = \frac{1}{n^4}\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^n\sum_{t=1}^n E(v_{i,k}(\theta_k^*)d_{j,k}(\theta_k^*) - e_{ij})^T(v_{s,k}(\theta_k^*)d_{t,k}(\theta_k^*) - e_{st}). \quad (37)$$

By a conditioning argument and independence among samples, it is straightforward to show that if the set $\{i,j,s,t\}$ has three or four unique elements, then $E(v_{i,k}(\theta_k^*)d_{i,k}(\theta_k^*) - e_{ij})^T(v_{s,k}(\theta_k^*)d_{t,k}(\theta_k^*) - e_{ij}) = 0$. Thus the RHS of Equation (37) has at most $\mathcal{O}(n^2)$ non-zero elements and each of those non-zero elements can be bounded using Hölder's inequality. Thus $E\|n\Omega_{k,1}^{(2)}\|_2^2 \leq C$. By similar argument we can show that $E\|n\Omega_{k,1}^{(2)}\|_2^4 \leq C$ if $v, v_1 \geq 4$. By independence among different $\Omega_{k,1}^{(2)}$, we can directly show that $E\|\frac{1}{K}\sum_{k=1}^K \tilde{H}_k(\theta_k^*)Q_k(\theta_k^*)\Omega_{k,1}^{(2)}\|_2^2 \leq \frac{C}{n^2K}$. For $\Omega_{k,2}$ appeared in Equation (35), $\Omega_{k,2} = \{(\Delta_k\otimes\Delta_k) - (\frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*))\otimes(\frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*))\} + \{(\frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*)) \otimes (\frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*)) - \frac{1}{n}Ed_{1,k}(\theta_k^*) \otimes d_{1,k}(\theta_k^*)\} = \Omega_{k,2}^{(1)} + \Omega_{k,2}^{(2)}$. We can show that $E\|\frac{1}{K}\sum_{k=1}^K \tilde{H}_k(\theta_k^*)Q_k(\theta_k^*)H_{3,k}(\theta_k^*)\Omega_{k,2}^{(2)}\|_2^2 \leq \frac{C}{n^2K}$ using similar argument as that when we bound $\Omega_{k,1}^{(2)}$. On the other hand, since $\|a\otimes a - b\otimes b\|_2^2 \leq 2\|a-b\|_2^2(\|a\|_2^2 + \|b\|_2^2)$, we have $E\left(\|\Omega_{k,2}^{(1)}\|_2^2 1_{\mathcal{E}_k}\right) \leq 2E\left(\|\Delta_k - \frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*)\|_2^2 I_{\mathcal{E}_k}(\|\Delta_k\|_2^2 + \|\frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*)\|_2^2)\right) \leq CE\left((\|\Delta_k\|_2^4 + \|\nabla^2_{\theta_k} M_{n,k}(\theta_k^*) - \nabla^2_{\theta_k} M_k(\theta_k^*)\|_2^2\|\Delta_k\|_2^2)(\|\Delta_k\|_2^2 + \|\frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*)\|_2^2)\right) \leq \frac{C}{n^3}$. The last inequality follows from a direct application of Hölder's inequality.

For $\Omega_{k,3}$ in (35), following the proof of Lemma 12 in Zhang et al. (2013), we can show that $\|\hat{H}_{3,k}(\theta_k^*) - H_{3,k}(\theta_k^*)\|_2 \leq \frac{C}{n}\sum_{i=1}^n(G(X_{k,i}) + G)$, which shows $E\left(\|\hat{H}_{3,k}(\theta_k^*) - H_{3,k}(\theta_k^*)\|_2^{2v}\right) =$

$\mathcal{O}(\frac{1}{n^v})$. Combined with the lipschitz continuity of $\nabla^3_{\theta_k} M(X_{k,1},\theta_k)$ with respect to θ_k in U_k, we can show that $E\left(\|\Omega_{k,3}\|_2^2 1_{\mathcal{E}_k}\right) \leq \frac{C}{n^3}$. For $\Delta_k I(\mathcal{E}_k^C)$, using Hölder's inequality we have that

$$E\left(\|\Delta_k I(\mathcal{E}_k^C)\|_2^2\right) \leq \sqrt{E\left(\|\Delta_k\|_2^4\right) P(\mathcal{E}_k^C)} = \mathcal{O}(\frac{1}{n^{v_2/2+1}}).$$

Combined with $E\left(\|\frac{1}{nK}\sum_{k=1}^K\sum_{i=1}^n \tilde{H}_k(\theta_k^*)d_{i,k}(\theta_k^*)\|_2^2\right) = \mathcal{O}((nK)^{-1})$, which is a direct application of Lemma 7 in Zhang et al. (2013), we have

$$E\left(\|\hat{\phi}^{pdWD} - \phi^*\|_2^2\right) \leq \frac{C_1}{nK} + \frac{C_2}{n^2K} + \frac{C_3}{n^3} + \frac{C_4}{n^{v_2/2+1}}.$$

■

Lemma B.9 *Under Assumptions 1 - 4 and 7 - 8, and Assumption 5 with* $v, v_1 \geq 4$ *,*

$$E\left(\|\hat{B}_k(\hat{\theta}_k)I_{\mathcal{E}_{k,bc}} - B_k(\theta_k^*)\|_2^2\right) \leq \frac{C}{n_k}.$$

Proof Denote $\Delta_k = \hat{\theta}_k - \theta_k^*$. By the definition of the event $\mathcal{E}_{k,bc}$, we already have that $\|\hat{B}_k(\hat{\theta}_k)I_{\mathcal{E}_{k,bc}} - B_k(\theta_k^*)\|_2^2 \leq Cn^2$. Below we first control the $\|Q_k(\theta_k^*) - \hat{Q}_k(\hat{\theta}_k)\|_2$ term. Note that $Q_k(\theta_k), \hat{Q}_k(\theta_k)$ are exactly $-L_k^{-1}(\theta_k), -\hat{L}_k(\theta_k)^{-1}$ defined in the proof of Theorem 3, thus under the event $\{\|\hat{L}_k(\hat{\theta}_k) - L_k(\theta_k^*)\|_2 \leq \frac{\rho_-}{2}\}$, we have $\|Q_k(\theta_k^*) - \hat{Q}_k(\hat{\theta}_k)\|_2 \leq \frac{2}{\rho_-^2}\|\hat{L}_k(\hat{\theta}_k) - L_k(\theta_k^*)\|_2$. Besides,

$$\|\hat{L}_k(\hat{\theta}_k) - L_k(\theta_k^*)\|_2 \leq \frac{1}{n}\sum_{i=1}^n G(X_{k,i})\|\Delta_k\|_2 + \|\hat{L}_k(\theta_k^*) - L_k(\theta_k^*)\|_2. \tag{38}$$

So if we define $\mathcal{E}_{Q,k} = \{\|\Delta_k\|_2 \leq \frac{\rho_-}{8G}, G_{n,k} \leq 2G, \|\hat{L}_k(\theta_k^*) - L_k(\theta_k^*)\|_2 < \frac{\rho_-}{4}\}$, then under this event we have $\|\hat{Q}_k(\hat{\theta}_k)\|_2 \leq \|Q_k(\theta_k^*) - \hat{Q}_k(\hat{\theta}_k)\|_2 + \|Q_k(\theta_k^*)\|_2 \leq \frac{1}{\rho_-} + \rho_-$. Using union bound and Markov's inequality it is easy to show $P(\mathcal{E}_{Q,k}) = 1 - \mathcal{O}(\frac{1}{n^{v_2}})$ with $v_2 = min\{v, v_1\}$. Thus we have $E\left(1_{\mathcal{E}_{Q,k}}\|Q_k(\theta_k^*) - \hat{Q}_k(\hat{\theta}_k)\|_2^2\right) \leq C_1E\left(\|\Delta_k\|_2^2\right) + C_2E\left(\|\hat{L}_k(\theta_k^*) - L_k(\theta_k^*)\|_2^2\right) = \mathcal{O}(\frac{1}{n})$. It is noted that

$$\begin{aligned}
& \|\hat{B}_k(\hat{\theta}_k) - B_k(\theta_k^*)\|_2^2 \\
\leq\ & 2\|\hat{Q}_k(\hat{\theta}_k)\frac{1}{n}\sum_{i=1}^n \hat{v}_{i,k}(\hat{\theta}_k)\hat{d}_{i,k}(\hat{\theta}_k) - Q_k(\theta_k^*)Ev_{i,k}(\theta_k^*)d_{i,k}(\theta_k^*)\|_2^2 \\
& + \frac{1}{2}\|\hat{Q}_k(\hat{\theta}_k)\hat{H}_{3,k}(\hat{\theta}_k)\frac{1}{n}\sum_{i=1}^n \hat{d}_{i,k}(\hat{\theta}_k)\otimes\hat{d}_{i,k}(\hat{\theta}_k) - Q_k(\theta_k^*)H_{3,k}(\theta_k^*)Ed_{i,k}(\theta_k^*)\otimes d_{i,k}(\theta_k^*)\|_2^2 \\
:=\ & 2\Omega_{k,1} + \frac{1}{2}\Omega_{k,2}.
\end{aligned}$$

Then we can bound those two terms respectively. For $\Omega_{k,1}$, under the event $\mathcal{E}_{Q,k}$ we have

$$\begin{aligned}
& \Omega_{k,1} 1_{\mathcal{E}_{Q,k}} \\
\leq\ & 2\big(\|\hat{Q}_k(\hat{\theta}_k)\{\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{i,k}(\hat{\theta}_k)\hat{d}_{i,k}(\hat{\theta}_k) - Ev_{i,k}(\theta_k^*)d_{i,k}(\theta_k^*)\}\|_2^2 1_{\mathcal{E}_{Q,k}} \\
+\ & 1_{\mathcal{E}_{Q,k}}\|Q_k(\theta_k^*) - \hat{Q}_k(\hat{\theta}_k)\|_2^2 \|Ev_{i,k}(\theta_k^*)d_{i,k}(\theta_k^*)\|_2^2\big) \\
\leq\ & C_1\|\frac{1}{n}\sum_{i=1}^{n}(\hat{v}_{i,k}(\hat{\theta}_k)\hat{d}_{i,k}(\hat{\theta}_k) - \hat{v}_{i,k}(\theta_k^*)d_{i,k}(\theta_k^*))\|_2^2 1_{\mathcal{E}_{Q,k}} + C_2\|\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{i,k}(\theta_k^*)d_{i,k}(\theta_k^*) - Ev_{i,k}(\theta_k^*)d_{i,k}(\theta_k^*)\|_2^2 \\
& + C_3 1_{\mathcal{E}_{Q,k}}\|Q_k(\theta_k^*) - \hat{Q}_k(\hat{\theta}_k)\|_2^2. \qquad (39)
\end{aligned}$$

By Lemma 7 in Zhang et al. (2013), we have $E\left(\|\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{i,k}(\theta_k^*)d_{i,k}(\theta_k^*) - Ev_{i,k}(\theta_k^*)d_{i,k}(\theta_k^*)\|_2^2\right) = \mathcal{O}(\frac{1}{n})$. Besides, using previous results we can also prove that $E\left(1_{\mathcal{E}_{Q,k}}\|Q_k(\theta_k^*) - \hat{Q}_k(\hat{\theta}_k)\|_2^2\right) = \mathcal{O}(\frac{1}{n})$. So we only need to show $E\left(\Omega_{k,1}^{(1)}\right) = \mathcal{O}(\frac{1}{n})$, where $\Omega_{k,1}^{(1)} = \|\frac{1}{n}\sum_{i=1}^{n}(\hat{v}_{i,k}(\hat{\theta}_k)\hat{d}_{i,k}(\hat{\theta}_k) - \hat{v}_{i,k}(\theta_k^*)d_{i,k}(\theta_k^*))\|_2^2 1_{\mathcal{E}_{Q,k}}$.

