Communication-efficient Distributed Newton-like Optimization with Gradients and $M$-estimators

Ziyan Yin

Abstract

In modern data science, it is common that large-scale data are stored and processed parallelly across a great number of locations. For reasons including confidentiality concerns, only limited data information from each parallel center is eligible to be transferred. To solve these problems more efficiently, a group of communication-efficient methods are being actively developed. We propose two communication-efficient Newton-type algorithms, combining the $M$-estimator and the gradient collected from each data center. They are created by constructing two Fisher information estimators globally with those communication-efficient statistics. Enjoying a higher rate of convergence, this framework improves upon existing Newton-like methods. Moreover, we present two bias-adjusted one-step distributed estimators. When the square of the center-wise sample size is of a greater magnitude than the total number of centers, they are as efficient as the global $M$-estimator asymptotically. The advantages of our methods are illustrated by extensive theoretical and empirical evidences.

1 Introduction

The statistical inference and optimization problem under a distributed setting is a popular topic in modern data science applications. For example, online marketing data are often too big to be stored within one hard drive; medical records are usually stored and processed separately at regional healthcare centers, and the inter-hospital sharing of the patient-level information is highly regulated and restricted. In these scenarios, despite the fact that more accurate inference can be obtained by combining data from all centers, practical realities prevent these centers from sharing the raw data stored therein. Among others, two foremost concerns are the data confidentiality and the technical feasibility. Consequently, only
limited statistics satisfying some restrictive rules are eligible to be infrequently transferred among centers.

Recent years have witnessed a flurry of important technological and methodological developments of the communication-efficient methods to handle such distributed data problems. One broadly applied class is the so-called divide-and-conquer strategy which takes average of certain statistics calculated at each center separately. Recent advances include Zhang et al. (2013); Rosenblatt and Nadler (2016); Lee et al. (2017); Zhang et al. (2015); Battey et al. (2018); Volgushev et al. (2019); Chen et al. (2020); Wang et al. (2019); Fan et al. (2019), etc.

However, one of the major limitations of these one-shot averaging methods is that, their efficacy is heavily limited by the center-wise sample size. In other words, the dataset cannot be split across too many centers, and adding extra data centers to the existing dataset does not always improve accuracy. Zhang et al. (2013) showed when the center-wise sample size is fixed, the mean squared error of the simple average of $M$-estimators can hardly be reduced by adding more data centers.

Recently, another communication-efficient framework has received considerable attention. Wang et al. (2017) and Jordan et al. (2019) proposed to use some surrogate loss function that can be evaluated efficiently at a preselected data center (also known as the local center). The advantage of this framework is that only gradients are collected from parallel centers, upon receiving an initial estimator. More specifically, this strategy replaces the higher-order derivatives of the global loss function (that requires data from all data centers) with those local surrogates. Due to its convenience and promising properties, this strategy has been broadly investigated and applied in different directions; for example, see Fan et al. (2021a); Duan et al. (2021); Yu et al. (2020, 2021); Li and Zhao (2021); Wang et al. (2019); Chen et al. (2021), among others.

Although the estimator proposed in Wang et al. (2017) and Jordan et al. (2019) appears efficient in some cases, when the number of centers $m$ and the center-wise sample size $n$ are of the same magnitude, the accuracy of the local surrogates becomes the bottleneck, limiting further improvement. The local high-order derivatives remain barely changed no matter how many parallel centers there are. More specifically, in each iteration, this framework uses the local Hessian estimator as the “learning rate”, which always has a fixed $O_P(n^{-1/2})$ gap to the true Fisher information matrix.

In light of this finding, in this paper, we are motivated to design a more efficient distributed estimation procedure by constructing some more accurate Fisher information estimators. We find in each parallel center, the Fisher information can be seen as the “coefficient” of a “linear” map from the gradient evaluated at the truth to the estimation error of the corresponding $M$-estimator. Although this “true” gradient is unavailable, when
a given estimator is close enough to the truth, the “linearity” between the errors of the \( M \)-estimators and the gradients evaluated at that given estimator still holds with negligible discrepancy. This kind of pseudo-linearity indicates that, its pseudo-coefficient — the true Fisher information matrix — can be estimated via linear regression. Thus, we propose two Fisher information estimators by combining the \( M \)-estimators and gradients collected from parallel centers. We show that under regular existence and continuity assumptions, the distances of both our estimators to their estimands are \( O_p(n^{-1} + m^{-1}) \), which outperform the local Hessian estimator used in [Wang et al. (2017)] and [Jordan et al. (2019)].

With these \( M \)-estimator/gradient-enhanced Fisher information estimators, we propose two iterative distributed algorithms correspondingly. Both have higher rates of convergence than the traditional method does. Additionally, we present two bias-adjusted one-step distributed estimators using the aforementioned Fisher information estimators. Their distances to the global \( M \)-estimator are reduced by removing quadratic terms of the estimation error of the initial estimator. As a result, these two refined one-shot estimators have the same efficiency asymptotically as the global \( M \)-estimator does when \( m = o(n^2) \).

We introduce some notations. Let \( F_X(u) \) and \( F_X^{-1}(q) \) be the cumulative distribution function and the quantile function of a random variable \( X \). For a vector \( a = (a_1, \ldots, a_m)^\top \), \( \text{supp}(a) = \{i : a_i \neq 0\} \), \( \|a\|_q = (\sum_i |a_i|^q)^{1/q} \), \( \|a\|_\infty = \max_i |a_i| \), and \( \|a\|_0 = \#\{i : a_i \neq 0\} \). Also, \( X \sim F \) denotes a random element \( X \) converges in distribution to \( F \). For a matrix \( A \), \( A_{ij} \) denote the entry in the \( i \)th row and \( j \)th column of \( A \). The Kronecker product is denoted by \( \otimes \).

2 Motivation and Problem Set-Up

We first state the structure and storage of the dataset. Let \( \{X_i\}_{i=1}^N \) denote \( N \) independent and identically distributed samples with marginal distribution \( P \) over some sample space \( \mathcal{X} \). For any parameter \( \theta \) containing in some convex space \( \Theta \subseteq \mathbb{R}^d \), define a convex and three-times differentiable loss function \( L : \Theta \times \mathbb{R}^d \to \mathbb{R} \), such that the true parameter \( \theta_0 \) is a minimizer of the population risk, that is

\[
\theta_0 \in \arg\min_{\theta \in \Theta} \mathbb{E}_P \{L(\theta, X)\}.
\]

We consider the evenly distributed setting where \( \{X_i\}_{i=1}^N \) are uniformly stored in one local (or central) center \( L \) (or \( M_1 \)) and \( m - 1 \) global (or parallel) centers \( \{M_i\}_{i=2}^m \) with \( \mathcal{G} = \bigcup_{i=1}^{m} M_i \); therefore, \( N = mn \). Suppose we have full access to the data stored in \( L \) which also plays the role of processing and delivering the final results. Despite the requirement
that the data are evenly stored with equal size $n$, our findings can be potentially generalized to unequal sample size cases without compromising the spirit.

Let $\hat{\theta}_i$ be the $M$-estimator from the $i$th center $\mathcal{M}_i$, and the average $\bar{\theta} = m^{-1} \sum_{i=2}^{m} \hat{\theta}_i$. Our findings can be potentially generalized to unequal sample size cases without compromising the spirit.

Let $\hat{\theta}_i$ be the $M$-estimator from the $i$th center $\mathcal{M}_i$, and the average $\bar{\theta} = m^{-1} \sum_{i=2}^{m} \hat{\theta}_i$. We refer the readers to [Zhang et al. (2013)] for a thorough picture of properties of $M$-estimators under this distributed setting.

Newton’s method is a powerful tool for optimization problems. However, when the dataset is split across many centers, and the inter-center communication cost is a major concern, it cannot be applied directly because the Hessian matrices can be huge and hard to transfer. One approach is the so-called communication-efficient surrogate likelihood (CSL) framework proposed in In Jordan et al. (2019) and Wang et al. (2017). They replaced the unavailable global Hessian with the local Hessian matrix $\hat{H}^{(l)}(\theta) = n^{-1} \sum_{j \in L} \nabla^2 L(\theta; X_j)$.

Then, Newton-like methods are applied upon receiving gradients from parallel centers. This method is efficient, and the communication cost is low. However, it depends heavily on the quality and quantity of the local data — $\hat{H}^{(l)}(\theta)$ remains almost unchanged no matter how large the global sample size is. That is, $\| \hat{H}^{(l)}(\theta) - I_0 \|_2 = O_P(\| \theta - \theta_0 \|_2) + O_P(n^{-1/2})$.

The estimation error of $\hat{H}^{(l)}(\theta)$ depends on $\| \theta - \theta_0 \|_2$ and $n^{-1/2}$, neither of which decreases with a larger $m$. Also, the whole optimization procedure may be ruined, if we accidentally choose a “bad” data center to be the local one.

Thus, we are trying to find communication-efficient alternatives of $\hat{H}^{(l)}(\theta)$, and we hope their errors would decrease when $m$ increases. This brings the estimators of the Fisher information matrix $I_0 = \mathbb{E}_P \{ \nabla^2 L(\theta_0; X) \}$ and its inverse $I_0^{-1}$:

\[
\hat{I}_0(\theta) = - \left\{ \sum_{i=1}^{m} (\hat{\theta}_i - \bar{\theta})(\hat{\theta}_i - \bar{\theta})^\top \right\}^{-1} \left[ \sum_{i=1}^{m} (\hat{\theta}_i - \bar{\theta}) \{ l_i(\theta) - \bar{I}(\theta) \}^\top \right],
\]

\[
\hat{\Omega}(\theta) = - \left[ \sum_{i=1}^{m} \{ l_i(\theta) - \bar{I}(\theta) \} \{ l_i(\theta) - \bar{I}(\theta) \}^\top \right]^{-1} \left[ \sum_{i=1}^{m} \{ l_i(\theta) - \bar{I}(\theta) \} (\hat{\theta}_i - \bar{\theta})^\top \right],
\]

where $l_i(\theta) = n^{-1} \sum_{j \in \mathcal{M}_i} \nabla L(\theta; X_j)$ is the gradient evaluated at $\theta$ in the $i$th center, and $\bar{I}(\theta) = m^{-1} \sum_{i=1}^{m} l_i(\theta)$ is their average. Their construction needs no high-order derivative; we only collect gradients and $M$-estimators. We name $\hat{I}_0(\theta)$ and $\hat{\Omega}(\theta)^{-1}$ the $M$-estimator/gradient (MG) and gradient/$M$-estimator (GM) Fisher information estimators, and will show they are closer to $I_0$ than $\hat{H}^{(l)}(\theta)$.
2.1 Estimation Accuracy of the M-Estimator/Gradient Fisher Information Estimators

Let $\Delta = \theta - \theta_0$, and define two cross-products involving the first and second derivatives of $L(\theta; X)$:

\[
Q_{11} = \mathbb{E}_P \left[ \nabla L(\theta_0; X_1) \nabla L(\theta_0; X_1)^\top \right],
\]
\[
Q_{12} = \mathbb{E}_P \left[ \nabla L(\theta_0; X_1) \otimes \{\nabla^2 L(\theta_0; X_1) - I_0\} \right].
\]

**Proposition 2.1.** Let $U(\rho) = \{\theta \in \Theta; \|\theta - \theta_0\|_2 \leq \rho\}$ be some Euclidean ball of radius $\rho > 0$ in which Assumptions 1, 2, 3, and 4 hold. Suppose $n$ and $m$ are large enough, such that $\hat{I}_0(\theta)$ and $\hat{\Omega}(\theta)$ exist almost surely.

(1) For any $\theta \in U(\rho)$,

\[
\hat{I}_0(\theta) - I_0 = I_0 Q_{11}^{-1} (I_d \otimes \Delta^\top) Q_{12} + R_D,
\]

where $\|R_D\|_2 = O_P \{ (n^{3/2} \|\Delta\|_2^3 + 1)(n^{-1} + m^{-1}) \}$;

(2) When $\theta \in U(\rho)$ and $\|\Delta\|_2 = o_P(n^{-1/4})$,

\[
\hat{\Omega}(\theta) - I_0^{-1} = Q_{11}^{-1} (I_d \otimes \Delta^\top) Q_{12} I_0^{-1} + R_G,
\]

where $\|R_G\|_2 = O_P(n \|\Delta\|_2^4 + m^{-1} + n^{-1})$.

It is clear that the estimation errors of $\hat{I}_0(\theta)$ and $\hat{\Omega}(\theta)$ consist of both $\Delta$-constant parts and sub-$\Delta$ parts. When $\|\Delta\|_2 = O_P(n^{-1/2})$, both $\|\hat{I}_0(\theta) - I_0\|_2$ and $\|\hat{\Omega}(\theta) - I_0^{-1}\|_2$ are $O(\|\Delta\|_2 + O_P(n^{-1} + m^{-1})$. By contrast, using the local Hessian leads to the error $\|\hat{H}^{(1)}(\theta) - I_0\|_2 = O_P(\|\Delta\|_2 + O_P(n^{-1/2})$. For a given $\Delta$, our proposed Fisher information estimators converge to their estimands stochastically faster than the local Hessian does, if $n = o(m^2)$. Also, for the $\Delta$-constant part, a local estimation can be conducted to further reduce their errors and speed up the whole process; see Theorem 3.1.