$$\begin{aligned}
\Omega_{k,1}^{(1)} &\leq 2\|\frac{1}{n}\sum_{i=1}^{n}(\hat{v}_{i,k}(\hat{\theta}_k) - \hat{v}_{i,k}(\theta_k^*))\hat{d}_{i,k}(\hat{\theta}_k)\|_2^2 1_{\mathcal{E}_{Q,k}} + 2\|\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{i,k}(\theta_k^*)(\hat{d}_{i,k}(\hat{\theta}_k) - d_{i,k}(\theta_k^*))\|_2^2 1_{\mathcal{E}_{Q,k}} \\
&:= 2(\Omega_{k,1}^{(2)} + \Omega_{k,1}^{(3)}).
\end{aligned}$$

Since $\|\hat{v}_{i,k}(\hat{\theta}_k) - \hat{v}_{i,k}(\theta_k^*)\|_2 \leq G(X_{k,i})\|\Delta_k\|_2$,

$$\Omega_{k,1}^{(2)} \leq C\|\Delta_k\|_2^2(\frac{1}{n}\sum_{i=1}^{n}G(X_{k,i})\|\nabla_{\theta_k}M(X_{k,i};\hat{\theta}_k)\|_2)^2.$$

Under the event $\mathcal{E}_k$, we have the expansion

$$\nabla_{\theta_k}M(X_{k,i};\hat{\theta}_k) = \nabla_{\theta_k}M(X_{k,i};\theta_k^*) + \int_0^1 \nabla_{\theta_k}^2 M(X_{k,i};\theta_k^* + t\Delta_k)dt\Delta_k,$$

which implies

$$\begin{aligned}
& \|\nabla_{\theta_k}M(X_{k,i};\hat{\theta}_k)\|_2 \\
\leq\ & \|\nabla_{\theta_k}M(X_{k,i};\theta_k^*)\|_2 + C_1 G(X_{k,i}) + C\|\nabla_{\theta_k}^2 M(X_{k,i};\theta_k^*) - \nabla_{\theta_k}^2 M_k(\theta_k^*)\|_2 + C_3. \quad (40)
\end{aligned}$$

Since $v_2 \geq 4$, by Hölder's inequality with three terms we have

$$\begin{aligned}
E\left(\Omega_{k,1}^{(2)} 1_{\mathcal{E}_k}\right) &\leq CE\left(\|\Delta_k\|_2^2 \frac{1}{n}\sum_{i=1}^{n}G^2(X_{k,i})\|\nabla_{\theta_k}M(X_{k,i};\hat{\theta}_k)\|_2^2 1_{\mathcal{E}_k}\right) \\
&\leq C\frac{1}{n}\sum_{i=1}^{n}\{E\left(\|\Delta_k\|_2^6\right)EG\left((X_{k,i})^6\right)E\left(\|\nabla_{\theta_k}M(X_{k,i};\hat{\theta}_k)1_{\mathcal{E}_k}\|_2^6\right)\}^{1/3} = \mathcal{O}(\frac{1}{n}).
\end{aligned}$$

We can similarly show that $E\left(\Omega_{k,1}^{(3)} 1_{\mathcal{E}_k}\right) = \mathcal{O}(\frac{1}{n})$. Now we turn to bound $\Omega_{k,2}$. Following the proof of Lemma 12 in Zhang et al. (2013), we can show that

$$\|\hat{H}_{3,k}(\theta_k^*) - H_{3,k}(\theta_k^*)\|_2 \le \frac{1}{n}\sum_{i=1}^n (G(X_{k,i}) + G), \tag{41}$$

which implies $E\left(\|\hat{H}_{3,k}(\theta_k^*) - H_{3,k}(\theta_k^*)\|_2^{2v}\right) = \mathcal{O}(\frac{1}{n^v})$. Besides, using Assumption 8 and Lemma B.7, under the event $\mathcal{E}_k$ we have

$$\|\hat{H}_{3,k}(\hat{\theta}_k) - \hat{H}_{3,k}(\theta_k^*)\|_2 \le \frac{p}{n}\sum_{i=1}^n A(X_{k,i})\|\Delta_k\|_2.$$

If we define the event

$$\mathcal{E}_{third,k} = \{\frac{1}{n}\sum_{i=1}^n G(X_{k,i}) \le 2G, \frac{1}{n}\sum_{i=1}^n A(X_{k,i}) \le 2A, \|\hat{H}_{3,k}(\theta_k^*) - H_{3,k}(\theta_k^*)\|_2 \le 1\}, \tag{42}$$

then $P(\mathcal{E}_{third,k}) = 1 - \mathcal{O}(\frac{1}{n^v})$ and under this event $\|\hat{H}_{3,k}(\hat{\theta}_k)\|_2 \le C$. Now under the event $\mathcal{E}_k \cap \mathcal{E}_{Q,k} \cap \mathcal{E}_{third,k}$ we have that

$$\begin{aligned}
& \Omega_{k,2} 1_{\mathcal{E}_k \cap \mathcal{E}_{Q,k} \cap \mathcal{E}_{third,k}} \\
\le\ & 2\|\hat{Q}_k(\hat{\theta}_k)\hat{H}_{3,k}(\hat{\theta}_k)\frac{1}{n}\sum_{i=1}^n \hat{d}_{i,k}(\hat{\theta}_k) \otimes \hat{d}_{i,k}(\hat{\theta}_k) - Q_k(\theta_k^*)H_{3,k}(\theta_k^*)\frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*) \otimes d_{i,k}(\theta_k^*)\|_2^2 1_{\mathcal{E}_k \cap \mathcal{E}_{Q,k} \cap \mathcal{E}_{third,k}} \\
& + C\|\frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*) \otimes d_{i,k}(\theta_k^*) - E d_{1,k}(\theta_k^*) \otimes d_{1,k}(\theta_k^*)\|_2^2 1_{\mathcal{E}_k \cap \mathcal{E}_{Q,k}} \\
\le\ & C_1\|(\hat{Q}_k(\hat{\theta}_k)\hat{H}_{3,k}(\hat{\theta}_k) - Q_k(\theta_k^*)H_{3,k}(\theta_k^*))\frac{1}{n}\sum_{i=1}^n \hat{d}_{i,k}(\hat{\theta}_k) \otimes \hat{d}_{i,k}(\hat{\theta}_k)\|_2^2 1_{\mathcal{E}_k \cap \mathcal{E}_{Q,k} \cap \mathcal{E}_{third,k}} \\
& + C_2\|\frac{1}{n}\sum_{i=1}^n \hat{d}_{i,k}(\hat{\theta}_k) \otimes \hat{d}_{i,k}(\hat{\theta}_k) - \frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*) \otimes d_{i,k}(\theta_k^*)\|_2^2 1_{\mathcal{E}_k \cap \mathcal{E}_{Q,k}} \\
& + C\|\frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*) \otimes d_{i,k}(\theta_k^*) - E d_{1,k}(\theta_k^*) \otimes d_{1,k}(\theta_k^*)\|_2^2 \\
:=\ & C_1\Omega_{k,2}^{(1)} + C_2\Omega_{k,2}^{(2)} + C\Omega_{k,2}^{(3)}
\end{aligned}$$

Using lemma 7 in Zhang et al. (2013), we have $E\left(\Omega_{k,2}^{(3)}\right) = \mathcal{O}(\frac{1}{n})$. Now we consider $\Omega_{k,2}^{(1)}$. Note

$$\begin{aligned}
& \Omega_{k,2}^{(1)} \\
\le\ & C(\|\hat{H}_{3,k}(\hat{\theta}_k) - H_{3,k}(\theta_k^*)\|_2^2 1_{\mathcal{E}_{third,k}} + \|\hat{Q}_k(\hat{\theta}_k) - Q_k(\theta_k^*)\|_2^2 1_{\mathcal{E}_{Q,k}})(\frac{1}{n}\sum_{i=1}^n \|\nabla_{\theta_k} M(X_{k,i};\hat{\theta}_k)\|_2^2 1_{\mathcal{E}_k})^2,
\end{aligned}$$

$$\begin{aligned}
E\left(\Omega_{k,2}^{(1)}\right) \ \le\ & C(\sqrt{E\left(\|\hat{H}_{3,k}(\hat{\theta}_k) - H_{3,k}(\theta_k^*)\|_2^4 1_{\mathcal{E}_{third,k}}\right)} + \sqrt{E\left(\|\hat{Q}_k(\hat{\theta}_k) - Q_k(\theta_k^*)\|_2^4 1_{\mathcal{E}_{Q,k}}\right)}) \cdot \\
& \sqrt{E(\frac{1}{n}\sum_{i=1}^n \|\nabla_{\theta_k} M(X_{k,i};\hat{\theta}_k)\|_2^2 1_{\mathcal{E}_k})^4} = \mathcal{O}(\frac{1}{n}).
\end{aligned}$$

Besides, we can also show that $E\left(\Omega_{k,2}^{(2)}\right) = \mathcal{O}(\frac{1}{n})$. In summary, now we have shown that $E\left(\|\hat{B}(\hat{\theta}_k) - B_k(\theta_k^*)\|_2^2 1_{\mathcal{E}_k \cap \mathcal{E}_{Q,k} \cap \mathcal{E}_{third,k}}\right) = \mathcal{O}(\frac{1}{n})$. If we define the event $\tilde{\mathcal{E}}_k = \mathcal{E}_k \cap \mathcal{E}_{Q,k} \cap \mathcal{E}_{third,k}$, then we can show that $P(\tilde{\mathcal{E}}_k) = 1 - \frac{1}{n^{v_2}}$. Besides, by subadditivity of the probability measure and Markov's inequality, we have that

$$
\begin{aligned}
P(\mathcal{E}_{k,bc}^c) &\leq P(\frac{1}{n}\|\hat{B}_k(\hat{\theta}_k) - B_k(\theta_k^*)\|_2 > \frac{r}{2}) + P(\|\hat{\theta}_k - \theta_k^*\|_2 + \frac{1}{n}\|B_k(\theta_k^*)\|_2 > \frac{r}{2}) \\
&\leq P(\frac{1}{n}\|\hat{B}_k(\hat{\theta}_k) - B_k(\theta_k^*)\|_2 > \frac{r}{2}, \tilde{\mathcal{E}}_k) + P(\tilde{\mathcal{E}}_k^c) + \frac{C}{n^{v_2}} \leq \frac{C_1}{n^3} + \frac{C_2}{n^{v_2}}.
\end{aligned}
$$

Thus $E\|\hat{B}_k(\hat{\theta}_k)1_{\mathcal{E}_{k,bc}} - B_k(\theta_k^*)\|_2^2 \leq 2E\|\hat{B}_k(\hat{\theta}_k) - B_k(\theta_k^*)\|_2^2 1_{\tilde{\mathcal{E}}_k \cap \mathcal{E}_{k,bc}} + 2E\|B_k(\theta_k^*)\|_2^2 1_{\mathcal{E}_{k,bc}^C \cap \tilde{\mathcal{E}}_k} + C_1 n^2 \frac{1}{n^{v_2}} \leq \frac{C_1}{n} + \frac{C_2}{n^3} + \frac{C_3}{n^{v_2-2}} + \frac{C_4}{n^{v_2}} = \mathcal{O}(\frac{1}{n})$. The result follows. ■

The asymptotic normality of the IdWD estimator is established in the following lemma.