Remarkably, in generalized linear models (e.g., logistic regression), $Q_{11} = I_0$ and $Q_{12} = 0$. Our proposed estimators would outperform $\hat{H}^{(1)}(\theta)$ without any adjustment.

2.2 Decomposition of the Estimation Errors

Proposition 2.1 comes from two key properties: the mean-value theorem describing the gap between two Hessian matrices with different parameters, and the asymptotic “linearity”
between the $M$-estimator and the application of the gradient at $\theta_0$.

The first statement is an application of the multivariate mean-value theorem of the distance between $n^{-1} \sum_{j \in \mathcal{M}_i} \nabla^2 L(\theta; X_j)$ and $n^{-1} \sum_{j \in \mathcal{M}_i} \nabla^2 L(\theta_0; X_j)$: if the third derivative of $L(\theta; X)$ exists and is $m(X)$-Lipschitz continuous with respect to $\theta$, then, for any $\theta \in U(\rho)$,

\[
n^{-1} \sum_{j \in \mathcal{M}_i} \nabla^2 L(\theta; X_j) - n^{-1} \sum_{j \in \mathcal{M}_i} \nabla^2 L(\theta_0; X_j) \approx \{I_d \otimes \Delta^\top\} Q,
\]

\[
n^{-1} \sum_{j \in \mathcal{M}_i} \int_0^1 \nabla^2 L\{\theta_0 + t\Delta; X_j\} \, dt - n^{-1} \sum_{j \in \mathcal{M}_i} \nabla^2 L(\theta_0; X_j) \approx \frac{1}{2} \{I_d \otimes \Delta^\top\} Q,
\]

where

\[
Q = \begin{pmatrix}
\nabla^2 \left[ \frac{\partial}{\partial \theta_1} \mathbb{E}_P \{L(\theta_0; X)\} \right]
& \ldots \\
\vdots \\
\nabla^2 \left[ \frac{\partial}{\partial \theta_d} \mathbb{E}_P \{L(\theta_0; X)\} \right]
\end{pmatrix}.
\]

The second property is the “linear” association between the $i$th sample gradient at $\theta_0$ and the corresponding $M$-estimator $\hat{\theta}_i$ at $\mathcal{M}_i$. At the $i$th center, let $d_i = \hat{\theta}_i - \theta_0$ and $\bar{H}_i = n^{-1} \sum_{j} \int_0^1 \nabla^2 L(\theta_0 + td_i; X_{ij}) \, dt$. Then,

\[
l_i(\theta_0) = -\bar{H}_i d_i = -I_0 d_i + (I_0 - \bar{H}_i) d_i.
\]  \hfill (3)

Under the conditions given in Zhang et al. (2013), $\|d_i\|=O_p(n^{-1/2})$ and $\|I_0 - \bar{H}_i\|=O_p(n^{-1/2})$. In (3), $I_0 d_i$ becomes subordinate $\|I_0 - \bar{H}_i\|d_i$ becomes subordinate $\|I_0 - \bar{H}_i\|d_i$ becomes subordinate $\|I_0 - \bar{H}_i\|d_i$. All these facts lead to the “linearity” between $l_i(\theta_0)$ and $d_i$:

\[
l_i(\theta_0) \approx -I_0 d_i.
\]

In practice, only $l_i(\theta)$ is available, so we consider $g_i = l_i(\theta) - I_0 \Delta$ as the practical implementation of $l_i(\theta_0)$. The association between $g_i$ and $d_i$ can be quantified by expanding $g_i$ around $\theta_0$.

Let $\tilde{H}_i = n^{-1} \sum_j \int_0^1 \nabla^2 L(\theta_0 + t\Delta; X_{ij}) \, dt$. Applying (3),

\[
g_i = l_i(\theta_0) + \tilde{H}_i \Delta - I_0 \Delta = -I_0 d_i + e_i,
\]

where the residual $e_i = (I_0 - \bar{H}_i) d_i + (\tilde{H}_i - I_0) \Delta$. The form $g_i = -I_0 d_i + e_i$ indicates a
"regression" estimator of $I_0$:

$$
\hat{I}_0(\theta) = -\left( \sum d_i d_i^\top \right)^{-1} \left( \sum d_i g_i^\top \right),
$$

(4)

and its estimation error is

$$
\hat{I}_0(\theta) - I_0 = -\left( \sum d_i d_i^\top \right)^{-1} \left( \sum d_i e_i^\top \right).
$$

(5)

In (5), the denominator $m^{-1} \sum d_i d_i^\top$ is less a concern: the pseudo linearity $l_i(\theta_0) \approx -I_0 d_i$ leads to the fact that $m^{-1} \sum n d_i d_i^\top \approx m^{-1} I_0^{-1} \sum n l_i(\theta_0) l_i(\theta_0)^\top I_0^{-1} \sim I_0^{-1} Q_{11} I_0^{-1}$.

On the other hand, for the numerator, by definition,

$$
m^{-1} \sum d_i e_i^\top = m^{-1} \sum d_i d_i^\top (I_0 - H_i) + m^{-1} \sum d_i d_i^\top (H_i - \bar{H}_i) + m^{-1} \sum d_i \Delta^\top (H_i - I_0) + m^{-1} \sum d_i \Delta^\top (\bar{H}_i - H_i).
$$

Bounding each part, we have

$$
m^{-1} \sum_j d_i e_i^\top \approx n^{-1} (I_d \otimes \Delta^\top) Q_{12} + R_D',
$$

where $\|R_D'\| = O_p(n^{-2} + n^{-3/2} m^{-1/2}) + O_p(n^{-1/2} m^{-1/2} \|\Delta\|_2^2)$. Together with the denominator,

$$
\hat{I}_0(\theta) - I_0 = \left( \sum n d_i d_i^\top \right)^{-1} \left( \sum n d_i e_i^\top \right) \approx I_0 Q^{-1}_{11} I_0 (I_d \otimes \Delta^\top) Q_{12} + R''_D,
$$

with $\|R''_D\|_2 = O_p(n^{-1} + n^{-1/2} m^{-1/2} + n^{1/2} m^{-1/2} \|\Delta\|_2^2)$.

For practical applications, we implement the sample-based estimators $\hat{g}_i$ and $\hat{d}_i$ to approximate the unknown $g_i$ and $d_i$, which gives us $\hat{I}_0(\theta)$. Meanwhile, another implication of (4) is $d_i = -I_0^{-1} g_i + I_0^{-1} e_i$, and this brings the estimator of $I_0^{-1} - \hat{\Omega}(\theta)$, whose distance to $I_0^{-1}$ can be quantified following the similar steps.
3 Main Results

In this section, we present both iterative and one-shot Newton-like optimization algorithms using the GM and MG Fisher information estimators.

3.1 The Iterative Distributed Estimators

Based on Proposition 2.1, \( \hat{\mathbf{I}}_0(\theta) \) and \( \hat{\Omega}(\theta)^{-1} \) can be applied to replace the unattainable global Hessian in each iteration of Newton’s method. We propose our \( M \)-estimator/gradient Newton-type optimization methods in Algorithms 1 and 2.

Algorithm 1 Distributed Newton-Like Optimization with MG estimators

1: Collect all \( M \)-estimators \( \{\hat{\theta}_i\}_{i=1}^m \) and compute their average \( \hat{\theta} = m^{-1} \sum_i \hat{\theta}_i; \)
2: Input the initial estimator \( \tilde{\theta}_0 \) and the total number of iterations \( T; \)
3: for \( t = 0, \ldots, T - 1 \) do
4: Broadcast the current estimator \( \tilde{\theta}_t \) to all \( m \) parallel centers \( \{\mathcal{M}_i\}_{i=1}^m; \)
5: for \( i = 1, \ldots, m \) do
6: At center \( \mathcal{M}_i \), compute the gradient \( l_i(\tilde{\theta}_t) = n^{-1} \sum_{j \in \mathcal{M}_i} \nabla L(\tilde{\theta}_t; X_j); \)
7: Return \( l_i(\tilde{\theta}_t) \) to the local center \( \mathcal{L}; \)
8: end for
9: At \( \mathcal{L}, \) compute the global gradient \( \hat{\mathbf{I}}(\tilde{\theta}_t) = m^{-1} \sum_i l_i(\tilde{\theta}_t); \)
10: Construct the MG Fisher information estimator:
\[
\hat{\mathbf{I}}_0(\tilde{\theta}_t) = -\left\{ \sum_i (\tilde{\theta}_i - \bar{\theta})(\tilde{\theta}_i - \bar{\theta})^\top \right\}^{-1} \left[ \sum_i (\tilde{\theta}_i - \bar{\theta}) \left\{ l_i(\tilde{\theta}_t) - \hat{\mathbf{I}}(\tilde{\theta}_t) \right\}^\top \right];
\]
11: Update \( \tilde{\theta}_{t+1} = \tilde{\theta}_t - \hat{\mathbf{I}}_0(\tilde{\theta}_t)^{-1} \hat{\mathbf{I}}(\tilde{\theta}_t); \)
12: end for
13: return \( \tilde{\theta}_T. \)

Under conditions listed in Section 3.3, these methods converge stochastically faster than the existing method does.

Define
\[
Q_{12} \circ \mathbf{u} = (\mathbf{I}_d \otimes \mathbf{u}^\top)Q_{12}\mathbf{u}, Q \circ \mathbf{u} = (\mathbf{I}_d \otimes \mathbf{u}^\top)Q\mathbf{u}, \mathbf{u} \in \mathbb{R}^d.
\]
Algorithm 2 Distributed Newton-Like Optimization with GM estimators

1: Collect all $M$-estimators $\left\{ \hat{\theta}_i \right\}_{i=1}^m$ and compute their average $\bar{\theta} = m^{-1} \sum_i \hat{\theta}_i$;
2: Input the initial estimator $\tilde{\theta}_0$ and the total number of iterations $T$;
3: for $t = 0, \ldots, T - 1$ do
4: Broadcast the current estimator $\tilde{\theta}_t$ to all $m$ parallel centers $\{M_i\}_{i=1}^m$;
5: for $i = 1, \ldots, m$ do
6: At center $M_i$, compute the gradient $l_i(\tilde{\theta}_t) = n^{-1} \sum_{j \in M_i} \nabla L(\tilde{\theta}_t; X_j)$;
7: Return $l_i(\tilde{\theta}_t)$ to the local center $L$;
8: end for
9: At $L$, compute the global gradient $I(\tilde{\theta}_t) = m^{-1} \sum_i l_i(\tilde{\theta}_t)$;
10: Construct the GM Fisher information estimator:
$$\hat{\Omega}(\tilde{\theta}_t) = - \left[ \sum_i \left\{ l_i(\tilde{\theta}_t) - I(\tilde{\theta}_t) \right\} \left\{ l_i(\tilde{\theta}_t) - I(\tilde{\theta}_t) \right\}^\top \right]^{-1} \left[ \sum_i \left\{ l_i(\tilde{\theta}_t) - I(\tilde{\theta}_t) \right\} (\hat{\theta}_t - \bar{\theta}) \right] ;$$
11: Update $\tilde{\theta}_{t+1} = \tilde{\theta}_t - \hat{\Omega}(\tilde{\theta}_t)I(\tilde{\theta}_t)$;
12: end for
13: return $\tilde{\theta}_T$.

Let $\theta^*$ be the “oracle” $M$-estimator:
$$\theta^* = \arg \min_{\theta \in \Theta} N^{-1} \sum_{j \in G} L(\theta; X_j),$$
and $\Delta^* = \theta^* - \theta_0$.

Theorem 3.1. Let $\Delta_t = \tilde{\theta}_t - \theta_0$ and $\Delta_t' = \tilde{\theta}_t' - \theta_0$ for $t = 0, \ldots, T$. When the initial estimators $\tilde{\theta}_0$ and $\tilde{\theta}_0'$ belong to $U(\rho)$, and $\|\Delta_0\|_2 = O_P(n^{-1/2})$ and $\|\Delta'_0\|_2 = O_P(n^{-1/2})$, consider the estimators $\tilde{\theta}_{t+1}$ and $\tilde{\theta}'_{t+1}$ defined in Algorithms 1 and 2:
$$\tilde{\theta}_{t+1} - \theta^* = Q_{11}^{-1}(Q_{12} \circ \Delta_t) - 2^{-1} I_0^{-1}(Q \circ \Delta_t) + R_1^*,$$
$$\tilde{\theta}'_{t+1} - \theta^* = Q_{11}^{-1}(Q_{12} \circ \Delta'_t) - 2^{-1} I_0^{-1}(Q \circ \Delta'_t) + R_2^*,$$
where the residual terms

\[ \|R^*_1\|_2 = O_P\{(m^{-1} + n^{-1})(\|\Delta_t\|_2 + \|\Delta^*\|_2) + m^{-1}n^{-1}\} \]

\[ \|R^*_2\|_2 = O_P\{(m^{-1} + n^{-1})(\|\Delta'_t\|_2 + \|\Delta^*\|_2) + m^{-1}n^{-1}\}. \]

Theorem 3.1 indicates that when \( \|\Delta_t\|_2 = O_P(n^{-1/2}) \) and \( \|\Delta'_t\|_2 = O_P(n^{-1/2}) \), in probability \( \widetilde{\theta}_{t+1} \) and \( \widetilde{\theta}'_{t+1} \) converge stochastically faster than the CSL method does in each iteration. That is, under the conditions in Theorem 3.1, after each iteration,

\[ \|\Delta_{t+1}\|_2 = O(\|\Delta_t\|_2^2) + n^{-1/2}O_P(n^{-1} + m^{-1}). \]

The error of the updated estimator is bounded by the square of its predecessor’s estimation error, which is of the same order of magnitude as Newton’s method with the global Hessian. The same conclusion can be drawn for the GM estimator \( \theta'_{t+1} \) as well.

Also, our algorithms have no concern about accidentally choosing a “bad” local center. The whole iteration process depends on the initial estimator and the global data quality, instead of the local one. Therefore, when \( n \) and \( m \) are both large enough, they tend to be more stable.

On the other hand, when \( \Delta_t \) is the dominating error contributor rather than \( n \) or \( m \), the explicit forms of those \( \Delta_t \)-constant terms — \( Q_{11}^{-1}(Q_{12} \circ \Delta_t) \) and \( 2^{-1}I_0^{-1}(Q \circ \Delta_t) \) — give us a chance to make an adjustment for them with only the local data to achieve better accuracy. Following this idea, we propose two one-shot algorithms in the next section, using the local \( M \)-estimator as the initial estimator.

For the choices of the initial estimator, a popular one is the \( M \)-estimator of the local center \( \mathcal{L} \), whose \( l_2 \) error is \( O_P(n^{-1/2}) \). Another one is the average of \( M \)-estimators, \( \theta \), whose error is \( \|\theta - \theta_0\|_2 = O_P(n^{-1} + n^{-1/2}m^{-1/2}) \) (see Zhang et al. (2013)).

### 3.2 Bias-Adjusted One-Step Estimators

Often, in practice, communication among data centers is limited and one-step estimation is much more preferred. In this case, a common choice of the initial estimator is the \( n^{1/2} \)-consistent local \( M \)-estimator obtained within \( \mathcal{L} \). When \( m = o(n^2) \), Theorem 3.1 indicates the \( \Delta_0 \)-constant terms \( I_0^{-1}(Q \circ \Delta_0) \) and \( Q_{11}^{-1}(Q_{12} \circ \Delta_0) \) are the dominating contributors of \( \Delta_1 \). To adjust for these two terms, we propose to estimate \( Q_{11}, Q, \) and \( Q_{12} \) locally in \( \mathcal{L} \) with any \( n^{1/2} \)-consistent estimators; while estimation of \( \Delta_0 \) requires some more accurate estimator \( \theta^A \), and this can be done after one iteration of collecting gradients and \( M \)-estimators from parallel centers.
Proposition 3.1. For some estimator $\theta^A \in U(\rho)$, define $\Delta^A = \theta^A - \theta_0$, and let $\hat{\Delta}_0 = \overline{\theta} - \theta^A$. Construct the local estimators of $I_0$, $Q_{11}$, and $Q_{12}$ with $\theta^A$:

$$
\hat{H}^A = n^{-1} \sum_{j \in \mathcal{L}} \nabla^2 L(\theta^A; X_j)
$$
$$
\hat{Q} = n^{-1} \sum_{j \in \mathcal{L}} \nabla^3 L(\theta^A; X_j)
$$

$$
\hat{Q}_{11} = n^{-1} \sum_{j \in \mathcal{L}} \left\{ \nabla L(\theta^A; X_j) - n^{-1} \sum_{j \in \mathcal{L}} \nabla L(\theta^A; X_j) \right\} \left\{ \nabla L(\theta^A; X_j) - n^{-1} \sum_{j \in \mathcal{L}} \nabla L(\theta^A; X_j) \right\}^T
$$
$$
\hat{Q}_{12} = n^{-1} \sum_{j \in \mathcal{L}} \left\{ \nabla L(\theta^A; X_j) - n^{-1} \sum_{j \in \mathcal{L}} \nabla L(\theta^A; X_j) \right\} \otimes \left\{ \nabla^2 L(\theta^A; X_j) - n^{-1} \sum_{j \in \mathcal{L}} \nabla^2 L(\theta^A; X_j) \right\}
$$

Under the conditions in Theorem 3.1, when $\|\Delta^A\|_2 = O_P(n^{-1/2})$,

$$
\left\| (\hat{Q} \circ \hat{\Delta}_0) - (Q \circ \Delta_0) \right\|_2 = O_P(\|\Delta^A\|_2^2) + O_P(\|\Delta^A\|_2 \|\Delta_0\|_2) + O_P(n^{-1/2} \|\Delta_0\|_2^2)
$$
$$
\left\| (\hat{Q}_{12} \circ \hat{\Delta}_0) - (Q_{12} \circ \Delta_0) \right\|_2 = O_P(\|\Delta^A\|_2^2) + O_P(\|\Delta^A\|_2 \|\Delta_0\|_2^2) + O_P(n^{-1/2} \|\Delta_0\|_2^2)
$$
$$
\left\| (\hat{H}^A)^{-1}(\hat{Q} \circ \hat{\Delta}_0) - I_0^{-1}(Q \circ \Delta_0) \right\|_2 = O_P(\|\Delta^A\|_2^2) + O_P(\|\Delta^A\|_2 \|\Delta_0\|_2) + O_P(n^{-1/2} \|\Delta_0\|_2^2)
$$
$$
\left\| \hat{Q}_{11}^{-1}(\hat{Q}_{12} \circ \hat{\Delta}_0) - Q_{11}^{-1}(Q_{12} \circ \Delta_0) \right\|_2 = O_P(\|\Delta^A\|_2^2) + O_P(\|\Delta^A\|_2 \|\Delta_0\|_2) + O_P(n^{-1/2} \|\Delta_0\|_2^2).
$$

Among other choices of $\theta^A$, we suggest to use the average of $M$-estimators $\overline{\theta}$ with the error $\|\overline{\theta} - \theta_0\|_2 = O_P(n^{-1/2}m^{-1/2} + n^{-1})$. It is more stable and does no depends on the local center. Otherwise, if $m$ is small or the local data quality is satisfactory, the one-step updated estimator $\overline{\theta}_1$ or $\overline{\theta}'_1$ can be taken as well.

Theorem 3.2. When the conditions in Theorem 3.1 hold, take the local M-estimator as $\overline{\theta}_0$, and set $\theta^A = \overline{\theta}$.

(1) Consider the one-step bias-adjusted distributed estimator based on $\hat{I}_0(\overline{\theta}_0)$:

$$
\tilde{\theta}_{os} = \overline{\theta}_0 - \hat{I}_0(\overline{\theta}_0)^{-1}\hat{I}(\overline{\theta}_0) - \hat{Q}_{11}^{-1}(\hat{Q}_{12} \circ \hat{\Delta}_0) + \frac{1}{2}(\hat{H}^A)^{-1}(\hat{Q} \circ \hat{\Delta}_0).
$$

Then, $\|\tilde{\theta}_{os} - \theta^A\|_2 = O_P(n^{-3/2} + m^{-1}n^{-1/2})$. 

11
(2) Consider the one-step bias-adjusted distributed estimator based on \( \hat{\Omega}(\widehat{\theta}_0) \):

\[
\tilde{\theta}_\text{os}' = \tilde{\theta}_0 - \hat{\Omega}(\widehat{\theta}_0) \hat{I}(\widehat{\theta}_0) - \hat{Q}_{11}^{-1} (\hat{Q}_{12} \circ \hat{\Delta}_0) + \frac{1}{2} (\hat{H}^A)^{-1}(\hat{Q} \circ \hat{\Delta}_0).
\]

Then, \( \|\tilde{\theta}_\text{os}' - \theta^*\|_2 = O_P(n^{-3/2} + m^{-1}n^{-1/2}) \).

It is well known that the oracle estimator \( \theta^* \) converges in distribution to \( N(0, I_0^{-1}Q_{11}^{-1}) \). Theorem 3.2 indicates that both \( \tilde{\theta}_\text{os} \) and \( \tilde{\theta}_\text{os}' \) have the same limiting distributions when \( m = o(n^2) \), which means they both achieve the optimal asymptotic efficiency as the oracle estimator \( \theta^* \) does. Thus, Gaussian approximation can be applied with any consistent estimators of \( I_0 \) and \( Q_{11} \) (for example, see Proposition 3.1) for statistical inference and construction of confidence intervals.

**Corollary 3.1.** Consider \( \tilde{\theta}_\text{os} \) and \( \tilde{\theta}_\text{os}' \) defined in Theorem 3.2. Under the conditions in Theorem 3.2, when \( m = o(n^2) \),

\[
(nm)^{1/2} \{ \hat{\Omega}(\tilde{\theta}_0) \hat{Q}_{11} \hat{\Omega}(\tilde{\theta}_0)^\top \}^{-1/2} (\tilde{\theta}_\text{os} - \theta_0) \sim N(0, I)
\]

\[
(nm)^{1/2} \{ \hat{\Omega}(\tilde{\theta}_0) \hat{Q}_{11} \hat{\Omega}(\tilde{\theta}_0)^\top \}^{-1/2} (\tilde{\theta}_\text{os}' - \theta_0) \sim N(0, I).
\]

### 3.3 Technical Assumptions

In this section, we list the convexity and identifiability assumptions used in our technical analysis.

**Assumption 1.** The parameter space \( \Theta \subset \mathbb{R}^d \) is a compact convex set, with \( \theta \in \text{int}(\Theta) \) and \( l_2 \)-radius \( r_0 = \max_{\theta \in \Theta} \| \theta - \theta_0 \|_2 \).

**Assumption 2.** That \( \theta_0 \in \Theta \) is the unique minimizer of the population risk \( \mathbb{E}_P \{ L(\theta; X_i) \} \). For any \( \delta > 0 \), there exists \( \varepsilon > 0 \), such that

\[
\lim \inf_{n \to \infty} \mathbb{P} \left[ \inf_{\|\theta - \theta_0\|_2 \geq \delta} \left\{ n^{-1} \sum_{i=1}^{n} L(\theta; X_i) - n^{-1} \sum_{i=1}^{n} L(\theta_0; X_i) \right\} \geq \varepsilon \right] = 1.
\]

**Assumption 3.** The population risk is twice-differentiable, and there exist finite constants \( \lambda_-, \lambda_+, \lambda_1, \lambda_2 > 0 \) such that \( \lambda_- I \preceq I_0 \preceq \lambda_+ I \),

\[
\mathbb{E}_P\{ \| \nabla L(\theta_0; X) \|_2^6 \} \leq \lambda_1^6 \text{ and } \mathbb{E}_P\{ \| \nabla^2 L(\theta_0; X) \|_2^6 \} \leq \lambda_2^6.
\]
Additionally, for all $\theta', \theta \in U$,
\[ \|\nabla^2 L(\theta'; X) - \nabla^2 L(\theta; X)\|_2 \leq h(X)\|\theta' - \theta\|_2 \]
We assume $\max(\mathbb{E}_P\{h(X)^6\}, \mathbb{E}_P[|h(X) - \mathbb{E}_P\{h(X)\}|^6]) \leq \lambda_h^6$ for some finite positive constant $\lambda_h$.

We further assume the loss function is three-times differentiable with enough continuity in $U(\rho)$. Recall the definition of the third derivative tensor operator of $L(\theta; X)$, $T(\theta; X) = \nabla^3 L(\theta; X)$, for $u \in \mathbb{R}^d$:
\[ T(\theta; X) = \nabla^3 L(\theta; X) = \begin{bmatrix} \nabla^2 \{ \frac{\partial}{\partial \theta_1} L(\theta; X) \} \\ \vdots \\ \nabla^2 \{ \frac{\partial}{\partial \theta_d} L(\theta; X) \} \end{bmatrix}. \]
Note that $Q$ defined in (2.2) is the expectation of $T(\theta_0, X)$. The definition of $T(\theta, X)$ indicates, for $\theta \in U(\rho)$,
\[ \nabla^2 L(\theta; X) - \nabla^2 L(\theta_0; X) = \int_0^1 \{I_d \otimes (\theta - \theta_0)^\top\} T(\theta_0 + t(\theta - \theta_0), X) \, dt. \]
When $\theta$ is close to $\theta_0$, and $T(\theta, X)$ is smooth enough with respect to $\theta$, $\nabla^2 L(\theta; X) - \nabla^2 L(\theta_0; X) \approx \{I_d \otimes (\theta - \theta_0)^\top\} T(\theta_0, X)$. The condition is specified as follows.