Lemma B.10 *Under Assumptions 1 - 4 and 7 - 8, and Assumption 5 with* $v, v_1 \geq 4$ *, if* $K = o(n^2)$*,*

$$
(\hat{\phi}^{IdWD} - \phi^*)^T \{\sum_{k=1}^K n_k H_k(\theta_k^*)^{-1}\}(\hat{\phi}^{IdWD} - \phi^*) \xrightarrow{d} \chi_{p_1}^2.
$$

Proof Note that

$$
\begin{aligned}
&\hat{\phi}^{IdWD} - \phi^* \qquad (43)\\
=\ & \{\frac{1}{K}\sum_{k=1}^K H_k(\theta_k^*)^{-1}\}^{-1} \frac{1}{K}\sum_{k=1}^K \tilde{H}_k(\theta_k^*)\big((\Delta_k - \frac{1}{n}B_k(\theta_k^*)) + (\frac{1}{n}B_k(\theta_k^*)) - \frac{1}{n}\hat{B}_k(\hat{\theta}_k)1_{\mathcal{E}_{k,bc}})\big).
\end{aligned}
$$

Since we have shown that $E\|\frac{1}{n}B_k(\theta_k^*) - \frac{1}{n}\hat{B}_k(\hat{\theta}_k)1_{\mathcal{E}_{k,bc}}\|_2^2 = \mathcal{O}(\frac{1}{n^3})$, by Markov's inequality we can show that $\sqrt{N}(\frac{1}{n}B_k(\theta_k^*)) - \frac{1}{n}\hat{B}_k(\hat{\theta}_k)1_{\mathcal{E}_{k,bc}}) = o_p(1)$ when $K = o(n^2)$. So we consider the first part in the RHS of the above equation. Following (36),

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^K \tilde{H}_k(\theta_k^*)(\Delta_k - \frac{1}{n}B_k(\theta_k^*)) \qquad (44)\\
=\ & \frac{1}{Kn}\sum_{k=1}^K\sum_{i=1}^n \tilde{H}_k(\theta_k^*)d_{i,k}(\theta_k^*) + \frac{1}{K}\sum_{k=1}^K \tilde{H}_k(\theta_k^*)\{Q_k(\theta_k^*)(\Omega_{k,1} + \frac{1}{2}H_{3,k}(\theta_k^*)\Omega_{k,2} + \Omega_{k,3})\}I(\cap_{k=1}^K \mathcal{E}_k) \\
&+ \frac{1}{K}\sum_{k=1}^K \tilde{H}_k(\theta_k^*)(\Delta_k + \frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*))I((\cap_{k=1}^K \mathcal{E}_k)^c).
\end{aligned}
$$

Using results in the proof of Lemma B.8, when $K = o(n^2)$, $\sqrt{N}\frac{1}{K}\sum_{k=1}^K \tilde{H}_k(\theta_k^*)\{Q_k(\theta_k^*)(\Omega_{k,1} + \frac{1}{2}H_{3,k}(\theta_k^*)\Omega_{k,2} + \Omega_{k,3})\} = o_p(1)$, where $\Omega_{k,i}$ for $1 \leq i \leq 3$ is defined in (35). Besides, since $P((\cap_{k=1}^K \mathcal{E}_k)^c) = \frac{K}{n^{v_2}}$ for some $v_2 = min\{v, v_1\} \geq 4$, $\sqrt{N}\frac{1}{K}\sum_{k=1}^K \tilde{H}_k(\theta_k^*)(\Delta_k + \frac{1}{n}\sum_{i=1}^n d_{i,k}(\theta_k^*))I((\cap_{k=1}^K \mathcal{E}_k)^c) = o_p(1)$ when $K = o(n^2)$. Then, we can apply Lindeberg-Feller's central limit theorem to establish the asymptotic normality of $\{\frac{1}{K}\sum_{k=1}^K H_k(\theta_k^*)^{-1}\}^{-1/2}\frac{1}{\sqrt{nK}}\sum_{k=1}^K \tilde{H}_k(\theta_k^*)d_{i,k}(\theta_k^*)$, which implies the limiting $\chi_{p_1}^2$ distribution of $\hat{\phi}^{IdWD}$ using (43) and (44). ■

B.2 Proof of Proposition 1

Proof The consistency of the local estimator $\hat{\theta}_k$ is directly implied by Lemma 6 of Zhang et al. (2013). Below we consider the consistency of the global estimator $\hat{\theta}$. Define the objective $\bar{M}(X,\theta) = (1/K)\sum_{k=1}^{K} M(x_k,\theta_k)$, where $X = (x_1^T, x_2^T, ..., x_K^T)^T$. Then we can formulate an statistical optimization problem with the population objective as $\bar{M}^*(\theta) = (1/K)E\left(\sum_{k=1}^{K} M(X_{k,1};\theta_k)\right)$. Denote $\bar{M}_i(X_i,\theta) = (1/K)\sum_{k=1}^{K} M(X_{k,i},\theta_k)$ and $\bar{M}(X,\theta) = (1/n)\sum_{i=1}^{n} \bar{M}_i(X_i,\theta)$. Now we can directly check that $E\left(\|\nabla_\theta \bar{M}_i(X_i,\theta^*)\|_2^{2v_1}\right) \le R^{2v_1}$ and $E\left(\|\nabla_\theta^2 \bar{M}_i(X_i,\theta^*) - \nabla_\theta^2 \bar{M}^*(\theta^*)\|_2^{2v}\right) \le L^{2v}$. Besides, for all $\theta,\theta' \in U$ with $U = \{\theta | \|\theta-\theta^*\|_2 \le \rho\}$, we have

$$\|\nabla_\theta^2 \bar{M}(X,\theta) - \nabla_\theta^2 \bar{M}(X,\theta')\|_2 \le (\frac{1}{K}\sum_{k=1}^{K} G(x_k))\|\theta-\theta'\|_2,$$

where $(1/K)E\left((\sum_{k=1}^{K} G(X_{k,1}))^{2v}\right) \le G^{2v}$. Also note that we can directly prove

$$\nabla_\theta^2 \bar{M}^*(\theta^*) \succeq \begin{pmatrix} \rho_- I_{p_1\times p_1} & \mathbf{0} \\ \mathbf{0} & \frac{\rho_-}{K} I_{Kp_2\times KP_2} \end{pmatrix} \succeq \frac{\rho_-}{K} I_{(p_1+Kp_2)\times(p_1+Kp_2)}.$$

Now we can apply Lemma 6 of Zhang et al. (2013) to obtain the consistency of $\hat{\theta}$. ■

B.3 Proof of Theorem 2

Proof See the proof of Lemma B.4. ■

B.4 Proof of Theorem 2

Proof Note that

$$\|\hat{\phi}-\phi^*\|_2 \quad \le \quad \|(\frac{1}{K}\sum_{k=1}^{K}\hat{H}_k(\hat{\theta}_k)^{-1})^{-1}\|_2 \|\frac{1}{K}\sum_{k=1}^{K}\hat{H}_k(\hat{\theta}_k)^{-1}(\hat{\phi}_k-\phi^*)\|_2.$$

Since $H_k(\theta_k^*)^{-1} \succeq \frac{\rho_-^2}{\rho_\sigma} I_{p_1\times p_1} \triangleq \rho_h I_{p_1\times p_1}$, by Lemma B.3, the event $\mathcal{IH}_K = \{\|\hat{H}_k(\hat{\theta}_k)^{-1} - H_k(\theta_k^*)^{-1}\|_2 \le \frac{c}{2}, k=1,...,K\}$ implies $\|\{\frac{1}{K}\sum_{k=1}^{K}\hat{H}_k(\hat{\theta}_k)^{-1}\}^{-1} - \{\frac{1}{K}\sum_{k=1}^{K}H_k(\theta_k^*)^{-1}\}^{-1}\|_2 \le \frac{2}{c^2}\|\frac{1}{K}\sum_{k=1}^{K}\hat{H}_k(\hat{\theta}_k)^{-1} - \frac{1}{K}\sum_{k=1}^{K}H_k(\theta_k^*)^{-1}\|_2$. Using Lemma B.3 again with $H_k(\theta_k^*) \succeq cI_{p_1\times p_1}$ as assumed in Assumption 7, the event $\mathcal{H}_K = \{\|\hat{H}_k(\hat{\theta}_k) - H_k(\theta_k^*)\|_2 \le \frac{c}{2}, k=1,...,K\}$ implies $\|\hat{H}_k(\hat{\theta}_k)^{-1} - H_k(\theta_k^*)^{-1}\|_2 \le \frac{2}{c^2}\|\hat{H}_k(\hat{\theta}_k) - H_k(\theta_k^*)\|_2, k=1,2,\cdots,K$. Now for any $\epsilon>0$, define $\epsilon_H = min\{\epsilon,c\}/4$, then under the event

$$\mathcal{H}_K^\epsilon = \{\|\hat{H}_k(\hat{\theta}_k) - H_k(\theta_k^*)\|_2 \le \epsilon_H, k=1,...,K\} \tag{45}$$

we have $\|\{\frac{1}{K}\sum_{k=1}^{K}\hat{H}_k(\hat{\theta}_k)^{-1}\}^{-1}-\{\frac{1}{K}\sum_{k=1}^{K}H_k(\theta_k^*)^{-1}\}^{-1}\|_2\leq\epsilon$. Now using the boundedness of $\|\hat{\phi}^{WD}-\phi^*\|$, we have

$$\begin{aligned}
& E\left(\|\hat{\phi}^{WD}-\phi^*\|_2^2\right) \\
\leq\; & C_1E\left(\|\frac{1}{K}\sum_{k=1}^{K}\hat{H}_k(\hat{\theta}_k)^{-1}(\hat{\phi}_k-\phi^*)\|_2^2 I(\mathcal{H}_K^\epsilon)\right) \qquad (46)\\
& +C_2E\left(\|\frac{1}{K}\sum_{k=1}^{K}(\hat{\phi}_k-\phi_k^*)\|_2^2 I(\|\frac{1}{K}\sum_{k=1}^{K}\hat{H}_k(\hat{\theta}_k)^{-1}(\hat{\phi}_k-\phi^*)\|_2^2 I(\mathcal{H}_K^\epsilon)\geq C_4)\right)+C_3P((\mathcal{H}_K^\epsilon)^c).
\end{aligned}$$

To derive the upper bound of $E\left(\|\hat{\phi}^{WD}-\phi^*\|_2^2\right)$, we only need to separately bound the three terms on the RHS of (46). Let us first consider bounding $P((\mathcal{H}_K^\epsilon)^c)$. Denote

$$\begin{aligned}
\hat{L}_k(\theta_k) &= \nabla^2_{\theta_k}M_{n,k}(\theta_k), \quad L_k(\theta_k)=\nabla^2_{\theta_k}M_k(\theta_k),\\
\hat{V}_k(\theta_k) &= \hat{L}_k(\theta_k)^{-1}\hat{\Sigma}_{S,k}(\theta_k)\hat{L}_k(\theta_k)^{-1} \quad \text{and} \quad V_k(\theta_k)=L_k(\theta_k)^{-1}\Sigma_{S,k}(\theta_k)L_k(\theta_k)^{-1}.
\end{aligned}$$

By definition of $\hat{H}_k(\theta_k)$ and the triangle's inequality, we have

$$\|\hat{H}_k(\hat{\theta}_k)-H_k(\theta_k^*)\|_2 \leq \|\hat{V}_k(\hat{\theta}_k)-\hat{V}_k(\theta_k^*)\|_2+\|\hat{V}_k(\theta_k^*)-V_k(\theta_k^*)\|_2. \qquad (47)$$

Hence, we can bound those two terms on the RHS of (47) separately. Note that

$$\begin{aligned}
& \|\hat{V}_k(\theta_k)-V_k(\theta_k)\|_2=\|\hat{L}_k(\theta_k)^{-1}\hat{\Sigma}_{S,k}(\theta_k)\hat{L}_k(\theta_k)^{-1}-L_k(\theta_k)^{-1}\Sigma_{S,k}(\theta_k)L_k(\theta_k)^{-1}\|_2\\
\leq\; & 2(\|\hat{L}_k(\theta_k)^{-1}-L_k(\theta_k)^{-1}\|_2^2+\|L_k(\theta_k)^{-1}\|_2^2)\|\hat{\Sigma}_{S,k}(\theta_k)-\Sigma_{S,k}(\theta_k)\|_2 \qquad (48)\\
& +(\|\hat{L}_k(\theta_k)^{-1}-L_k(\theta_k)^{-1}\|_2+2\|L_k(\theta_k)^{-1}\|_2)\|\Sigma_{S,k}(\theta_k)\|_2\|\hat{L}_k(\theta_k)^{-1}-L_k(\theta_k)^{-1}\|_2.
\end{aligned}$$

Then, under the event $\mathcal{L}_K^\epsilon=\{\|\hat{L}_k(\theta_k^*)-L_k(\theta_k^*)\|_2\leq min\{\epsilon\rho_-^2/2,\rho_-/2\},k=1,...,K\}$ with ρ_- being the lower bound of the eigenvalues of $L_k(\theta_k^*)$ as assumed in Assumption 4, we have

$$\begin{aligned}
& \|\hat{V}_k(\theta_k^*)-V_k(\theta_k^*)\|_2 \qquad (49)\\
\leq\; & 2(\epsilon^2+\frac{1}{\rho_-^2})\|\hat{\Sigma}_{S,k}(\theta_k^*)-\Sigma_{S,k}(\theta_k^*)\|_2+(\epsilon+\frac{2}{\rho_-})\frac{2\rho_\sigma}{\rho_-^2}\|\hat{L}_k(\theta_k^*)-L_k(\theta_k^*)\|_2, k=1,...,K.
\end{aligned}$$

Similar to (48), we have

$$\begin{aligned}
& \|\hat{V}_k(\hat{\theta}_k)-\hat{V}_k(\theta_k^*)\|_2\\
\leq\; & 2(\|\hat{L}_k(\hat{\theta}_k)^{-1}-\hat{L}_k(\theta_k^*)^{-1}\|_2^2+\|\hat{L}_k(\theta_k^*)^{-1}\|_2^2)\|\hat{\Sigma}_{S,k}(\hat{\theta}_k)-\hat{\Sigma}_{S,k}(\theta_k^*)\|_2\\
& +(\|\hat{L}_k(\hat{\theta}_k)^{-1}-\hat{L}_k(\theta_k^*)^{-1}\|_2+2\|\hat{L}_k(\theta_k^*)^{-1}\|_2)\|\hat{\Sigma}_{S,k}(\theta_k^*)\|_2\|\hat{L}_k(\hat{\theta}_k)^{-1}-\hat{L}_k(\theta_k^*)^{-1}\|_2.
\end{aligned}$$

Define an event

$$\begin{aligned}
\mathcal{M}_K = & \Big\{\|\hat{L}_k(\theta_k^*)^{-1}\|_2\leq\frac{2}{\rho_-},\|\hat{\Sigma}_{S,k}(\theta_k^*)\|_2\leq 2\rho_\sigma,\\
& \|\hat{L}_k(\hat{\theta}_k)-\hat{L}_k(\theta_k^*)\|_2\leq min\{\frac{\rho_-}{4},\frac{\epsilon\rho_-^2}{8}\},k=1,...,K\Big\}.
\end{aligned}$$

Then, under this event for $k = 1, 2, ..., K$ we have

$$\begin{aligned} & \|\hat{V}_k(\hat{\theta}_k) - \hat{V}_k(\theta_k^*)\|_2 \\ \leq \; & 2(\epsilon^2 + \frac{4}{\rho_-^2})\|\hat{\Sigma}_{S,k}(\hat{\theta}_k) - \hat{\Sigma}_{S,k}(\theta_k^*)\|_2 + (\epsilon + \frac{4}{\rho_-})\frac{16\rho_\sigma}{\rho_-^2}\|\hat{L}_k(\hat{\theta}_k) - \hat{L}_k(\theta_k^*)\|_2 \\ \leq \; & (C_1 B_{n,k} + C_2 G_{n,k})\|\hat{\theta}_k - \theta_k^*\|_2, \end{aligned} \tag{50}$$

where $B_{n,k} = (1/n)\sum_{i=1}^n B(X_{k,i})$ and $G_{n,k} = (1/n)\sum_{i=1}^n G(X_{k,i})$. Note that $\|\hat{L}_k(\theta_k^*)^{-1}\|_2 \leq \frac{2}{\rho_-}$ and $\|\hat{\Sigma}_{S,k}(\theta_k^*)\| \leq 2\rho_\sigma$ are implied by $\|\hat{L}_k(\theta_k^*) - L_k(\theta_k^*)\|_2 \leq \frac{\rho_-}{2}$ and $\|\hat{\Sigma}_{S,k}(\theta_k^*) - \Sigma_{S,k}(\theta_k^*)\|_2 \leq \frac{\rho_\sigma}{2}$, respectively. Thus, we define the event

$$\begin{aligned} \mathcal{U}_K \;=\; & \Big\{ B_{n,k} \leq 2B, G_{n,k} \leq 2G, \|\hat{L}_k(\theta_k^*) - L_k(\theta_k^*)\|_2 \leq C_1, \\ & \qquad \|\hat{\Sigma}_{S,k}(\theta_k^*) - \Sigma_{S,k}(\theta_k^*)\|_2 \leq C_2, k = 1, ..., K \Big\}, \end{aligned}$$

which satisfies $\mathcal{M}_K \cup \mathcal{L}_K^\epsilon \subset \mathcal{U}_K$ and under $\mathcal{U}_K$ we have $\|\hat{H}_k(\hat{\theta}_k) - H_k(\theta_k^*)\|_2 \leq C\|\hat{\theta}_k - \theta_k^*\|_2 + \frac{\epsilon_H}{2}$. Furthermore, we define the event

$$\mathcal{A}_K = \mathcal{U}_K \cap (\cap_{k=1}^K \mathcal{E}_k) \cap (\{\|\hat{\theta}_k - \theta_k^*\|_2 \leq \frac{\epsilon_H}{2C}, k = 1, ..., K\}). \tag{51}$$

By Lemma 6 in Zhang et al. (2013), under the event $\cap_{k=1}^K \mathcal{E}_k$, the event $\{\|\hat{\theta}_k - \theta_k^*\|_2 \leq \epsilon_H/(2C), k = 1, ..., K\}$ is implied by the event $\{\|\nabla_{\theta_k} M_{n,k}(\theta_k^*)\|_2 \leq (1-\rho)\rho_- \epsilon_H/(4C), k = 1, ...K\}$. Now with the union bound and Lemma 7 in Zhang et al. (2013), we can obtain that

$$P((\mathcal{H}_K^\epsilon)^c) \leq P(\mathcal{A}_K^c) \leq C\frac{K}{n^{\bar{v}}}, \tag{52}$$

where $\bar{v} = min\{v, \frac{v_1}{2}\}$. It is noted that we need the existence of higher-order moments of the score (first-order derivative of the M-function) due to the estimation of its covariance matrix $\Sigma_{S,k}(\theta_k^*)$ in the construction of the estimated optimal weights.

Next we consider bounding $E\|\frac{1}{K}\sum_{k=1}^K \hat{H}_k(\hat{\theta}_k)^{-1}(\hat{\phi}_k - \phi^*)\|_2^2 I(\mathcal{H}_K^\epsilon)$ in (46). Recall the definition of $\mathcal{H}_K^\epsilon$ in (45), we can naturally decompose the event into $\mathcal{H}_K^\epsilon = \cap_{k=1}^K \mathcal{H}_K^{\epsilon,(k)}$, where $\mathcal{H}_K^{\epsilon,(k)} = \{\|\hat{H}_k(\hat{\theta}_k) - H_k(\theta_k^*)\|_2 \leq \epsilon_H\}$. It is noted that under the event $\mathcal{H}_K^{\epsilon,(k)}$, we have

$$\|\hat{H}_k(\hat{\theta}_k)^{-1}\|_2 \leq \frac{2}{c^2}\|\hat{H}_k(\hat{\theta}_k) - H_k(\theta_k^*)\|_2 + \|H_k(\theta_k^*)^{-1}\|_2 \leq C.$$

Since elements of $\{\hat{H}_k(\hat{\theta}_k)^{-1}(\hat{\phi}_k - \phi^*) I(\mathcal{H}_K^{\epsilon,(k)})\}_{k=1}^K$ are independent with one another, we decompose the term as follows:

$$E\left(\|\frac{1}{K}\sum_{k=1}^K \hat{H}_k(\hat{\theta}_k)^{-1}(\hat{\phi}_k - \phi^*)\|_2^2 I(\mathcal{H}_K^\epsilon)\right) \leq \max_{1\leq k\leq K}\left(\frac{C}{K}E\left(\|\hat{\phi}_k - \phi^*\|_2^2\right) + \|E\left(\hat{H}_k(\hat{\theta}_k)^{-1}(\hat{\phi}_k - \phi^*) I(\mathcal{H}_K^{\epsilon,(k)})\right)\|_2^2\right). \tag{53}$$

By the proof of Theorem 1 in Zhang et al. (2013), we have that $E\|\hat{\phi}_k - \phi^*\|_2^2 \le \frac{C_1}{n} + \frac{C_2}{n^2}$. Besides, for the second term in the RHS of inequality (53), we have

$$\begin{aligned}
&\|E\left(\hat{H}_k(\hat{\theta}_k)^{-1}(\hat{\phi}_k - \phi^*)I(\mathcal{H}_K^{\epsilon,(k)})\right)\|_2^2 \\
\le\ & 2\|E\left((\hat{H}_k(\hat{\theta}_k)^{-1} - H_k(\theta_k^*)^{-1})(\hat{\phi}_k - \phi^*)I(\mathcal{H}_K^{\epsilon,(k)})\right)\|_2^2 + 2\|E\left(H_k(\theta_k^*)^{-1}(\hat{\phi}_k - \phi^*)I(\mathcal{H}_K^{\epsilon,(k)})\right)\|_2^2 \\
\le\ & C_1 E\left(\|\hat{H}_k(\hat{\theta}_k)^{-1} - H_k(\theta_k^*)^{-1}\|_2^2 I(\mathcal{H}_K^{\epsilon,(k)})\|\hat{\phi}_k - \phi^*\|_2^2\right) + C\|E\left((\hat{\theta}_k - \theta_k^*)I(\mathcal{H}_K^{\epsilon,(k)})\right)\|_2^2.
\end{aligned}$$

Using Equations (47), (49) and (50), we can show that for $2\bar{v} \ge v' \ge 1$,

$$\begin{aligned}
&\|\hat{H}_k(\hat{\theta}_k)^{-1} - H_k(\theta_k^*)^{-1}\|_2^{v'} I(\mathcal{H}_K^{\epsilon,(k)}) \le C_1\|\hat{V}_k(\hat{\theta}_k) - V_k(\theta_k^*)\|_2^{v'} I(\mathcal{H}_K^{\epsilon,(k)}) \\
\le\ & C_2(\|\hat{\Sigma}_{S,k}(\theta_k^*) - \Sigma_{S,k}(\theta_k^*)\|_2^{v'} + \|\hat{L}_k(\theta_k^*) - L_k(\theta_k^*)\|_2^{v'} + \|\hat{\theta}_k - \theta_k^*\|_2^{v'}).
\end{aligned}$$

We immediately have $E\left(\|\hat{H}_k(\hat{\theta}_k)^{-1} - H_k(\theta_k^*)^{-1}\|_2^{v'} I(\mathcal{H}_K^{\epsilon,(k)})\right) = \mathcal{O}(1/n^{v'/2})$. Thus by Hölder's inequality we have

$$\begin{aligned}
&E\left(\|\hat{H}_k(\hat{\theta}_k)^{-1} - H_k(\theta_k^*)^{-1}\|_2^2 I(\mathcal{H}_K^{\epsilon,(k)})\|\hat{\phi}_k - \phi^*\|_2^2\right) \\
\le\ & \sqrt{E\left(\|\hat{H}_k(\hat{\theta}_k)^{-1} - H_k(\theta_k^*)^{-1}\|_2^4 I(\mathcal{H}_K^{\epsilon,(k)})\right) E\left(\|\hat{\phi}_k - \phi^*\|_2^4\right)} = \mathcal{O}(\frac{1}{n^2}) + \mathcal{O}(\frac{1}{n^3}).
\end{aligned}$$