**Assumption 4.** The risk function is three-times differentiable, and $T(\theta, X)$ is $m(X)$-Lipschitz. That is, for all $\theta', \theta \in U(\rho)$ and $u \in \mathbb{R}^d$,
\[ \| (I_d \otimes u^\top) \{T(\theta', X) - T(\theta, X)\} \|_2 \leq m(X)\|\theta - \theta\|_2\|u\|_2, \]
with $\max(\mathbb{E}_P\{m(X)^4\}, \mathbb{E}_P[\{m(X) - \mathbb{E}\{m(X)\}\}^4]) \leq \lambda_m^4$ for some finite positive constant $\lambda_m$. Additionally, at $\theta_0$, $\|T(\theta_0, X)\|_2$ is bounded by $g(X)$ for some function $g(X)$, and $\mathbb{E}_P\{g(X)^4\} \leq \lambda_3^4$.

Assumption 4 also implicates Assumption 3. That is, under Assumption 3, $\|T(\theta, X)\|_2 \leq g(X) + \rho m(X)$. We can choose $h(X) = g(X) + \rho m(X)$, so that for any $\theta', \theta \in U(\rho)$,
\[ \|\nabla^2 L(\theta'; X) - \nabla^2 L(\theta; X)\|_2 \leq \int_0^1 \|T(\theta + t(\theta' - \theta), X)(\theta' - \theta)\|_2 \, dt \]
\[ \leq h(X)\|\theta' - \theta\|_2. \]
4 Simulations

In this section, we validate and visualize our methodology with extensive synthetic examples. Our presentation starts with evaluating the accuracy of \( \hat{I}_0(\theta) \) and \( \hat{\Omega}(\theta) \). In the second part, the convergence rates of \( \hat{\theta}_t \) and \( \hat{\theta}'_t \) proposed in Algorithms 1 and 2 are compared with existing methods, when multiple rounds of communications among centers are eligible. Then, in the cases of one-step estimation, we demonstrate the distributions and the empirical coverage probabilities of \( \hat{\theta}_{os} \) and \( \hat{\theta}'_{os} \) proposed in Theorem 3.2 with different \( n \) and \( m \).

Herein, suppose \( X \) has two components: \( X = (Y, S^\top)^\top \), where \( Y \) is the response variable, and \( S \) is the predictor. Let \( \theta_0 \) be the true coefficient describing the association between \( Y \) and \( S \). We show examples of Poisson regression and logistic regression with canonical links, with \( S \) coming in different sizes and different distributions. It is our pragmatic experience that the distribution of the design matrices makes no substantial difference, as long as those moment assumptions hold.

4.1 Estimation Accuracy for Fisher Information

The estimation accuracy of \( \hat{I}_0(\theta) \) and \( \hat{\Omega}(\theta) \) is quantified by the relative distance to \( I_0 \) and \( I_0^{-1} \), that is, the Fisher information estimation error — \( \delta_1(A) = \|I_0\|^2_2 \|I_0 - A\|_2 \) — and the inverse Fisher information estimation error — \( \delta_2(A) = \|I_0^{-1}\|^2_2 \|I_0^{-1} - A^{-1}\|_2 \) — for some non-trivial matrix \( A \).

Proposition 2.1 indicates \( \delta_1 \) and \( \delta_2 \) are affected by the error \( \Delta \), the center-wise sample size \( n \), and the total number of centers \( m \). To demonstrate a comprehensive picture, we sample \( \theta \) from \( N(\theta_0, \sigma^2 I) \) with \( \sigma^2 = 16^{-1}, 256^{-1}, \) and \( 65536^{-1} \). Also, \( n \) and \( m \) take 100, 200, 400, and 800 separately. For each combination of \( \{\sigma^2, n, m\} \), 10000 simulations are repeated.

The results are visualized in Figure 1. We compare four estimators of \( I_0 \) and their inverses for \( I_0^{-1} \): the local estimator \( H^{(l)}(\theta) = n^{-1} \sum_{j \in \mathcal{L}} \nabla^2 L(\theta; X_j) \) (LC, black lines), the global estimator \( H^{(g)}(\theta) = N^{-1} \sum_{j \in \mathcal{G}} \nabla^2 L(\theta; X_j) \) (GL, green lines), the proposed MG Fisher information estimator \( \hat{I}_0(\theta) \) (MG, red dashed lines), and the GM Fisher information...
estimator $\hat{\Omega}(\theta)^{-1}$ (GM, red solid lines).

From Figure 1, we observe that with fixed $\sigma^2$ and $m$, the estimation errors of both proposed methods drop as $n$ increases; while $H^{(l)}(\theta)$ and $H^{(g)}(\theta)$ remain almost the same, when $\theta$ is far from $\theta_0$. It is not surprising that our proposed estimators are closer to the truth. Recall that in generalized linear models, $Q_{12} = 0$; in this scenario, Proposition 2.1 indicates that the impacts of $\Delta$ on $\hat{I}_0(\theta)$ and $\hat{\Omega}(\theta)^{-1}$ diminish, as $n$ and $m$ increase.

On the other hand, if $\|\theta - \theta_0\|_2$ is small, $\delta_1\{H^{(l)}(\theta)\}$ and $\delta_2\{H^{(g)}(\theta)\}$ also drop and converge to their global counterparts. Similar phenomena can be observed in $\hat{I}_0(\theta)$ and $\hat{\Omega}(\theta)$ as well. But both red lines drop much faster as $n$ increases. This is consistent with our conclusions in Proposition 2.1.

4.2 Performance of Iterative Distributed Estimation

Our second presentation focuses on the iterative estimators $\tilde{\theta}_t$ and $\tilde{\theta}'_t$ proposed in Algorithm 1 and 2. To verify our conclusion, we consider a Poisson model:

$$P(Y_i = y | S_i, \theta_0) = \frac{1}{y!} \exp\{yS_i^T \theta_0 - \exp(S_i^T \theta_0)\},$$

where $d = 4$, $\theta_0 = d^{-1/2}1$, and $S_i \sim N(0, I)$.

The initial estimator $\tilde{\theta}_0$ is set to be the average $M$-estimator $\overline{\theta}$. Three iterations are performed. We take $n$ and $m$ to be 100, 200, 400, and 800 respectively, and compare the relative $l_2$ distance to the oracle estimator $\theta^*$, $\delta_o(\theta) = \|\theta^*\|^{-1}_2\|\theta^* - \theta\|_2$, at different rounds of iterations. For each combination of $n$ and $m$, 10000 simulations are performed.

Figure 2 presents the comparisons. In each sub-figure, the back line (CSL) stands for the CSL method proposed in [Jordan et al. (2019)]. The green line (GL) stands for the ideal Newton estimator when we have all the data. The red dashed line and the red solid line stand for $\tilde{\theta}_t$ (MG) and $\tilde{\theta}'_t$ (GM) proposed in Algorithm 1 and 2.

This figure tells that our methods converge to $\theta^*$ much faster than CSL does. In each setting, regardless of the ratio between $n$ and $m$, our methods are able to keep the relative errors below 5% after two rounds of communications, and < 1% after three rounds. By contrast, the performance of CSL depends heavily on $n$; that is, the local data quality has great impact on CSL. This confirms our conclusion and discussion in Theorem 3.1.
4.3 Performance of One-Step Distributed Estimation

The last presentation focuses on the refined one-step distributed estimators \( \tilde{\theta}_{os} \) and \( \tilde{\theta}'_{os} \) discussed in Theorem 3.2 and Corollary 3.1. We state that they have the same asymptotic efficiency as \( \theta^* \) does when \( m = o(n^2) \). To examine this, with different \( m \) and \( n \), we demonstrate plots of the proposed estimators versus the oracle estimator and the empirical confidence interval coverage. The results in this section focus on the first element of \( \theta_0 \).

We take \( n \) and \( m \) to be 100, 200, 400, 800, and 1600 respectively. Four one-step methods are compared: the one-step CSL methods proposed in Jordan et al. (2019); the average M-estimator discussed in Zhang et al. (2013); our two proposed methods, \( \tilde{\theta}_{os} \) and \( \tilde{\theta}'_{os} \). Three models are involved:

- “Model 1” is logistic regression (6) with \( \theta_0 = d^{-1/2}1 \), and \( S_i \sim N(0, I) \);
- “Model 2” is logistic regression (6) with \( \theta_0 = d^{-1/2}1 \), and elements of \( S_i \) are mutually independent and follow \( \exp(1) \);
- “Model 3” is Poisson regression (7) with \( \theta_0 = d^{-1/2}1 \), and \( S_i \sim N(0, I) \).

Figure 3 examines the approximation performance of the proposed methods. The x-axis is the oracle estimator. And the estimators from the four methods are on the y-axis. We have a few observations. When the ratio \( nm^{-1/2} \) increases, the distance between our proposed methods and \( \theta^* \) is getting smaller and smaller. By contrast, due to the impact from the local center, the accuracy of CSL drops rapidly when \( m \) increases. Also, it is interesting to mention that there exists an “eternal” gap between \( \tilde{\theta} \) and the oracle estimator. This stands for the bias of \( \tilde{\theta} \), which depends on \( n \) only, and cannot be reduced by adding more data centers.

The 95% confidence intervals of the first element of \( \theta_0 \) are constructed based on \( \tilde{\theta}_{os} \) and \( \tilde{\theta}'_{os} \), and their empirical coverage is evaluated in Table 1. The normal approximation methods proposed in Corollary 3.1 are applied. These results are consistent with Figure 3: when \( n \) is large and \( m = o(n^2) \), our methods provide close-to-nominal coverage. It is justifiable to request a large \( n \), since both proposed methods are based on the large-sample theory.
Table 1: Empirical coverage probabilities for the first element of $\theta_0$. Methods given in Corollary 3.1 are used to construct 95% confidence intervals.

| $m$  | $n$  | $d = 2$ Model 1 | $d = 2$ Model 2 | $d = 2$ Model 3 | $d = 4$ Model 1 | $d = 4$ Model 2 | $d = 4$ Model 3 |
|------|------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|      |      | $\hat{\theta}_{os}$ | $\hat{\theta}_{os}$ | $\hat{\theta}_{os}$ | $\hat{\theta}_{os}$ | $\hat{\theta}_{os}$ | $\hat{\theta}_{os}$ |
| 100  | 100  | 90  | 86  | 87  | 77  | 83  | 86  | 76  | 74  | 64  | 73  | 75  |
|      | 200  | 93  | 92  | 92  | 88  | 90  | 91  | 89  | 88  | 85  | 82  | 85  | 86  |
|      | 400  | 94  | 94  | 94  | 92  | 93  | 93  | 92  | 91  | 89  | 90  | 91  |      |
|      | 800  | 95  | 94  | 94  | 94  | 94  | 94  | 93  | 93  | 92  | 93  | 93  |      |
|      | 1600 | 95  | 95  | 95  | 95  | 95  | 95  | 94  | 94  | 94  | 95  | 95  | 94  |
|      | 100  | 88  | 81  | 83  | 69  | 79  | 83  | 78  | 69  | 71  | 58  | 68  | 71  |
|      | 200  | 93  | 90  | 91  | 85  | 87  | 90  | 87  | 85  | 82  | 77  | 79  | 85  |
|      | 400  | 94  | 94  | 93  | 90  | 93  | 93  | 92  | 91  | 89  | 87  | 88  | 90  |
|      | 800  | 95  | 95  | 95  | 94  | 94  | 94  | 93  | 93  | 92  | 92  | 93  |      |
|      | 1600 | 94  | 94  | 95  | 95  | 94  | 95  | 94  | 94  | 94  | 94  | 94  |      |
|      | 100  | 84  | 74  | 81  | 60  | 71  | 79  | 73  | 60  | 63  | 47  | 57  | 64  |
|      | 200  | 91  | 87  | 89  | 78  | 82  | 88  | 84  | 81  | 78  | 70  | 74  | 82  |
|      | 400  | 93  | 92  | 93  | 89  | 90  | 93  | 90  | 91  | 87  | 85  | 85  | 90  |
|      | 800  | 95  | 94  | 94  | 93  | 93  | 94  | 94  | 94  | 91  | 90  | 91  | 93  |
|      | 1600 | 95  | 95  | 95  | 95  | 95  | 95  | 94  | 94  | 94  | 93  | 94  | 94  |
|      | 100  | 81  | 65  | 77  | 50  | 66  | 71  | 67  | 50  | 57  | 38  | 50  | 56  |
|      | 200  | 88  | 84  | 86  | 71  | 78  | 85  | 79  | 74  | 73  | 61  | 67  | 76  |
|      | 800  | 94  | 94  | 94  | 92  | 92  | 94  | 91  | 92  | 90  | 90  | 89  | 92  |
|      | 1600 | 95  | 95  | 95  | 94  | 94  | 94  | 93  | 93  | 92  | 93  | 93  | 94  |
|      | 100  | 76  | 53  | 71  | 40  | 58  | 62  | 59  | 38  | 49  | 27  | 42  | 46  |
|      | 200  | 86  | 76  | 82  | 61  | 72  | 81  | 74  | 66  | 65  | 50  | 57  | 68  |
|      | 1600 | 91  | 89  | 89  | 80  | 82  | 89  | 84  | 84  | 79  | 74  | 74  | 83  |
|      | 800  | 93  | 93  | 93  | 89  | 89  | 93  | 91  | 91  | 87  | 86  | 86  | 92  |
|      | 1600 | 94  | 94  | 95  | 93  | 93  | 94  | 93  | 94  | 92  | 93  | 91  | 93  |
Figure 1: Estimation accuracy for $\mathbf{I}_0$ and $\mathbf{I}_0^{-1}$. In each setting of $\sigma^2$, the first row stands for the relative estimation errors regarding $\mathbf{I}_0$, and the second row for $\mathbf{I}_0^{-1}$. Four estimators are compared: black lines for the local Hessian $\mathbf{H}^{(l)}(\theta)$ (LC), green lines for the global Hessian $\mathbf{H}^{(g)}(\theta)$ (GL), the red solid lines for $\hat{\Omega}(\theta)^{-1}$ (GM), and the red dashed lines for $\hat{\mathbf{I}}_0(\theta)$ (MG). Their inverses are used to estimate $\mathbf{I}_0^{-1}$ in the second row.
Figure 2: Relative $l_2$ distance to the oracle estimator. Back lines stand for the CSL method, and green lines for the Newton estimator (GL) with the global Hessian. Red solid lines stand for $\tilde{\theta}_t$ (GM), and red dashed lines for $\tilde{\theta}_t'$ (MG).
Figure 3: Plots of the one-step estimators versus the oracle estimator. The oracle estimator is on the x-axis. The y-axis includes four estimators: the green line stands for the average of $M$-estimators (AVG), the black line for the one-shot CSL method (CSL-OS), the red solid line for $\tilde{\theta}_{os}$ (GM-OS), and the red dashed line for $\tilde{\theta}_{os}'$ (MG-OS). Inference for the first element of $\theta_0$ in Model 1 is presented.
References

Acosta, A. D. (1981). Inequalities for $b$-valued random vectors with applications to the strong law of large numbers. *Ann. Probab.*, 9(1):157–161.