On the other hand, we have that

$$\begin{aligned}
\|E\left((\hat{\phi}_k - \phi^*)I(\mathcal{H}_K^{\epsilon,(k)})\right)\|_2^2 &\le 2\|E(\hat{\phi}_k - \phi^*)\|_2^2 + 2\|E\left((\hat{\phi}_k - \phi^*)(1 - I(\mathcal{H}_K^{\epsilon,(k)}))\right)\|_2^2 \\
&\le 2\|E(\hat{\phi}_k - \phi^*)\|_2^2 + 2\sqrt{E\left(\|\hat{\phi}_k - \phi^*\|_2^4\right) P((\mathcal{H}_K^{\epsilon,(k)})^C)} \\
&= \mathcal{O}(\frac{1}{n^2}) + \sqrt{\mathcal{O}(\frac{1}{n^2})\mathcal{O}(\frac{1}{n^{\bar{v}}})} = \mathcal{O}(\frac{1}{n^2}),
\end{aligned}$$

where the equation $\|E(\hat{\phi}_k - \phi^*)\|_2^2 = \mathcal{O}(1/n^2)$ follows from the proof of Theorem 1 Zhang et al. (2013), and now we conclude that $\|E\left(\hat{H}_k(\hat{\theta}_k)^{-1}(\hat{\phi}_k - \phi^*)I(\mathcal{H}_K^{\epsilon,(k)})\right)\|_2^2 \le C/n^2$. In summary we have

$$E\left(\|\frac{1}{K}\sum_{k=1}^{K}\hat{H}_k(\hat{\theta}_k)^{-1}(\hat{\phi}_k - \phi^*)\|_2^2 I(\mathcal{H}_K^{\epsilon})\right) \le \frac{C_1}{nK} + \frac{C_2}{n^2K} + \frac{C_3}{n^2} + \frac{C_4}{n^3}. \tag{54}$$

At last we consider bounding $E\left(\|\frac{1}{K}\sum_{k=1}^{K}(\hat{\phi}_k - \phi_k^*)\|_2^2 I(\|\frac{1}{K}\sum_{k=1}^{K}\hat{H}_k(\hat{\theta}_k)^{-1}(\hat{\phi}_k - \phi^*)\|_2^2 \ge C_4)\right)$ in (46) . Denote this term as $\tilde{R}$. Using previous results we can similarly show that

$$\tilde{R} = \mathcal{O}(\frac{1}{nK}) + \mathcal{O}(\frac{1}{n^2}). \tag{55}$$

With Equations (52), (54) and (55), the proof is complete. ■

B.5 Proof of Theorem 3

Proof With the results in Lemma B.4 and Lemma B.5, the proof follows from a direct application of the Slutsky's lemma. ■

B.6 Proof of Theorem 9

Proof

To apply Theorem 1 in Yuan and Jennrich (2000) Yuan and Jennrich (2000), we need to check the uniform convergence of $\frac{1}{n}\sum_{i=1}^{n}\left(\psi_{\phi}^{\lambda}(X_{k,i},\theta_k)^T \quad \psi_{\lambda}^{\lambda}(X_{k,i};\theta_k)^T\right)^T$. This is actually the last p_2 columns of $\nabla^2_{\theta_k} M_{n,k}(\theta_k)$ for $\theta_k \in U_k$, so we only need to show the uniform convergence of $\nabla^2_{\theta_k} M_{n,k}(\theta_k)$ in U_k. By Assumption 5, $\nabla^2_{\theta_k} M(x;\theta_k)$ is Lipschitz continuous w.r.t. θ_k for $\theta_k \in U_k$, then we can directly apply Corollary 3.1 of Newey (1991)Newey (1991) to establish the required uniform convergence.

Now we are to show $\hat{\lambda}_k^{(2)} \xrightarrow{P} \lambda_k^*$. Following the proof of Lemma 6 in Zhang et al. (2013) Zhang et al. (2013), we can first show that under the event $\mathcal{E}_k$, $M_{n,k}(\theta_k)$ is $(1-\rho)\rho_-$-strongly convex on the ball $\tilde{U}_k = \{\|\theta_k - \theta_k^*\|_2 \le \rho_k\}$, where $\rho_k \le min\{\frac{\rho\rho_-}{4G}, \rho\}$. Define the event $\mathcal{E}_{WD,k} = \{\|\hat{\phi}^{WD} - \phi^*\|_2 < \rho_k/2\}$, then under this event, $\tilde{\theta}_k^* = (\hat{\phi}^{WD}, \lambda_k^*) \in \tilde{U}_k$. For any $\theta_k^{'} = (\hat{\phi}^{WD}, \lambda_k) \in \Theta_k$, if $\theta_k^{'} \notin \tilde{U}_k$, then under $\mathcal{E}_{WD,k}$, there exists $w_0 \in [0,1]$ such that $\theta_{k,0}^{'} = w_0\theta_k^{'} + (1-w_0)\tilde{\theta}_k^*$ lies on the surface of the ball $\tilde{U}_k$, and thus $\|\theta_{k,0}^{'} - \tilde{\theta}_k^*\|_2 = w_0\|\theta_k^{'} - \tilde{\theta}_k^*\|_2 \in (\frac{\rho_k}{2}, \frac{3\rho_k}{2})$. Now under $\mathcal{E}_{WD,k}$ we have that

$$\begin{aligned} M_{n,k}(\theta_k^{'}) &\ge M_{n,k}(\theta_{k,0}^{'}) + < \nabla_{\theta_k} M_{n,k}(\theta_{k,0}^{'}), \theta_k^{'} - \theta_{k,0}^{'} > \\ &\ge M_{n,k}(\tilde{\theta}_k^*) + < \nabla_{\theta_k} M_{n,k}(\tilde{\theta}_k^*), \theta_k^{'} - \tilde{\theta}_k^* > + \frac{1}{2}(1-\rho)\rho_-\frac{\rho_k^2}{4} \\ &\quad + < \nabla_{\theta_k} M_{n,k}(\theta_{k,0}^{'}) - \nabla_{\theta_k} M_{n,k}(\tilde{\theta}_k^*), \theta_k^{'} - \theta_{k,0}^{'} > \\ &\ge M_{n,k}(\tilde{\theta}_k^*) + < \nabla_{\theta_k} M_{n,k}(\tilde{\theta}_k^*), \theta_k^{'} - \tilde{\theta}_k^* > + \frac{1}{2}(1-\rho)\rho_-\frac{\rho_k^2}{4}, \end{aligned} \tag{56}$$

where the first inequality holds due to the convexity of $M_{n,k}(\theta_k)$ on U_k and the second holds due to the strong convexity on $\tilde{U}_k$. The last inequality holds due to $\theta_k^{'} - \tilde{\theta}_k^* = \frac{1-w_0}{w_0}(\theta_{k,0}^{'} - \tilde{\theta}_k^*)$. When $\theta_k^{'} \in \tilde{U}_k$, with strong convexity Equation (56) still holds with $\frac{\rho_k^2}{4}$ changed to $\|\theta_k^{'} - \tilde{\theta}_k^*\|_2^2 = \|\lambda_k - \lambda_k^*\|_2^2$. In any case the following relationship holds under the event $\mathcal{E}_{WD,k}$:

$$M_{n,k}(\theta_k^{'}) \ge M_{n,k}(\tilde{\theta}_k^*) + < \nabla_{\theta_k} M_{n,k}(\tilde{\theta}_k^*), \theta_k^{'} - \tilde{\theta}_k^* > + \frac{1}{2}(1-\rho)\rho_- min\{\frac{\rho_k^2}{4}, \|\lambda_k - \lambda_k^*\|_2^2\}.$$

Rewriting the inequality we obtain that

$$\begin{aligned} & min\{\|\lambda_k - \lambda_k^*\|_2^2, \frac{\rho_k^2}{4}\} \le \frac{2}{(1-\rho)\rho_-}[M_{n,k}(\theta_k^{'}) - M_{n,k}(\tilde{\theta}_k^*) + < \nabla_{\theta_k} M_{n,k}(\tilde{\theta}_k^*), \theta_k^{'} - \tilde{\theta}_k^* >] \\ \le\ & \frac{2}{(1-\rho)\rho_-}[M_{n,k}(\theta_k^{'}) - M_{n,k}(\tilde{\theta}_k^*) + \|\nabla_{\theta_k} M_{n,k}(\tilde{\theta}_k^*)\|_2\|\theta_k^{'} - \tilde{\theta}_k^*\|_2]. \end{aligned} \tag{57}$$

Now if we denote $\theta^{'}_{k,1} = (\hat{\phi}^{WD}, \hat{\lambda}^{(2)}_k)$ and set $\theta^{'}_k = \kappa\theta^{'}_{k,1} + (1-\kappa)\tilde{\theta}^*_k$ for any fixed $\kappa \in [0,1]$, we will have

$$min\{\kappa\|\hat{\lambda}^{(2)}_k - \lambda^*_k\|_2, \frac{\rho_k^2}{4\kappa\|\hat{\lambda}^{(2)}_k - \lambda^*_k\|_2}\} \leq \frac{2(M_{n,k}(\kappa\theta^{'}_{k,1} + (1-\kappa)\tilde{\theta}^*_k) - M_{n,k}(\tilde{\theta}^*_k))}{\kappa(1-\rho)\rho_-\|\hat{\lambda}^{(2)}_k - \lambda^*_k\|_2} + \frac{2\|\nabla_{\theta_k} M_{n,k}(\tilde{\theta}^*_k)\|_2}{(1-\rho)\rho_-}.$$

By definition we have $M_{n,k}(\theta^{'}_{k,1}) \leq M_{n,k}(\tilde{\theta}^*_k)$ and thus by convexity we have

$$min\{\kappa\|\hat{\lambda}^{(2)}_k - \lambda^*_k\|_2, \frac{\rho_k^2}{4\kappa\|\hat{\lambda}^{(2)}_k - \lambda^*_k\|_2}\} \leq \frac{2\|\nabla_{\theta_k} M_{n,k}(\tilde{\theta}^*_k)\|_2}{(1-\rho)\rho_-}.$$

Define the event $\mathcal{E}_{s,k} = \{\frac{2\|\nabla_{\theta_k} M_{n,k}(\tilde{\theta}^*_k)\|_2}{(1-\rho)\rho_-} \leq \frac{\rho_k}{2}\}$, then under this event we have

$$min\{\kappa\|\hat{\lambda}^{(2)}_k - \lambda^*_k\|_2, \frac{\rho_k^2}{4\kappa\|\hat{\lambda}^{(2)}_k - \lambda^*_k\|_2}\} \leq \frac{\rho_k}{2}.$$

If $\|\hat{\lambda}^{(2)}_k - \lambda^*_k\|_2 > \frac{\rho_k}{2}$, we can set $\kappa = \frac{\rho_k}{2\|\hat{\lambda}^{(2)}_k - \lambda^*_k\|_2}$, which leads to a contradiction. Thus we have $\|\hat{\lambda}^{(2)}_k - \lambda^*_k\|_2 \leq \frac{\rho_k}{2}$. Then using Equation (57) we have

$$\|\hat{\lambda}^{(2)}_k - \lambda^*_k\|_2 < \frac{2\|\nabla_{\theta_k} M_{n,k}(\tilde{\theta}^*_k)\|_2}{(1-\rho)\rho_-}. \tag{58}$$

Since $\hat{\phi}^{WD}$ is consistent, we have $P(\mathcal{E}_{WD,k}) \to 1$. Besides, we already know that $P(\mathcal{E}_k) \to 1$. Due to the form of the event $\mathcal{E}_{s,k}$ and inequality (58), it remains to show $\|\nabla_{\theta_k} M_{n,k}(\tilde{\theta}^*_k)\|_2 = o_P(1)$ to establish the consistency of $\hat{\lambda}^{(2)}_k$. Note

$$\|\nabla_{\theta_k} M_{n,k}(\tilde{\theta}^*_k)\|_2 \leq \|\nabla_{\theta_k} M_{n,k}(\tilde{\theta}^*_k) - \nabla_{\theta_k} M_{n,k}(\theta^*_k)\|_2 + \|\nabla_{\theta_k} M_{n,k}(\theta^*_k)\|_2$$

and the latter term is of $O_p(\frac{1}{n^{v_1}})$. Using the consistency of $\hat{\phi}^{WD}$ we can show $\|\nabla_{\theta_k} M_{n,k}(\tilde{\theta}^*_k) - \nabla_{\theta_k} M_{n,k}(\theta^*_k)\|_2 = o_p(1)$. Besides, since $\hat{\phi}^{WD}$ is $\sqrt{N}$-consistent and $K \to \infty$, then $\sqrt{n}(\hat{\phi}^{WD} - \phi^*) = o_p(1)$ and the asymptotic normality of $\sqrt{n}\big(\frac{1}{n}\sum_{i=1}^n \psi_\lambda(X_{k,i};\theta^*_k) + \Psi^{\phi}_{\lambda}(\theta^*_k)(\hat{\phi}^{WD} - \phi^*)\big)$ is implied by the asymptotic normality of $\sqrt{n}\big(\frac{1}{n}\sum_{i=1}^n \psi_\lambda(X_{k,i};\theta^*_k)\big)$ and Slutsky's lemma. Now we apply Theorem 1 in Yuan and Jennrich (2000) and the result follows. ■