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *Ann. Statist.*, 46(3):1352–1382.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Cen, S., Zhang, H., Chi, Y., Chen, W., and Liu, T. (2020). Convergence of distributed stochastic variance reduced methods without sampling extra data. *IEEE Transactions on Signal Processing*, 68:3976–3989.

Chang, C., Deng, Y., Jiang, X., and Long, Q. (2020). Multiple imputation for analysis of incomplete data in distributed health data networks. *Nature Communications*, 11(1):5467.

Chen, L. H., Goldstein, L., and Shao, Q.-M. (2011). *Normal approximation by Stein’s method*, volume 2. Springer.

Chen, R. Y., Gittens, A., and Tropp, J. A. (2012). The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference: A Journal of the IMA*, 1(1):2–20.

Chen, T., Giannakis, G., Sun, T., and Yin, W. (2018). Lag: Lazily aggregated gradient for communication-efficient distributed learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 5050–5060. Curran Associates, Inc.

Chen, X., Liu, W., Mao, X., and Yang, Z. (2020). Distributed high-dimensional regression under a quantile loss function. *Journal of Machine Learning Research*, 21(182):1–43.

Chen, X., Liu, W., and Zhang, Y. (2021). First-order newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, 0(0):1–17.
Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309 – 2352.

Dinan, T., Beider, P., and Wylie, D. (2019). The national flood insurance program: Is it financially sound? *Risk Management and Insurance Review*, 22(1):15–38.

Dobriban, E. and Sheng, Y. (2021). Distributed linear regression by averaging. *The Annals of Statistics*, 49(2):918 – 943.

Duan, R., Boland, M. R., Liu, Z., Liu, Y., Chang, H. H., Xu, H., Chu, H., Schmid, C. H., Forrest, C. B., Holmes, J. H., Schuemie, M. J., Berlin, J. A., Moore, J. H., and Chen, Y. (2019). Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*, 27(3):376–385.

Duan, R., Luo, C., Schuemie, M. J., Tong, J., Liang, C. J., Chang, H. H., Boland, M. R., Bian, J., Xu, H., Holmes, J. H., Forrest, C. B., Morton, S. C., Berlin, J. A., Moore, J. H., Mahoney, K. B., and Chen, Y. (2020). Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. *Journal of the American Medical Informatics Association*, 27(7):1028–1036.

Duan, R., Ning, Y., and Chen, Y. (2021). Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, 109(1):67–83.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65.

Fan, J., Guo, Y., and Wang, K. (2021a). Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*, 0(0):1–11.

Fan, J., Han, F., and Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, 1(2):293–314.
Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2021b). *Statistical foundations of data science*. CRC Press.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148.

Fan, J., Wang, D., Wang, K., and Zhu, Z. (2019). Distributed estimation of principal eigenspaces. *Ann. Statist.*, 47(6):3009–3031.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Taylor & Francis Inc.

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

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

Kousky, C. and Michel-Kerjan, E. (2017). Examining flood insurance claims in the united states: Six key findings. *Journal of Risk and Insurance*, 84(3):819–850.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. (2017). Communication-efficient sparse regression. *J. Mach. Learn. Res.*, 18(1):115–144.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag, New York, NY, USA, second edition.
Li, B., Cen, S., Chen, Y., and Chi, Y. (2020a). Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. In Chiappa, S. and Calandra, R., editors, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of Proceedings of Machine Learning Research, pages 1662–1672, Online. PMLR.

Li, M. and Zhao, J. (2021). Communication-efficient distributed linear discriminant analysis for binary classification. Statistica Sinica.

Li, R., Chen, Y., Ritchie, M. D., and Moore, J. H. (2020b). Electronic health records and polygenic risk scores for predicting disease risk. Nature Reviews Genetics, 21(8):493–502.

Li, X., Li, R., Xia, Z., and Xu, C. (2020c). Distributed feature screening via componentwise debiasing. Journal of Machine Learning Research, 21(24):1–32.

Liu, Q. and Ihler, A. T. (2014). Distributed estimation, information loss and exponential families. In NIPS.

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B., and Tropp, J. A. (2014). Matrix concentration inequalities via the method of exchangeable pairs. Ann. Probab., 42(3):906–945.

Nadarajah, S. and Pogány, T. K. (2016). On the distribution of the product of correlated normal random variables. Comptes Rendus Mathematique, 354(2):201 – 204.

Rosenblatt, J. D. and Nadler, B. (2016). On the optimality of averaging in distributed statistical learning. Information and Inference: A Journal of the IMA, 5(4):379–404.

Röllin, A. (2013). Stein’s method in high dimensions with applications. Annales de l’Institut Henri Poincaré, Probabilités et Statistiques, 49(2):529 – 549.

Shamir, O., Srebro, N., and Zhang, T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. In International conference on machine learning, pages 1000–1008.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory, volume 6, pages 583–603. University of California Press.
Stein, C. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135 – 1151.

Su, W., Bogdan, M., and Candès, E. (2017). False discoveries occur early on the lasso path. *Ann. Statist.*, 45(5):2133–2150.

Su, W. J. (2018). When is the first spurious variable selected by sequential regression procedures? *Biometrika*, 105(3):517–527.

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for -penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166 – 1202.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer New York.

Volgushev, S., Chao, S.-K., and Cheng, G. (2019). Distributed inference for quantile regression processes. *Ann. Statist.*, 47(3):1634–1662.

Wang, J., Kolar, M., Srebro, N., and Zhang, T. (2017). Efficient distributed learning with sparsity. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3636–3645, International Convention Centre, Sydney, Australia. PMLR.

Wang, X., Yang, Z., Chen, X., and Liu, W. (2019). Distributed inference for linear support vector machine. *Journal of Machine Learning Research*, 20(113):1–41.

Yu, Y., Chao, S.-K., and Cheng, G. (2020). Simultaneous inference for massive data: distributed bootstrap. In *International Conference on Machine Learning*, pages 10892–10901. PMLR.
Yu, Y., Chao, S.-K., and Cheng, G. (2021). Distributed bootstrap for simultaneous inference under high dimensionality.

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

Zhang, Y., Duchi, J., and Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(102):3299–3340.

Zhang, Y., Duchi, J. C., and Wainwright, M. J. (2013). Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.*, 14(1):3321–3363.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563.
Appendix

Proof of Proposition 2.1

First, define $Q_{12}(u) = (I \otimes u^\top)Q_{12}$, $Q_{12}(u^{\otimes 2}) = Q_{12} \circ u = (I \otimes u^\top)Q_{12}u$, and $Q(u^{\otimes 2}) = Q \circ u = (I \otimes u^\top)Qu$. Let $\hat{d}_i = \theta_i - \bar{\theta}$ and $\hat{g}_i = l_i(\theta) - \bar{I}(\theta)$.

The First Statement

By definition,

$$l_i(\theta_0) = -I_0 \hat{d}_i + I_0 \hat{d}_i + l_i(\theta_0) = -I_0 \hat{d}_i + I_0 (d_i + \theta_0 - \bar{\theta}) + l_i(\theta_0)$$

$$\hat{g}_i = l_i(\theta_0) + (\bar{H}_i - \bar{H}) \Delta - \bar{I}(\theta_0)$$

Let $e_i = I_0(\theta_0 - \bar{\theta}) + \{I_0d_i + l_i(\theta_0)\} + (\bar{H}_i - \bar{H}) \Delta - \bar{I}(\theta_0)$, then

$$\hat{I}_0 - I_0 = \left( m^{-1} \sum \hat{d}_i \hat{d}_i^\top \right)^{-1} \left( m^{-1} \sum \hat{d}_i e_i^\top \right).$$

Recall the definition $d_i = \bar{\theta}_i - \theta_0$ and $\hat{d}_i = d_i + \theta_0 - \bar{\theta}$. Note that $\sum \hat{d}_i = \sum e_i = 0$, $m^{-1} \sum d_i = \bar{\theta} - \theta_0$, and

$$m^{-1} \sum \hat{d}_i e_i^\top = m^{-1} \sum (d_i + \theta_0 - \bar{\theta}) \left\{ I_0(\theta_0 - \bar{\theta}) + \{I_0d_i + l_i(\theta_0)\} + (\bar{H}_i - \bar{H}) \Delta - \bar{I}(\theta_0) \right\}^\top$$

$$= m^{-1} \sum \hat{d}_i \{ l_i(\theta_0) + I_0d_i \}^\top - (\bar{\theta} - \theta_0)(\bar{\theta} - \theta_0)^\top I_0 + m^{-1} \sum d_i \Delta^\top (\bar{H}_i - \bar{H}) - (\bar{\theta} - \theta_0) \bar{I}(\theta_0)^\top.$$

Lemma 4.12 indicates $\|\bar{I}(\theta_0)\|_2 = O_P(n^{-1/2}m^{-1/2})$, and Lemma 4.1 states $\|\bar{\theta} - \theta_0\|_2 = O_P\{(nm)^{-1/2} + n^{-1}\}$. Also, under Assumptions 1-3, Lemma 4.2 states $m^{-1} \sum d_i \{ l_i(\theta_0) + I_0d_i \}^\top_2 = O_P(m^{-1/2}n^{-3/2} + n^{-2})$. Under Assumptions 1-4, Lemma 4.6 states

$$m^{-1} \sum \bar{H}_i - \bar{H}) \Delta d_i^\top = -n^{-1}Q_{12}^{\top} \times (I_{d \times d} \otimes \Delta) \times I_0^{-1} + R_{d11},$$

27
where \( \|R_{d1}\|_2 = O_p(n^{-1/2}) \|\Delta\|_2^3 + O_p\{n^{-1} + (nm)^{-1/2}\} \|\Delta\|_2^2 + O_p\{n^{-3/2} + n^{-1}m^{-1/2}\} \|\Delta\|_2. \) Putting them together, we have

\[
\left\| m^{-1} \sum d_i e_i^\top + n^{-1}I_0^{-1}Q_{12}(\Delta) \right\|_2 \\
\leq \left\| m^{-1} \sum d_i (I_0(\theta_0) + I_0 d_i)^\top \right\|_2 + \left\| (\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^\top I_0 \right\|_2 + \\
\left\| m^{-1} \sum d_i \Delta^\top (\hat{H}_i - \bar{H}) + n^{-1}I_0^{-1}Q_{12}(\Delta) \right\|_2 + \left\| (\hat{\theta} - \theta_0)\bar{I}(\theta_0)^\top \right\|_2 \\
= O_p(n^{-1/2}) \|\Delta\|_2^3 + O_p\{n^{-1} + (nm)^{-1/2}\} \|\Delta\|_2^2 + O_p\{n^{-3/2} + n^{-1}m^{-1/2}\} \|\Delta\|_2 + \\
O_p(n^{-1}m^{-1} + n^{-2}).
\]

On the other hand, Lemma 4.3 indicates \( \|m^{-1} \sum n\hat{d}_i \hat{d}_i^\top - I_0^{-1}Q_{11}I_0^{-1}\|_2 = O_p(n^{-1/2} + m^{-1/2}). \) The matrix inverse transformation in continuous. Therefore, by the continuous mapping theorem,

\[
\left\| \left( m^{-1} \sum n\hat{d}_i \hat{d}_i^\top \right)^{-1} - (I_0^{-1}Q_{11}I_0^{-1})^{-1} \right\|_2 \\
\leq \left\| m^{-1} \sum n\hat{d}_i \hat{d}_i^\top \right\|_2 \left\| m^{-1} \sum n\hat{d}_i \hat{d}_i^\top - I_0^{-1}Q_{11}I_0^{-1} \right\|_2 O(1) \\
= O_p(n^{-1/2} + m^{-1/2}).
\]

Note that this part does not depend on \( \Delta. \)

\[
\hat{I}_0 - I_0 \\
= \left( m^{-1} \sum \hat{d}_i \hat{d}_i^\top \right)^{-1} \left( m^{-1} \sum \hat{d}_i e_i^\top \right) \\
= \left\{ \left( m^{-1} \sum n\hat{d}_i \hat{d}_i^\top \right)^{-1} - I_0 Q_{11}^{-1}I_0 + I_0 Q_{11}^{-1}I_0 \right\} \left\{ m^{-1} \sum n\hat{d}_i e_i^\top - I_0^{-1}Q_{12}(\Delta) + I_0^{-1}Q_{12}(\Delta) \right\} \\
= \left\{ \left( m^{-1} \sum n\hat{d}_i \hat{d}_i^\top \right)^{-1} - I_0 Q_{11}^{-1}I_0 \right\} \left( m^{-1} \sum n\hat{d}_i e_i^\top \right) + \\
( I_0 Q_{11}^{-1}I_0 ) \left\{ m^{-1} \sum n\hat{d}_i e_i^\top - I_0^{-1}Q_{12}(\Delta) \right\} + I_0 Q_{11}^{-1}Q_{12}(\Delta).
\]
Let
\[ R_{D1} = \left\{ \left( m^{-1} \sum n\hat{d}_i\hat{d}_i^\top \right)^{-1} - I_0 Q_{11}^{-1} I_0 \right\} \left( m^{-1} \sum n\hat{d}_i e_i^\top \right) \]
\[ R_{D2} = (I_0 Q_{11}^{-1} I_0) \left\{ m^{-1} \sum n\hat{d}_i e_i^\top - I_0^{-1} Q_{12}(\Delta) \right\}, \]

with
\[
\|R_{D1}\|_2 = O_p(1 + n^{1/2}m^{-1/2})\|\Delta\|^2 + O_p(n^{-1/2} + m^{-1/2})\|\Delta\|^2 + O_p(n^{-1} + m^{-1})\|\Delta\|_2 + O_p(m^{-3/2} + n^{-3/2}) + O_p(n^{-1/2} + m^{-1/2})\|Q_{12}(\Delta)\|_2
\]
\[
\|R_{D2}\|_2 = O_p(n^{1/2})\|\Delta\|^3 + O_p(1 + n^{1/2}m^{-1/2})\|\Delta\|^2 + O_p(n^{-1/2} + m^{-1/2})\|\Delta\|_2 + O_p(m^{-1} + n^{-1}).
\]

Define \( R_D = R_{D1} + R_{D2} \), and we complete the proof.

**The Second Statement of Proposition 2.1**

By the definition,
\[
\hat{g}_i = I_i(\theta_0) + (\tilde{H}_i - \tilde{H})\Delta - I(\theta_0)
\]
\[
= -I_0 d_i + I_0(\theta_0 - \theta) + \{I_0 d_i + I_1(\theta_0)\} + (\tilde{H}_i - \tilde{H})\Delta - I(\theta_0)
\]
\[
\hat{d}_i = -I_0^{-1}\hat{g}_i + (\theta_0 - \tilde{\theta}) + \{d_i + I_0^{-1}I_i(\theta_0)\} + I_0^{-1}(\tilde{H}_i - \tilde{H})\Delta - I_0^{-1}(\theta_0).
\]

Let \( e_i = (\theta_0 - \tilde{\theta}) + \{d_i + I_0 I_i(\theta_0)\} + I_0^{-1}(\tilde{H}_i - \tilde{H})\Delta - I_0^{-1}(\theta_0) \). Then,
\[
\hat{\Omega} - I_0^{-1} = \left( m^{-1} \sum \hat{g}_i \hat{g}_i^\top \right)^{-1} \left( m^{-1} \sum \hat{g}_i e_i^\top \right).
\]

First, Lemma 4.4 states, under Assumption 13
\[
m^{-1} \sum \hat{g}_i \hat{g}_i^\top = n^{-1} Q_{11} + R_G,
\]

29
where $\|R_G\|_2 = O_p(1)\|\Delta\|_2^2 + O_p(n^{-1/2})\|\Delta\|_2 + O_p(n^{-1})\|\Delta\|_2 + O_p(n^{-1/2})$. Therefore, by the continuous mapping theorem, when $\|\Delta\|_2 = o_p(n^{-1/4})$,

$$(m^{-1} \sum \hat{g}_i \hat{g}_i^\top)^{-1} \leadsto Q_{11}^{-1}.$$

Recall the definition $d_i = \hat{\theta}_i - \theta_0$ and $\tilde{d}_i = d_i + \theta_0 - \bar{\theta}$. Note that $\sum \hat{g}_i = \sum e_i = 0$, $m^{-1} \sum d_i = \theta - \theta_0$, and

$$m^{-1} \sum \hat{g}_i e_i^\top = m^{-1} \sum \{l_i(\theta_0) + (\bar{H}_i - \bar{\bar{H}})\Delta - \bar{l}(\theta_0)\} \times$$

$$\left[ (\theta_0 - \bar{\theta}) + \{d_i + I_0^{-1} l_i(\theta_0)\} + I_0^{-1}(\bar{H}_i - \bar{\bar{H}})\Delta - I_0^{-1}\bar{l}(\theta_0) \right] ^\top$$

$$= \bar{l}(\theta_0)(\theta_0 - \bar{\theta}) + m^{-1} \sum l_i(\theta_0)\{d_i + I_0^{-1} l_i(\theta_0)\} ^\top +$$

$$m^{-1} \sum l_i(\theta_0)\Delta ^\top(\bar{H}_i - \bar{\bar{H}})I_0^{-1} - \bar{l}(\theta_0)\bar{l}(\theta_0) ^\top I_0^{-1}$$

$$m^{-1} \sum (\bar{H}_i - \bar{\bar{H}})\Delta \{d_i + I_0^{-1} l_i(\theta_0)\} ^\top +$$

$$m^{-1} \sum (\bar{H}_i - \bar{\bar{H}})\Delta \Delta ^\top(\bar{H}_i - \bar{\bar{H}})I_0^{-1}.$$

Lemma 4.12 indicates $\|\bar{l}(\theta_0)\|_2 = O_p(n^{-1/2}m^{-1/2})$, and Lemma 4.1 states $\|\bar{\theta} - \theta_0\|_2 = O_p\{(nm)^{-1/2} + n^{-1}\}$. Lemma 4.5 indicates, under Assumptions 1-4,

$$m^{-1} \sum l_i(\theta_0)\Delta ^\top(\bar{H}_i - \bar{\bar{H}}) = n^{-1}(I_{d \times d} \otimes \Delta ^\top) \times Q_{12} + R_{22},$$

where $\|R_{22}\|_2 = O_p(n^{-1/2})\|\Delta\|_2^3 + O_p\{n^{-1} + (nm)^{-1/2}\}\|\Delta\|_2^2 + O_p(n^{-1}m^{-1/2})\|\Delta\|_2$.

Lemma 4.7 states when Assumptions 1-3 hold,

$$\left\| m^{-1} \sum (\bar{H}_i - \bar{\bar{H}})\Delta \Delta ^\top(\bar{H}_i - \bar{\bar{H}}) \right\|_2 = O_p(1)\|\Delta\|_2^2 + O_p(n^{-1})\|\Delta\|_2^2.$$

Lemma 4.14 indicates $E(\|d_i - d_{0,i}\|_2^2) = O(n^{-2})$. Under Assumptions 1-3, Lemma 4.18 states

$$l_i(\theta_0)(d_i - d_{0,i}) ^\top = W_i + \bar{W}_i,$$

with $E(\|W_i\|_2^2) = O(n^{-3})$, $E(\|W_i\|_2) = O(n^{-2})$, and $E(\|\bar{W}_i\|_2) = O(n^{-2})$. Also, by Lemma
\[ \mathbb{E} \left( \left\| m^{-1} \sum W_i \right\|_2^2 \right) = O(m^{-1}n^{-3} + n^{-4}). \]

Hence,
\[
\left\| m^{-1} \sum l_i(\theta_0)\{l_i(\theta_0) + I_0d_i\}^T \right\|_2 \leq \left\| m^{-1} \sum W_i \right\|_2 + \left\| m^{-1} \sum W_i \right\|_2 = O_p(m^{-1/2}n^{-3/2} + n^{-2}).
\]

Also, consider \( m^{-1} \sum (\tilde{H}_i - \bar{H})\Delta\{d_i + I_0^{-1}l_i(\theta_0)\}^T \). By the Cauchy–Schwarz inequality
\[
\left\| m^{-1} \sum (\tilde{H}_i - \bar{H})\Delta\{d_i + I_0^{-1}l_i(\theta_0)\} \right\|_2 \leq m^{-1} \sum \left\| (\tilde{H}_i - \bar{H})\Delta \right\|_2 \| d_i - d_{0,i} \|_2 \\
\leq \left\{ m^{-1} \sum \left\| (\tilde{H}_i - \bar{H})\Delta \right\|_2^2 \right\}^{1/2} \left\{ m^{-1} \sum \| d_i - d_{0,i} \|_2^2 \right\}^{1/2} \\
= O_p(n^{-1})\| \Delta \|_2^2 + O_p(n^{-3/2})\| \Delta \|_2.
\]

The last inequality comes from Lemmas 4.7 and 4.14. Putting them together, we have
\[
\left\| m^{-1} \sum \hat{g}_i e_i^T - n^{-1}Q_{12}(\Delta)I_0^{-1} \right\|_2 \\
= O_p(1)\| \Delta \|_2^4 + O_p(n^{-1/2})\| \Delta \|_2^3 + O_p\{n^{-1} + (nm)^{-1/2}\}\| \Delta \|_2^2 + \\
O_p(n^{-3/2} + n^{-1}m^{-1/2})\| \Delta \|_2 + O_p(n^{-1}m^{-1} + n^{-2}).
\]

We assume \( \| \Delta \|_2 = o_p(n^{-1/4}) \), so that \( m^{-1} \sum n\hat{g}_i \hat{g}_i^T \to Q_{11}, \| (m^{-1} \sum n\hat{g}_i \hat{g}_i^T)^{-1} - Q_{11}^{-1} \| = \)
The first estimator has the decomposition:

\[ o_p(1), \text{ and } \|m^{-1} \sum n \tilde{g}_i e_i^\top - Q_{12}(\Delta) I_0^{-1}\|_2 = o_p(1). \]

\[ \begin{align*}
\hat{\Omega} - I_0^{-1} &= \left(m^{-1} \sum \tilde{g}_i \tilde{g}_i^\top \right)^{-1} \left(m^{-1} \sum \tilde{g}_i e_i^\top \right) \\
&= \left\{ \left(m^{-1} \sum \tilde{g}_i \tilde{g}_i^\top \right)^{-1} - Q_{11}^{-1} \right\} \left(m^{-1} \sum n \tilde{g}_i e_i^\top - Q_{12}(\Delta) I_0^{-1} + Q_{12}(\Delta) I_0^{-1} \right) \\
&= \left\{ \left(m^{-1} \sum \tilde{g}_i \tilde{g}_i^\top \right)^{-1} - Q_{11}^{-1} \right\} \left(m^{-1} \sum n \tilde{g}_i e_i^\top \right) + \\
& \quad Q_{11}^{-1} \left\{ m^{-1} \sum n \tilde{g}_i e_i^\top - Q_{12}(\Delta) I_0^{-1} \right\} + Q_{11}^{-1} Q_{12}(\Delta) I_0^{-1}.
\end{align*} \]

Let

\[ R_{G1} = Q_{11}^{-1} \left\{ m^{-1} \sum n \tilde{g}_i e_i^\top - Q_{12}(\Delta) I_0^{-1} \right\}, \]

\[ R_{G2} = \left\{ \left(m^{-1} \sum \tilde{g}_i \tilde{g}_i^\top \right)^{-1} - Q_{11}^{-1} \right\} \left(m^{-1} \sum n \tilde{g}_i e_i^\top \right), \]

with

\[ \|R_{G1}\|_2 = O_p(n) \|\Delta\|_2^4 + O_p(n^{1/2}) \|\Delta\|_2^3 + O_p(1 + n^{1/2} m^{-1/2}) \|\Delta\|_2 + O_p(n^{-1/2} + m^{-1} + n^{-1}), \]

\[ \|R_{G2}\|_2 = o_p(\|R_{G1}\|_2) + \left\{ O_p(n^{1/2}) \|\Delta\|_2^5 + O_p(1) \|\Delta\|_2 + O_p(m^{-1/2}) \right\} \|Q_{12}(\Delta)\|_2. \]

Define \( R_G = R_{G1} + R_{G2} \), and we complete the proof.