B.7 Proof of Theorem 4

Proof Recall the definition of the event $\mathcal{H}^{\epsilon}_K$ defined in Formula (45) in the proof of Theorem 2, we can similarly define $\mathcal{H}^{\epsilon}_{K,j}$ to control the estimation error of $\{\hat{H}_{k,j}(\hat{\theta}_{k,j})\}_{k=1}^K, j = 1, 2$, and $1 - P(\mathcal{H}^{\epsilon}_{K,j}) \leq \frac{C}{n^{v}}$. Note

$$E\left(\|\hat{\phi}^{dWD} - \phi^*\|_2^2\right) \leq \frac{1}{2}\sum_{j=1}^2 E\left(\|\hat{\phi}^{dWD}_j - \phi^*\|_2^2\right) = E\left(\|\hat{\phi}^{dWD}_1 - \phi^*\|_2^2\right),$$

so it suffices to bound the last term. Under the event $\mathcal{H}_{K,1}^{\epsilon}$ and using boundedness of $\|\hat{\phi}_1^{dWD}-\phi^*\|_2$,

$$
\begin{aligned}
& E\left(\|\hat{\phi}_1^{dWD}-\phi^*\|_2^2\right) \\
\leq\ & C_1 E\left(\|\frac{1}{K}\sum_{k=1}^{K}\hat{H}_{k,1}(\hat{\theta}_{k,1})^{-1}(\hat{\phi}_{k,2}^{bc}-\phi^*)\|_2^2 I(\mathcal{H}_{K,1}^{\epsilon})\right) \\
& +C_2 E\left(\|\frac{1}{K}\sum_{k=1}^{K}(\hat{\phi}_{k,2}^{bc}-\phi^*)\|_2^2 I(\|\frac{1}{K}\sum_{k=1}^{K}\hat{H}_{k,1}(\hat{\theta}_{k,1})^{-1}(\hat{\phi}_{k,2}^{bc}-\phi^*)\|_2^2 I(\mathcal{H}_{K,1}^{\epsilon})>C_3)\right) \\
& +C_4 E\left(\|\hat{\phi}_1^{dWD}-\phi^*\|_2^2 I((\mathcal{H}_{K,1}^{\epsilon})^c)\right) \triangleq C_1 E\left(R_1\right)+C_2 E\left(R_2\right)+C_4 E\left(R_3\right). \qquad (59)
\end{aligned}
$$

In the following, we control R_1, R_2, R_3, respectively.

$$
\begin{aligned}
R_1 \ \leq\ & C_1\|\frac{1}{K}\sum_{k=1}^{K}\hat{\tilde{H}}_{k,1}(\hat{\theta}_{k,1})(\hat{\theta}_{k,2}-\theta_k^*-\frac{1}{n/2}B_k(\theta_k^*))\|_2^2 I(\mathcal{H}_{K,1}^{\epsilon}) \\
& +\frac{1}{n^2}\frac{C_2}{K}\sum_{k=1}^{K}\|\hat{B}_{k,2}(\hat{\theta}_{k,2})1_{\mathcal{E}_{k,bc,2}}-B_k(\theta_k^*)\|_2^2 I(\hat{\theta}_{k,2}^{bc}\in\Theta_k),
\end{aligned}
$$

where $\hat{\tilde{H}}_{k,1}(\hat{\theta}_{k,1}) = \begin{pmatrix}\hat{H}_{k,1}(\hat{\theta}_{k,1})^{-1} & \mathbf{0}\end{pmatrix}$. Using the result in Lemma B.9 , the expectation of the second term in the RHS is of $\mathcal{O}(\frac{1}{n^3})$. Denote the first term as $R_1^{(1)}$. We can decompose the event $\mathcal{H}_{K,1}^{\epsilon}$ as $\mathcal{H}_{K,1}^{\epsilon} = \cap_{k=1}^{K}\mathcal{H}_{K,1}^{(k),\epsilon}$ where $\mathcal{H}_{K,1}^{\epsilon,(k)} = \{\|\hat{H}_{k,1}(\hat{\theta}_{k,1})-H_k(\theta_k^*)\|_2 \leq \epsilon_H\}$. Then, we have

$$
\begin{aligned}
R_1^{(1)} \ =\ & \|\frac{1}{K}\sum_{k=1}^{K}\hat{\tilde{H}}_{k,1}(\hat{\theta}_{k,1})I(\mathcal{H}_{K,1}^{(k),\epsilon})(\hat{\theta}_{k,2}-\theta_k^*-\frac{1}{n/2}B_k(\theta_k^*))\|_2^2 I(\mathcal{H}_{K,1}^{\epsilon}) \\
\leq\ & \|\frac{1}{K}\sum_{k=1}^{K}\hat{\tilde{H}}_{k,1}(\hat{\theta}_{k,1})I(\mathcal{H}_{K,1}^{(k),\epsilon})(\hat{\theta}_{k,2}-\theta_k^*-\frac{1}{n/2}B_k(\theta_k^*))\|_2^2.
\end{aligned}
$$

Since $\{\hat{\tilde{H}}_{k,1}(\hat{\theta}_{k,1})I(\mathcal{H}_{K,1}^{(k),\epsilon})\}_{k=1}^{K}$ are independent of $\{\hat{\theta}_{k,2}\}_{k=1}^{K}$, following the proof of Lemma B.8, it can be similarly showed that $E\|\frac{1}{K}\sum_{k=1}^{K}\hat{\tilde{H}}_{k,1}(\hat{\theta}_{k,1})I(\mathcal{H}_{K,1}^{(k),\epsilon})(\hat{\theta}_{k,2}-\theta_k^*-\frac{1}{n/2}B_k(\theta_k^*))\|_2^2 \leq \frac{C_1}{nK}+\frac{C_2}{n^2K}+\frac{C_3}{n^3}$. For the R_2 term appeared in (59), since $\|\frac{1}{K}\sum_{k=1}^{K}(\hat{\phi}_{k,2}^{bc}-\phi^*)\|_2$ is bounded, then we can apply Hölder's inequality and Markov's inequality to obtain

$$
\begin{aligned}
& E\left(R_2\right) \\
\leq\ & CE\|\frac{1}{K}\sum_{k=1}^{K}(\hat{\phi}_{k,2}^{bc}-\phi^*)\|_2 I(\|\frac{1}{K}\sum_{k=1}^{K}\hat{H}_{k,1}(\hat{\theta}_{k,1})^{-1}(\hat{\phi}_{k,2}^{bc}-\phi^*)\|_2^2 I(\mathcal{H}_{K,1}^{\epsilon})>C_3) \\
\leq\ & C\sqrt{E\|\frac{1}{K}\sum_{k=1}(\hat{\phi}_{k,2}^{bc}-\phi^*)\|_2^2 P(\|\frac{1}{K}\sum_{k=1}^{K}\hat{H}_{k,1}(\hat{\theta}_{k,1})^{-1}(\hat{\phi}_{k,2}^{bc}-\phi^*)\|_2^2 I(\mathcal{H}_{K,1}^{\epsilon})>C_3)} \\
\leq\ & \frac{C_1}{nK}+\frac{C_2}{n^2K}+\frac{C_3}{n^3}.
\end{aligned}
$$

For the last R_3 term in Equation (59), using the boundedness of $\|\hat{\phi}_1^{dWD} - \phi^*\|_2^2$, we have that $E(R_3) \le CP((\mathcal{H}_{K,1}^{\epsilon})^c) \le \frac{C_5 K}{n^{\bar{v}}}$. Now the proof is complete.

■

B.8 Proof of Theorem 5

Proof Similar to Lemma B.5, we first prove that the following term is of $o_p(N^{-1/2})$:

$$\{\sum_{s=1}^{K}\hat{H}_{s,1}(\hat{\theta}_{s,1})^{-1}\}^{-1}\sum_{k=1}^{K}\hat{H}_{k,1}(\hat{\theta}_{k,1})^{-1}(\hat{\phi}_{k,2}^{bc}-\phi^*)-\{\sum_{s=1}^{K}H_s(\theta_s^*)^{-1}\}^{-1}\sum_{k=1}^{K}H_k(\theta_k^*)^{-1}(\hat{\phi}_{k,2}^{bc}-\phi^*)$$

for $K = o(n^2)$. Denote the LHS term of the above equation as R_H. We have proved in the previous theorem that $P(\mathcal{H}_{K,1}^{\epsilon}) \to 1$ when $\bar{v} \ge 2$, so we only need to show $R_H I(\mathcal{H}_{K,1}^{\epsilon}) = o_p(1)$.

$$\begin{aligned}
&\|R_H\|_2 I(\mathcal{H}_{K,1}^{\epsilon})\\
\le\ & \sqrt{N}\|\{\frac{1}{K}\sum_{s=1}^{K}\hat{H}_{s,1}(\hat{\theta}_{s,1})^{-1}\}^{-1}-\{\frac{1}{K}\sum_{s=1}^{K}H_s(\theta_s^*)^{-1}\}^{-1}\|_2 I(\mathcal{H}_{K,1}^{\epsilon})\|\frac{1}{K}\sum_{k=1}^{K}\hat{H}_{k,1}(\hat{\theta}_{k,1})^{-1}(\hat{\phi}_{k,2}^{bc}-\phi^*)\|_2\\
&+C\sqrt{N}\|\frac{1}{K}\sum_{k=1}^{K}(\hat{H}_{k,1}(\hat{\theta}_{k,1})^{-1}-H_k(\theta_k^*)^{-1})(\hat{\phi}_{k,2}^{bc}-\phi^*)\|_2 I(\mathcal{H}_{K,1}^{\epsilon})\\
=\ & R_{H,1}+CR_{H,2}
\end{aligned}$$

For $K = o(n^2)$, we have shown in the proof of Theorem 5 that

$$I(\mathcal{H}_{K,1}^{\epsilon})\|\frac{1}{K}\sum_{k=1}^{K}\hat{H}_{k,1}(\hat{\theta}_{k,1})^{-1}(\hat{\phi}_{k,2}^{bc}-\phi^*)\|_2=\mathcal{O}_p(\frac{1}{\sqrt{N}}).$$

Besides, we also have that

$$\|\{\frac{1}{K}\sum_{s=1}^{K}\hat{H}_{s,1}(\hat{\theta}_{s,1})^{-1}\}^{-1}-\{\frac{1}{K}\sum_{s=1}^{K}H_s(\theta_s^*)^{-1}\}^{-1}\|_2 I(\mathcal{H}_{K,1}^{\epsilon})=\mathcal{O}_p(\frac{1}{\sqrt{n}}).$$

Combining those two results we prove that $R_{H,1} = O_p(\frac{1}{\sqrt{n}}) = o_p(1)$. Now we turn to bound $R_{H,2}$. Note that

$$R_{H,2}\ \le\ \sqrt{N}\|\frac{1}{K}\sum_{k=1}^{K}(\hat{\tilde{H}}_{k,1}(\hat{\theta}_{k,1})-\tilde{H}_{k,1}(\theta_k^*))I(\mathcal{H}_{K,1}^{(k),\epsilon})(\hat{\theta}_{k,2}^{bc}-\theta_k^*)\|_2=R_{H,2}^{(1)}.$$