**Proof of Theorem 3.1**

**The first statement**

The first estimator has the decomposition:

\[ \begin{align*}
\tilde{\theta}_{t+1} - \theta_0 &= \tilde{I}_0^{-1} (I_0 - I_0) \Delta_t + \tilde{I}_0^{-1} (I_0 - \overline{H}(\theta))^T \Delta_t - \tilde{I}_0^{-1} 1(\theta_0) \\
\tilde{\theta}_{t+1} - \theta^* &= \tilde{I}_0^{-1} (I_0 - I_0) \Delta_t + \tilde{I}_0^{-1} (I_0 - \overline{H}(\theta))^T \Delta_t + \tilde{I}_0^{-1} (I_0 - \overline{H}(\theta)^*) (\overline{H}(\theta)^*)^{-1}(\theta_0),
\end{align*} \]

32
where $\mathbf{H}^* = \int_0^1 \sum_{j \in g} \nabla^2 L(\theta_0 + t \Delta^*; \mathbf{X}_j) dt$. Proposition 2.1 states when Assumptions 1-4 hold and $\hat{\theta}_t \in U(\rho)$, $\hat{I}_0 - I_0 = I_0 Q_{11}^{-1} Q_{12}(\Delta_t) + R_D$, where $\|R_D\|_2 = O_P(n^{1/2})\|\Delta_t\|^3_2 + O_P(1 + n^{1/2}m^{-1/2})\|\Delta_t\|^2_2 + O_P(n^{-1/2} + m^{-1/2})\|\Delta_t\|_2 + O_P(m^{-1} + n^{-1})$. When $\|\Delta_t\|_2 = O_P(n^{-1/2})$, $\|R_D\|_2 = O_P(m^{-1} + n^{-1})$, and $\|\hat{I}_0 - I_0\|_2 = O_P(\|\Delta_t\|_2 + m^{-1} + n^{-1})$. Consequently,

$$
\hat{I}_0^{-1}(\hat{I}_0 - I_0)\Delta_t = (\hat{I}_0^{-1} - I_0^{-1})\{I_0 Q_{11}^{-1} Q_{12}(\Delta_t) + R_D\} \Delta_t
$$

$$
= Q_{11}^{-1} Q_{12}(\Delta_t^2) + \hat{I}_0^{-1}(I_0 - \hat{I}_0)I_0^{-1}(\hat{I}_0 - I_0)\Delta_t + I_0^{-1}R_D \Delta_t.
$$

Let $R_{11} = \hat{I}_0^{-1}(I_0 - \hat{I}_0)I_0^{-1}(\hat{I}_0 - I_0)\Delta_t + I_0^{-1}R_D \Delta_t$, then $\|R_{11}\|_2 = O_P(\|\Delta_t\|^3_2 + O_P(m^{-1} + n^{-1})\|\Delta_t\|_2)$.

On the other hand, Lemma 4.8 gives $(I_0 - \hat{H}^{(g)})\Delta_t = -2^{-1} Q(\Delta_t^2) + R_B$, where $\|R_B\| = O_P(\|\Delta_t\|^2_2 + \|\Delta_t\|^2_2 + n^{-1/2}m^{-1/2}\|\Delta_t\|_2)$. Hence,

$$
\hat{I}_0^{-1}(I_0 - \hat{H}^{(g)}) \Delta_t = (\hat{I}_0^{-1} - I_0^{-1} + I_0^{-1})\{-2^{-1} Q(\Delta_t^2) + R_B\}
$$

$$
= -2^{-1} I_0^{-1} Q(\Delta_t^2) + I_0^{-1} R_B + \hat{I}_0^{-1}(I_0 - \hat{I}_0)I_0^{-1}(I_0 - \hat{H}^{(g)}) \Delta_t.
$$

Let $R_{22} = I_0^{-1} R_B + \hat{I}_0^{-1}(I_0 - \hat{I}_0)I_0^{-1}(I_0 - \hat{H}^{(g)}) \Delta_t$. We have $\|(I_0 - \hat{H}^{(g)}) \Delta_t\| = O_P(\|\Delta_t\|^2_2 + n^{-1/2}m^{-1/2}\|\Delta_t\|_2)$ and $\|R_{22}\|_2 = O_P(n^{-1/2}\|\Delta_t\|^2_2 + n^{-1/2}m^{-1/2}\|\Delta_t\|_2)$.

Consider the last term of $\Delta_{t+1} - \hat{I}_0^{-1} Q(\theta_0)$. Note that $\Delta^* = -(\hat{H}^{-1})^{-1}(\theta_0)$, and we state without proof that $\|\Delta^*\|_2 = O_P(n^{-1/2}m^{-1/2})$. Let $H_0^{(g)} = n^{-1} m^{-1} \sum_{j \in g} \nabla^2 L(\theta_0; \mathbf{X}_j)$. Lemma 4.13 and 4.12 indicate

$$
\|(\hat{H}^* - I_0) \Delta^*\|_2 \leq \|(\hat{H}^* - H_0^{(g)}) \Delta^*\|_2 + \|(H_0^{(g)} - I_0) \Delta^*\|_2 = O_P(n^{-1} m^{-1} + n^{-1/2}m^{-1/2}\|\Delta^*\|_2).
$$

Therefore, $\|I^{(g)}(\theta_0)\|_2 = \|\hat{H}^* \Delta^*\|_2 = O_P(\|\Delta^*\|_2 + n^{-1} m^{-1})$, and

$$
\|\hat{I}_0^{-1}(\hat{I}_0 - \hat{H}^*) (H^*)^{-1} I^{(g)}(\theta_0)\|_2 \leq \|I_0^{-1}\|_2 \left\{\|\hat{I}_0 - I_0\|_2 \|\Delta^*\|_2 + \|(I_0 - \hat{H}^*) \Delta^*\|_2\right\}
$$

$$
= O_P(n^{-1} m^{-1} + n^{-1/2}m^{-1/2}\|\Delta^*\|_2 + n^{-1} m^{-1}).
$$

Hence,

$$
\hat{\theta}_{t+1} - \theta^* = Q_{11}^{-1} Q_{12}(\Delta_t^2) - 2^{-1} I_0^{-1} Q(\Delta_t^2) + R_{11} + R_{22} + \hat{I}_0^{-1}(\hat{I}_0 - \hat{H}^*) (H^*)^{-1} I^{(g)}(\theta_0)
$$

$$
= Q_{11}^{-1} Q_{12}(\Delta_t^2) - 2^{-1} I_0^{-1} Q(\Delta_t^2) + R_1^*.
$$
where $\mathbf{R}_1^* = \mathbf{R}_{11} + \mathbf{R}_{22} - \mathbf{I}_0^{-1} (\mathbf{I}_0 - \mathbf{H}^*) \Delta^*$ with

$$
\| \mathbf{R}_1^* \|_2 \leq \| \mathbf{R}_{11} \|_2 + \| \mathbf{R}_{12} \|_2 + \| \mathbf{I}_0^{-1} \|_2 \left\{ \| \mathbf{I}_0 - \mathbf{I}_0 \|_2 \| \Delta^* \|_2 + \| (\mathbf{I}_0 - \mathbf{H}^*) \Delta^* \|_2 \right\} = O_P(m^{-1} + n^{-1}) (\| \Delta_t \|_2 + \| \Delta^* \|_2) + O_P(m^{-1}n^{-1}).
$$

The second statement

Similarly, let $\Delta'_{t+1} = \tilde{\theta}'_{t+1} - \theta_0$, and

$$
\begin{align*}
\tilde{\theta}'_{t+1} - \theta_0 &= (\mathbf{I} - \tilde{\Omega} \mathbf{H}^{(g)}(\theta)) \Delta_t - \tilde{\Omega} l^{(g)}(\theta_0) \\
&= \mathbf{I}_0^{-1}(\mathbf{I}_0 - \mathbf{H}^{(g)}(\theta_0)) \Delta_t - (\mathbf{I}_0 - \mathbf{I}_0^{-1}) \mathbf{H}^{(g)}(\theta_0) \Delta_t - \tilde{\Omega} l^{(g)}(\theta_0), \\
\tilde{\theta}'_{t+1} - \theta^* &= \mathbf{I}_0^{-1}(\mathbf{I}_0 - \mathbf{H}^{(g)}(\theta_0)) \Delta_t - (\mathbf{I}_0 - \mathbf{I}_0^{-1}) \mathbf{H}^{(g)}(\theta_0) \Delta_t - \{\mathbf{I} - (\mathbf{H}^*)^{-1}\} l^{(g)}(\theta_0).
\end{align*}
$$

For the first component, Lemma 4.8 gives $(\mathbf{I}_0 - \mathbf{H}^{(g)}(\theta_0)) \Delta_t = -2^{-1} \mathbf{Q}(\Delta_t^{\otimes 2}) + \mathbf{R}_H$, where $\| \mathbf{R}_H \|_2 = O_P(n^{-1/2}) \| \Delta_t \|_2^2 + O_P(n^{-1/2}m^{-1/2}) \| \Delta_t \|_2$. When $\| \Delta_t \|_2 = O_P(n^{-1/2})$, $\| (\mathbf{I}_0 - \mathbf{H}^{(g)}(\theta_0)) \Delta_t \| = O_P(1) \| \Delta_t \|_2^2 + O_P(n^{-1/2}m^{-1/2}) \| \Delta_t \|_2$. Therefore,

$$
\mathbf{I}_0^{-1}(\mathbf{I}_0 - \mathbf{H}^{(g)}(\theta_0)) \Delta_t = -2^{-1} \mathbf{I}_0^{-1} \mathbf{Q}(\Delta_t^{\otimes 2}) + \mathbf{I}_0^{-1} \mathbf{R}_H.
$$

Next, by Proposition 2.1 $\hat{\Omega} - \mathbf{I}_0^{-1} = \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}(\Delta_t) \mathbf{I}_0^{-1} + \mathbf{R}_G$, where $\| \mathbf{R}_G \|_2 = O_P(n^{-1/2} + m^{-1/2}) \| \Delta_t \|_2 + O_P(m^{-1} + n^{-1})$.

$$
(\hat{\Omega} - \mathbf{I}_0^{-1}) \mathbf{H}^{(g)}(\theta_0) \Delta_t = \{\mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}(\Delta_t) \mathbf{I}_0^{-1} + \mathbf{R}_G\}(\mathbf{H}^{(g)}(\theta_0) - \mathbf{I}_0 - \mathbf{I}_0) \Delta_t \\
= \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}(\Delta_t^{\otimes 2}) + \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}(\Delta_t) \mathbf{I}_0^{-1} (\mathbf{H}^{(g)}(\theta_0) - \mathbf{I}_0) \Delta_t + \mathbf{R}_G \mathbf{H}^{(g)}(\theta_0) \Delta_t,
$$

where

$$
\begin{align*}
\| \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}(\Delta_t) \mathbf{I}_0^{-1} (\mathbf{H}^{(g)}(\theta_0) - \mathbf{I}_0) \Delta_t \|_2 &= O_P(1) \| \Delta_t \|_2^3 + O_P(n^{-1/2}m^{-1/2}) \| \Delta_t \|_2^3 \\
\| \mathbf{R}_G \mathbf{H}^{(g)}(\theta_0) \Delta_t \|_2 &= O_P(n^{-1/2} + m^{-1/2}) \| \Delta_t \|_2^2 + O_P(m^{-1} + n^{-1}) \| \Delta_t \|_2.
\end{align*}
$$

Putting them together, we have

$$
\Delta'_{t+1} = -2^{-1} \mathbf{I}_0^{-1} \mathbf{Q}(\Delta_t^{\otimes 2}) + \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}(\Delta_t^{\otimes 2}) + \mathbf{R}_1' - \hat{\Omega} l^{(g)}(\theta_0),
$$

34
where $R'_t = I_0^{-1} R_H + Q_{11}^{-1} Q_{12} (\Delta_t) I_0^{-1} (\bar{H}^{(g)} - I_0) \Delta_t + R_G \bar{H}^{(g)} \Delta_t$, with $\|R'_t\|_2 = O_p(n^{-1/2} + m^{-1/2}) \|\Delta_t\|_2^2 + O_p(n^{-1} + m^{-1}) \|\Delta_t\|_2$.