From previous derivation, we know that when $K = o(n^2)$ the leading order term in $R_{H,2}^{(1)}$ is

$$R_{H,2}^{(2)}=\sqrt{N}\|\frac{1}{K}\sum_{k=1}^{K}(\hat{\tilde{H}}_{k,1}(\hat{\theta}_{k,1})-\tilde{H}_{k,1}(\theta_k^*))I(\mathcal{H}_{K,1}^{(k),\epsilon})\frac{1}{n/2}\sum_{i=1}^{n/2}d_{i,k}^{(2)}(\theta_k^*)\|_2,$$

where $d_{i,k}(\theta_k^*)^{(j)} = Q_k(\theta_k^*)\nabla_{\theta_k}M(X_{k,i}^{(j)};\theta_k^*)$, so we only need to show that $R_{H,2}^{(2)} = o_p(1)$. Denote $m = n/2$, then using the independence between $\hat{\tilde{H}}_{k,1}(\hat{\theta}_{k,1})$ and $d_{i,k}^{(2)}(\theta_k^*)$, we have that

$$\frac{1}{N}E(R_{H,2}^{(2)})^2\le\frac{1}{K^2}\sum_{k=1}^{K}E\left(\|(\hat{\tilde{H}}_{k,1}(\hat{\theta}_{k,1})-\tilde{H}_{k,1}(\theta_k^*))I(\mathcal{H}_{K,1}^{(k),\epsilon})\|_2^2\right)E\left(\|\frac{1}{m}\sum_{i=1}^{m}d_{i,k}^{(2)}(\theta_k^*)\|_2^2\right)+0=\mathcal{O}(\frac{1}{Nn}).$$

Then by Markov's inequality it's direct to show $R_{H,2}^{(2)} = o_p(1)$. Similarly we can show that

$$\sqrt{N}\left(\{\sum_{s=1}^{K}\hat{H}_{s,2}(\hat{\theta}_{s,2})^{-1}\}^{-1}\sum_{k=1}^{K}\hat{H}_{k,2}(\hat{\theta}_{k,2})^{-1}(\hat{\phi}_{k,1}^{bc}-\phi^*) - \{\sum_{s=1}^{K}H_s(\theta_s^*)^{-1}\}^{-1}\sum_{k=1}^{K}H_k(\theta_k^*)^{-1}(\hat{\phi}_{k,1}^{bc}-\phi^*)\right)$$

is $o_p(1)$ for $K = o(n^2)$. Now with Slutsky's lemma, it remains to establish the asymptotic normality of

$$\frac{1}{2}\sum_{j=1}^{2}\{\sum_{s=1}^{K}H_s(\theta_s^*)^{-1}\}^{-1}\sum_{k=1}^{K}H_k(\theta_k^*)^{-1}\hat{\phi}_{k,j}^{bc},$$

and the proof directly follows from the proof of Lemma B.10. ■

B.9 Proof of Corollary 8

Proof The coefficient 2 in the leading term is derived the same as the proof in Lemma B.9, and the remainders can be derived if we follow the proof of Lemma B.8. ■

C. Additional numerical results

C.1 Simulation results based on the errors in variables model

In this simulation experiment, we simulated the errors-in-variables Model (6) with the objective function (7) to compare the performance of the full sample, the split and conquer and the weighted distributed estimators: $\hat{\phi}_{full}$, $\hat{\phi}^{SaC}$ and $\hat{\phi}^{WD}$. The simulation was carried out by first generating IID $\{Z_{i,k}\}$ from $\mathcal{N}(\mu_Z, \sigma_Z^2)$, and then upon given a $Z_{i,k}$, $(X_{k,i}, Y_{i,k})^T$ were independently drawn from $\mathcal{N}\big((Z_{i,k}, \phi^* + \lambda_k^* Z_{i,k})^T, \sigma^2 I_{2\times 2}\big)$. We chose $\phi^* = 1, K = 2, \sigma^2 = 1$ and $n_1 = n_2 = 5 \times 10^4 = N/2$, and $\lambda_1^*, \lambda_2^*, \mu_Z$ and σ_Z^2 were those reported in Table 3 under four scenarios. As discussed in Section 2, the relative efficiency of $\hat{\phi}_{full}$ to $\hat{\phi}^{SaC}$ depends on the ratio $\sigma^2(E(Z))^2/(\text{var}(Z)E(Z^2))$ as shown in (8). We designed four scenarios according to the above ratio under $\lambda_1^* \neq \lambda_2^*$ and $EZ \neq 0$, respectively, which represented the settings where the full sample estimator $\hat{\phi}_{full}$ would be less (Scenario 1) or more (Scenario 2) efficient than the split and conquer estimator as predicted by the ratio, but not as efficient as the weighted distributed estimator $\hat{\phi}^{WD}$. Scenario 3 ($\lambda_1^* \neq \lambda_2^*, EZ = 0$) was the case when $\hat{\phi}_{full}$ and $\hat{\phi}^{WD}$ would be asymptotically equivalent, and both estimators would be more efficient than $\hat{\phi}^{SaC}$. Scenario 4 was the homogeneous case with $\lambda_1^* = \lambda_2^*$ in which all three estimators would have the same asymptotic efficiency. For all four scenarios, the ARE column of Table 3 confirmed the relative efficiency as predicted by the asymptotic variances in (8), and was well reflected in the comparison of the root mean square errors, as the bias is of smaller order as compared with that of the standard deviation and thus negligible.

C.2 Simulation results based on the logistic model

Figure 3 reports the absolute bias and root mean square errors of the estimators when $p_2 = 4$. Table 4 reports the empirical coverage and the average width of the CIs when $p_2 = 4$. Table 5 reports the average CPU time per simulation run based on 500 replications of the five estimators for a range of K for the logistic regression model with $p_2 = 4$. It is observed in Figure 3 that the bias of the estimators were smaller with $p_2 = 4$ compared to the results with $p_2 = 10$ in Figure

Table 3: Average root mean square error (RMSE) and the standard deviation (SD), multiplied by 10^2, of the full sample estimator $\hat{\phi}_{full}$, the SaC estimator $\hat{\phi}^{SaC}$ and the WD estimator $\hat{\phi}^{WD}$ under four scenarios for the errors-in-variables model (12) for $N = 10^5, K = 2$ and $n_1 = n_2$. AREs (asymptotic relative efficiency) of $\hat{\phi}_{full}$ to $\hat{\phi}^{SaC}$ are calculated from (8)

			$\hat{\phi}_{full}$		$\hat{\phi}^{SaC}$		$\hat{\phi}^{WD}$	
Scenario	$(\lambda_1^*, \lambda_2^*)$	ARE	RMSE	SD	RMSE	SD	RMSE	SD
Scenario 1	(0.25,3.25)	0.89	4.55	4.51	4.12	4.09	3.91	3.89
$(\mu_Z = 1, \sigma_Z^2 = 0.1)$	(0.5,3.5)	0.93	4.65	4.65	4.35	4.35	4.08	4.08
	(0.75,3.75)	0.97	4.52	4.52	4.40	4.38	4.13	4.13
Scenario 2	(0.25,2.25)	1.18	2.95	2.95	3.24	3.24	2.89	2.89
$(\mu_Z = 3, \sigma_Z^2 = 0.5)$	(0.75,2.75)	1.28	3.28	3.26	3.65	3.64	3.17	3.16
	(1.25,3.25)	1.31	3.71	3.71	4.16	4.07	3.64	3.61
Scenario 3	(0.25,2.25)	1.97	0.41	0.41	0.61	0.61	0.41	0.41
$(\mu_Z = 0, \sigma_Z^2 = 0.5)$	(0.75,2.75)	1.92	0.51	0.51	0.70	0.70	0.51	0.51
	(1.25,3.25)	1.68	0.64	0.64	0.82	0.82	0.64	0.64
Scenario 4	(0.5,0.5)	1	3.25	3.24	3.31	3.28	3.30	3.26
$(\mu_Z = 4, \sigma_Z^2 = 0.5)$	(1.0,1.0)	1	3.53	3.53	3.59	3.59	3.59	3.59
	(1.5,1.5)	1	4.06	4.03	4.08	4.07	4.06	4.06

1. As a consequence, the CIs based on the weighted distributed estimator had adequate coverage probabilities even when $K = 1000$.

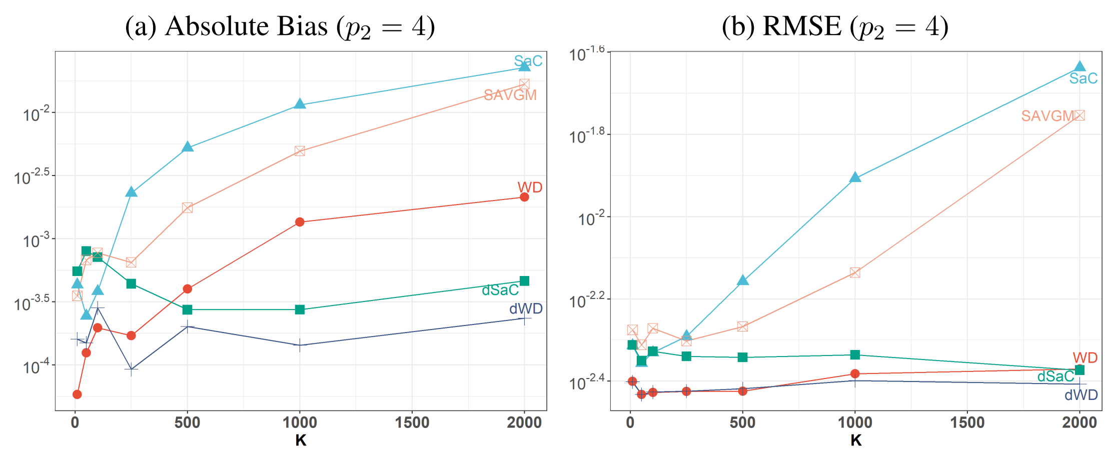


Figure 3: Average simulated bias (a) and the root mean square errors (RMSE) (b) of the weighted distributed (WD) (red circle), the split and conquer(SaC) (blue triangle), the debiased split and conquer (dSaC) (green square), the debiased weighted distributed (dWD) (purple cross), the subsampled average mixture SAVGM (pink square cross) estimators, with respect to the number of data block K for the logistic regression model with the dimension p_2 of the nuisance parameter λ_k being 4, and the full sample size $N = 2 \times 10^6$.

Table 4: Coverage probabilities and widths (in parentheses, multiplied by 100) of the $1-\alpha$ confidence intervals for the common parameter ϕ in the logistic regression model based on the asymptotic normality of the split and conquer (SaC), the weighted distributed (WD), the debiased split and conquer (dSaC) and the debiased weighted distributed (dWD) estimators with respect to the number of data blocks K. The dimension p_2 of the nuisance parameter λ_k is 4 and total sample size $N = 2 \times 10^6$

K		SaC			WD			dSaC			dWD	
	$1-\alpha$											
		0.99	0.95	0.90	0.99	0.95	0.90	0.99	0.95	0.90	0.99	0.95	0.90
10		0.99	0.96	0.92	0.99	0.97	0.91	0.99	0.96	0.92	0.99	0.96	0.91
		(2.45)	(1.87)	(1.57)	(2.03)	(1.55)	(1.30)	(2.45)	(1.87)	(1.57)	(2.03)	(1.55)	(1.30)
50		0.99	0.95	0.91	0.98	0.93	0.89	0.99	0.95	0.91	0.99	0.93	0.88
		(2.36)	(1.80)	(1.51)	(1.97)	(1.50)	(1.26)	(2.36)	(1.80)	(1.51)	(1.97)	(1.50)	(1.26)
100		0.98	0.94	0.91	0.99	0.95	0.91	0.99	0.95	0.91	0.99	0.95	0.91
		(2.36)	(1.79)	(1.51)	(1.96)	(1.49)	(1.25)	(2.36)	(1.79)	(1.51)	(1.96)	(1.49)	(1.25)
250		0.99	0.93	0.85	0.99	0.95	0.90	0.99	0.96	0.91	0.99	0.95	0.90
		(2.36)	(1.79)	(1.50)	(1.96)	(1.49)	(1.25)	(2.36)	(1.79)	(1.50)	(1.96)	(1.49)	(1.25)
500		0.91	0.77	0.66	0.99	0.95	0.88	0.99	0.96	0.90	0.99	0.95	0.89
		(2.36)	(1.80)	(1.51)	(1.96)	(1.49)	(1.25)	(2.36)	(1.80)	(1.51)	(1.96)	(1.49)	(1.25)
1000		0.65	0.41	0.28	0.99	0.94	0.88	0.99	0.94	0.88	0.99	0.93	0.88
		(2.38)	(1.81)	(1.52)	(1.96)	(1.49)	(1.25)	(2.38)	(1.81)	(1.52)	(1.97)	(1.50)	(1.25)
2000		0.01	0.01	0.00	0.99	0.91	0.81	0.98	0.94	0.88	0.99	0.94	0.90
		(2.42)	(1.84)	(1.55)	(1.96)	(1.50)	(1.25)	(2.42)	(1.84)	(1.55)	(1.98)	(1.50)	(1.26)

Table 5: Average CPU time for each replication based on $B = 500$ replications for the split and conquer (SaC), Zhang's SAVGM, the weighted distributed (WD), the debiased split and conquer (dSaC) and the debiased weighted distributed (dWD) estimators for the logistic regression model with respect to K. The dimension p_2 of the nuisance parameter λ_k is 4 and total sample size $N = 2 \times 10^6$

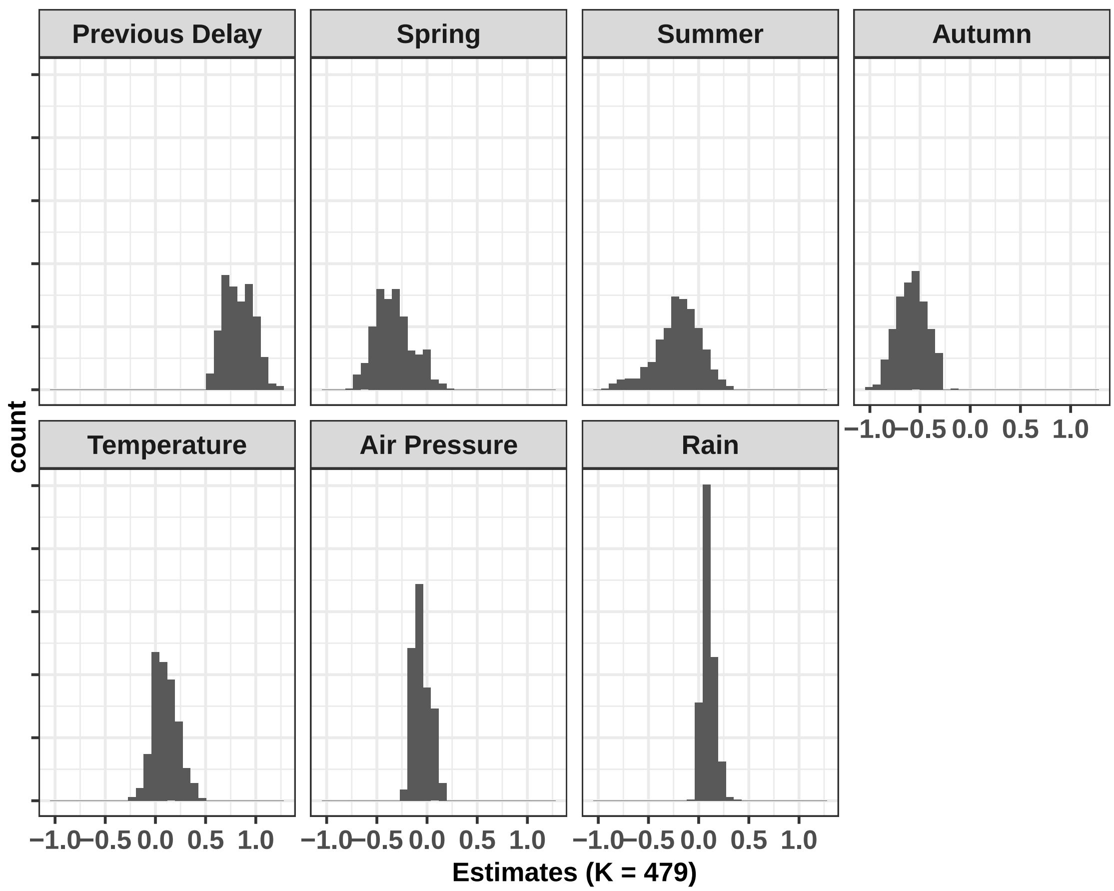
K	SaC	SAVGM	WD	dSaC	dWD
10	15.65	15.97	18.50	20.00	21.95
50	9.63	9.95	10.66	12.37	14.59
100	8.09	8.63	8.76	10.50	12.05
250	8.49	9.69	9.07	10.84	12.82
500	9.68	11.58	10.25	11.97	14.84
1000	11.67	13.81	12.32	13.93	19.08
2000	15.78	19.68	16.57	18.11	28.55

C.3 Pre-processing of the real data

The arrival delay of the previous flight that utilized the same plane was obtained by matching the tail number of the plane. The three meteorological factors (rain rate, close surface air pressure, and temperature) were obtained by matching this airline's on-time performance data with the ERA5 hourly data (https://cds.climate.copernicus.eu/). This dataset includes reanalysis from 1959 onwards whose temporal and spatial resolutions are one hour and $0.25^\circ \times 0.25^\circ$, respectively. We applied the $f(x) = log(1 + x)$ transformation to the rain variable due to its serious skewness. We also standardized each covariate in each of the data blocks before performing the logistic regression analysis.

We chose the parameter of the three meteorological factors as the common parameter based on Figure 4, which shows that the local estimates of those three parameters are the most concentrated.

Figure 4: Histogram of the parameter estimates across the data blocks



References

T. Ando. Concavity of certain maps on positive definite matrices and applications to hadamard products. *Linear Algebra and its Applications*, 26:203–241, 08 1979. doi: 10.1016/0024-3795(79)90179-4.

M. Bartlett. Approximate confidence intervals. *Biometrika*, 40:12–19, 01 1953. doi: 10.2307/2333091.

H. Battey, J. Fan, H. Liu, J. Lu, and Z. Zhu. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46:1352–1382, 06 2018. doi: 10.1214/17-AOS1587.

R. J. Carroll and D. Ruppert. The use and misuse of orthogonal regression in linear errors-in-variables models. *The American Statistician*, 50(1):1–6, 1996. ISSN 00031305. URL `http://www.jstor.org/stable/2685035`.

S. X. Chen and L. Peng. Distributed statistical inference for massive data. *The Annals of Statistics*, 49:2851–2869, 02 2021.

X. Chen and M. Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24:1655–1684, 10 2014. doi: 10.5705/ss.2013.088.

X. Chen, W. Liu, and Y. Zhang. Quantile regression under memory constraint. *The Annals of Statistics*, 47:3244–3273, 12 2019. doi: 10.1214/18-AOS1777.

R. Duan, Y. Ning, and Y. Chen. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, to appear, 02 2021. doi: https://doi.org/10.1093/biomet/asab007.

T. Evgeniou and M. Pontil. Regularized multi–task learning. *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, 08 2004. doi: 10.1145/1014052.1014067.

L. Hansen. Large sample properties generalized method of moments estimators. *Econometrica*, 50: 1029–1054, 02 1982. doi: 10.2307/1912775.

E. Haynsworth. Applications of an inequality for the schur complement. *Proceedings of the American Mathematical Society*, 24:512–516, 03 1970. doi: 10.1090/S0002-9939-1970-0255580-7.

M. Henmi and S. Eguchi. A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*, 91:929–941, 02 2004. doi: 10.1093/biomet/91.4.929.

C. Huang and X. Huo. A distributed one-step estimator. *Mathematical Programming*, 174:41–76, 11 2019. doi: 10.1007/s10107-019-01369-0.

M. Jordan, J. Lee, and Y. Yang. Communication-efficient distributed statistical learning. *Journal of the American Statistical Association*, 114:668–681, 05 2019. doi: 10.1080/01621459.2018.1429274.

P. Kairouz, H. McMahan, B. Avent, A. Bellet, M. Bennis, A. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. D'Oliveira, H. Eichner, S. El Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. Gibbons, and S. Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14:1–210, 01 2021. doi: 10.1561/9781680837896.

A. Kleiner, A. Talwalkar, P. Sarkar, and M. Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 76:795–816, 12 2011. doi: 10.1111/rssb.12050.

T. Lai and J. Wang. Edgeworth expansions for symmetric statistics with applications to bootstrap methods. *Statistica Sinica*, 3:517–542, 01 1993.

H. Li, B. Lindsay, and R. Waterman. Efficiency of projected score methods in rectangular array asymptotics. *Journal of the Royal Statistical Society Series B*, 65:191–208, 02 2003. doi: 10.1111/1467-9868.00380.

T. Li, A. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37:50–60, 05 2020. doi: 10.1109/MSP.2020.2975749.

N. Lin and R. Xi. Fast surrogates of U-statistics. *Computational Statistics & Data Analysis*, 54:16–24, 01 2010. doi: 10.1016/j.csda.2009.08.009.

T.-T. Lu and S.-H. Shiou. Inverses of 2 × 2 block matrices. *Computers & Mathematics With Applications - COMPUT MATH APPL*, 43:119–129, 01 2002. doi: 10.1016/S0898-1221(01)00278-4.

P. McCullagh. Quasi-likelihood functions. *The Annals of Statistics*, 11:59–67, 03 1983. doi: 10.1214/aos/1176346056.

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of Machine Learning Research*, 54:1273–1282, 20–22 Apr 2017. URL `https://proceedings.mlr.press/v54/mcmahan17a.html`.

W. Newey. Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59:1161–1167, 02 1991. doi: 10.2307/2938179.

O. Reiersol. Identifiability of a linear relation between variables which are subject to error. *Econometrica*, 18:375–389, 10 1950. doi: 10.2307/1907835.

P. Rilstone, V. Srivastava, and A. Ullah. The second-order bias and mean squared error of nonlinear estimators. *Journal of Econometrics*, 124:369–395, 12 1996. doi: 10.1016/0304-4076(96)89457-7.

S. Sengupta, S. Volgushev, and X. Shao. A subsampled double bootstrap for massive data. *Journal of the American Statistical Association*, 111:1222–1232, 08 2015. doi: 10.1080/01621459.2015.1080709.

V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar. Federated Multi-Task Learning. *Advances in Neural Information Processing Systems*, 05 2017.

L. Stefanski and D. Boos. The Calculus of M-Estimation. *The American Statistician*, 56:29–38, 02 2002. doi: 10.1198/000313002753631330.

A. van der Vaart. *Asymptotic Statistics*, chapter 5. Cambridge University Press, 01 1999. doi: 10.1017/CBO9780511802256.

S. Volgushev, S.-K. Chao, and G. Cheng. Distributed inference for quantile regression processes. *Annals of Statistics*, 47, 01 2017. doi: 10.1214/18-AOS1730.

Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10:1–19, 01 2019. doi: 10.1145/3298981.

A. Yaron, L. Hansen, and J. Heaton. Finite-Sample Properties of Some Alternative GMM Estimators. *Journal of Business & Economic Statistics*, 14:262–80, 02 1996. doi: 10.1080/07350015.1996.10524656.

K.-H. Yuan and R. Jennrich. Estimating equations with nuisance parameters: Theory and applications. *Annals of the Institute of Statistical Mathematics*, 52:343–350, 02 2000. doi: 10.1023/A:1004122007440.

Y. Zhang, J. Duchi, and M. Wainwright. Comunication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013. doi: 10.1109/CDC.2012.6426691.

T. Zhao, G. Cheng, and H. Liu. A partially linear framework for massive heterogeneous data. *The Annals of Statistics*, 44:1400–1437, 10 2014. doi: 10.1214/15-AOS1410.