Consider $-\{\hat{\Omega} - (\bar{H}^*)^{-1}\} l^{(g)}(\theta_0)$. Recall that $\|\hat{\Omega} - I_0^{-1}\|_2 = O_p(\|\Delta_t\|_2 + m^{-1} + n^{-1})$, and $\|l^{(g)}(\theta_0) + I_0 \Delta^*\|_2 = O_p(n^{-1} m^{-1} + n^{-1/2} m^{-1/2} \|\Delta^*\|_2)$.

$$\|\{\hat{\Omega} - (\bar{H}^*)^{-1}\} l^{(g)}(\theta_0)\|_2 \leq \|\hat{\Omega} - I_0^{-1}\|_2 \|l^{(g)}(\theta_0)\|_2 + \|I_0^{-1} (\bar{H}^* - I_0) \Delta^*\|_2$$

$$= O_p\{ (\|\Delta_t\|_2 + m^{-1} + n^{-1}) \|\Delta^*\|_2 + O_p(m^{-1} n^{-1}) \}.$$

Let $R_2 = R'_t - \{\hat{\Omega} - (\bar{H}^*)^{-1}\} l^{(g)}(\theta_0)$. Then,

$$\|R_2\| = O_p(n^{-1} + m^{-1})(\|\Delta_t\|_2 + \|\Delta^*\|_2) + O_p(m^{-1} n^{-1}).$$

**Proof of Proposition 3.1**

Recall that $Q_{11} = \mathbb{E}\{l(\theta_0; X_i) l(\theta_0; X_i)^\top\}$ and $Q_{12} = \mathbb{E}[l(\theta_0; X_i) \otimes \{H(\theta_0; X_i) - I_0\}]$. In this section, since all estimators are based on the local data, for parsimony, we drop the superscript $(l)$. Also, let $l = \nabla L$ and $H = \nabla^2 L$.

**Estimation of $I_0^{-1} Q(\Delta_0^{\otimes 2})$**

Consider $Q(\Delta_0^{\otimes 2})$. By definition and Lemma 4.9 and 4.10,

$$\|\{\nabla^3 L(\theta^4; X_i) - \nabla^3 L(\theta_0; X_i)\} (\hat{\Delta}_0^{\otimes 2})\|_2 \leq m(X_i) \|\hat{\Delta}_0\|_2^2 \|\Delta^4\|_2$$

$$\left\{n^{-1} \sum \nabla^3 L(\theta_0; X_i) - Q\right\} (\hat{\Delta}_0^{\otimes 2}) = 2R_2$$

$$\|Q(\hat{\Delta}_0^{\otimes 2} - \Delta_0^{\otimes 2})\|_2 \leq d\lambda_h \|\hat{\Delta}_0 + \Delta_0\|_2 \|\hat{\Delta}_0 - \Delta_0\|_2,$$
where \( \| \mathbf{R}_2 \|_2 \leq R_2 \| \hat{\mathbf{A}}_0 \|_2^2 \) with \( \mathbb{E}(\mathbf{R}_2^k) = O(n^{-k/2}) \). Recall the definition \( \hat{\mathbf{A}}_0 = \hat{\theta}_0 - \theta^A = \mathbf{A}_0 - \mathbf{A}^A \).

\[
\left\lVert \hat{\mathbf{Q}}(\hat{\mathbf{A}}_0^{\otimes 2}) - \mathbf{Q}(\mathbf{A}_0^{\otimes 2}) \right\rVert_2 
\leq n^{-1} \sum ||\{ \nabla^3 L(\theta^A; \mathbf{X}_i) - \nabla^3 L(\theta_0; \mathbf{X}_i) \}(\hat{\mathbf{A}}_0^{\otimes 2})||_2 + 
\left\lVert \left\{ n^{-1} \sum \nabla^3 L(\theta_0; \mathbf{X}_i) - \mathbf{Q} \right\}(\hat{\mathbf{A}}_0^{\otimes 2}) \right\rVert_2 + \left\lVert \mathbf{Q}(\hat{\mathbf{A}}_0^{\otimes 2} - \mathbf{A}_0^{\otimes 2}) \right\rVert_2 
= O_p(1) \| \hat{\mathbf{A}}_0 \|_2^2 \| \mathbf{A}_0 \|_2 + O_p(n^{-1/2}) \| \hat{\mathbf{A}}_0 \|_2 \| \theta^A - \hat{\theta}_0 - \mathbf{A}_0 \|_2 + O_p(n^{-1/2}) \| \mathbf{A}_0 \|_2 + O_p(n^{-1/2}) \| \mathbf{A}_0 \|_2^2 
= O_p(1) \| \mathbf{A}_0 \|_2 + O_p(1) \| \mathbf{A}_0 \|_2 + O_p(n^{-1/2}) \| \mathbf{A}_0 \|_2^2.
\]

When \( \| \mathbf{A}_0 \|_2 = O_p(n^{-1/2}) \) and \( \| \mathbf{A}_0 \|_2 = O_p(n^{-1/2}) \),

\[
\left\lVert \hat{\mathbf{Q}}(\hat{\mathbf{A}}_0^{\otimes 2}) - \mathbf{Q}(\mathbf{A}_0^{\otimes 2}) \right\rVert_2 = O_p(1) \| \mathbf{A}_0 \|_2 + O_p(1) \| \mathbf{A}_0 \|_2 + O_p(n^{-1/2}) \| \mathbf{A}_0 \|_2^2.
\]

Consider the local estimator of \( I_n^{-1} \), \( \hat{\mathbf{H}} = n^{-1} \sum \hat{\mathbf{H}}_i \) where \( \hat{\mathbf{H}}_i = \nabla^2 L(\theta^A; \mathbf{X}_i) \).

\[
\| \hat{\mathbf{H}} - \mathbf{I}_0 \|_2 \leq \| \hat{\mathbf{H}} - \mathbf{H}_0 \|_2 + \| \mathbf{I}_0 - \mathbf{H}_0 \|_2 = O_p(1) \| \mathbf{A}_0 \|_2 + O_p(n^{-1/2}) \tag{8}
\]
\[
\| \hat{\mathbf{H}}^{-1} - \mathbf{I}_0^{-1} \|_2 \leq \| \hat{\mathbf{H}}^{-1} - \mathbf{I}_0^{-1} \|_2 + \| \mathbf{I}^{-1} \|_2 \| \hat{\mathbf{H}} - \mathbf{I}_0 \|_2 = O_p(1) \| \mathbf{A}_0 \|_2 + O_p(n^{-1/2}) \tag{9}
\]

Therefore,

\[
\| \hat{\mathbf{H}}^{-1} \hat{\mathbf{Q}}(\hat{\mathbf{A}}_0^{\otimes 2}) - \mathbf{I}_0^{-1} \mathbf{Q}(\mathbf{A}_0^{\otimes 2}) \|_2 
\leq \| \hat{\mathbf{H}}^{-1} \hat{\mathbf{Q}}(\hat{\mathbf{A}}_0^{\otimes 2}) - \mathbf{I}_0^{-1} \hat{\mathbf{Q}}(\hat{\mathbf{A}}_0^{\otimes 2}) \|_2 + \| \mathbf{I}_0^{-1} \hat{\mathbf{Q}}(\hat{\mathbf{A}}_0^{\otimes 2}) - \mathbf{I}_0^{-1} \mathbf{Q}(\mathbf{A}_0^{\otimes 2}) \|_2 
= O_p(1) \| \mathbf{A}_0 \|_2 + O_p(1) \| \mathbf{A}_0 \|_2 + O_p(n^{-1/2}) \| \mathbf{A}_0 \|_2^2. \tag{10}
\]
Estimation of $Q_{11}^{-1}Q_{12} (\Delta_0^{\otimes 2})$

Recall that $l_i$ denotes $l_i^{(l)}$ and $I$ denotes $I^{(l)}$. By definition

$$
\hat{Q}_{11} = n^{-1} \sum \{ l_i (\theta^A) - I (\theta^A) \} \{ l_i (\theta^A) - I (\theta^A) \}^T
$$

$$
= n^{-1} \sum \{ l_i (\theta^A) - l_i + l_i \} \{ l_i (\theta^A) - l_i + l_i \}^T - \bar{I} (\theta^A) \bar{I} (\theta^A)^T
$$

$$
= n^{-1} \sum l_i \bar{I}^T + n^{-1} \sum \{ l_i (\theta^A) - l_i \} \bar{I}^T + n^{-1} \sum l_i \{ l_i (\theta^A) - l_i \}^T +
$$

$$
n^{-1} \sum \{ l_i (\theta^A) - l_i \} \{ l_i (\theta^A) - l_i \}^T - \bar{I} (\theta^A) \bar{I} (\theta^A)^T.
$$

By Lemma 4.16, we have

$$
E \left( \left\| n^{-1} \sum l_i l_i^T - \hat{Q}_{11} \right\|^2 \right) = O(n^{-1})
$$

$$
\left\| n^{-1} \sum l_i l_i^T - \hat{Q}_{11} \right\|_2 = O_P(n^{-1/2}).
$$

Let $\bar{H}_i = \int_0^1 \nabla^2 L (\theta_0 + t \Delta^A; X_i^{(l)}) \, dt$, then $l_i (\theta^A) = l_i (\theta_0) + \bar{H}_i \Delta^A$ and

$$
\| l_i (\theta^A) - l_i (\theta_0) \|_2 = \| (\bar{H}_i - H_i) \Delta^A + H_i \Delta^A \|_2 \leq h(X_i) \| \Delta^A \|_2^2 + \| H_i \|_2 \| \Delta^A \|_2
$$

$$
\| \bar{I} (\theta^A) - \bar{I} (\theta_0) \|_2 \leq n^{-1} \sum \| l_i (\theta^A) - l_i (\theta_0) \|_2 = O_P(1) \| \Delta^A \|_2^2 + O_P(1) \| \Delta^A \|_2.
$$
Therefore,

\[
\left\| n^{-1} \sum \{ l_i(\theta^A) - l_i \} \hat{l}_i^\top \right\|_2 \\
\leq n^{-1} \sum h(X_i)\|l_i\|_2 \|\Delta^A\|_2^2 + n^{-1} \sum \|l_i\|_2 \|H_i\|_2 \|\Delta^A\|_2 \\
= O_{\tilde{p}}(1)\|\Delta^A\|_2^4 + O_{\tilde{p}}(1)\|\Delta^A\|_2 \\
\left\| n^{-1} \sum \{ l_i(\theta^A) - l_i(\theta_0) \} \hat{l}_i \right\|_2 \\
\leq n^{-1} \sum \|l_i(\theta^A) - l_i(\theta_0)\|_2^2 \\
\leq n^{-1} \sum 2h(X_i)^2\|\Delta^A\|_2^4 + n^{-1} \sum 2 \|H_i\|_2^2 \|\Delta^A\|_2^2 \\
= O_{\tilde{p}}(1)\|\Delta^A\|_2^4 + O_{\tilde{p}}(1)\|\Delta^A\|_2^2 \\
\|I(\theta^A)I(\theta^A)^\top\|_2 \\
= \|\hat{I}(\theta^A) - \hat{I} + \hat{I}\|_2^2 \\
= O_{\tilde{p}}(1)\|\Delta^A\|_2^4 + O_{\tilde{p}}(1)\|\Delta^A\|_2^2 + O_{\tilde{p}}(n^{-1}).
\]

Putting them together, we have

\[
\|\hat{Q}_{11} - Q_{11}\|_2 = O_{\tilde{p}}(1)\|\Delta^A\|_2^4 + O_{\tilde{p}}(1)\|\Delta^A\|_2^2 + O_{\tilde{p}}(1)\|\Delta^A\|_2 + O_{\tilde{p}}(n^{-1/2}). \quad (11)
\]

Consider $Q_{12}$. By Lemma 4.19, $\|n^{-1} \sum l_i \otimes (H_i - I_0) - Q_{12}\| = O_{\tilde{p}}(n^{-1/2})$. By definition,

\[
\hat{Q}_{12} = n^{-1} \sum \{ l_i(\theta^A) - l_i(\theta_0) \} \otimes \{ \hat{H}_i - \hat{H} \} \\
= n^{-1} \sum \{ l_i(\theta^A) - l_i(\theta_0) \} \otimes \{ \hat{H}_i - I_0 + I_0 - \hat{H} \} + \]

\[
= n^{-1} \sum l_i(\theta_0) \otimes \{ \hat{H}_i - H_i + H_i - I_0 + I_0 - \hat{H} \} + \]

\[
= n^{-1} \sum l_i(\theta^A) \otimes \{ \hat{H}_i - I_0 \} + \hat{I}(\theta^A) \otimes \{ I_0 - \hat{H} \} + \]

\[
= n^{-1} \sum l_i(\theta_0) \otimes \{ \hat{H}_i - H_i \} + n^{-1} \sum l_i(\theta_0) \otimes (H_i - I_0).
\]

Recall that

\[
\|l_i(\theta^A) - l_i(\theta_0)\|_2 = \|(\hat{H}_i - H_i)\Delta^A + H_i\Delta^A\|_2 \leq h(X_i)\|\Delta^A\|_2^2 + \|H_i\|_2 \|\Delta^A\|_2 \\
\|\hat{H}_i - I_0\|_2 \leq \|\hat{H}_i - H_i\|_2 + \|H_i - I_0\|_2 \leq h(X_i)\|\Delta^A\|_2 + \|H_i - I_0\|_2.
\]

38
Then,
\[
\|I(\theta^A) \otimes \{I_0 - \hat{H}\} \| \leq \|I(\theta^A) - I(\theta_0) + I(\theta_0)\|_2 \|I_0 - H + H - \hat{H}\|_2
\]
\[
\leq \left\{ \|I(\theta_0)\|_2 + n^{-1} \sum h(X_i)\|\Delta^A\|_2^2 + n^{-1} \sum \|H_i\|_2\|\Delta^A\|_2 \right\} \left\{ \|I_0 - H\|_2 + n^{-1} \sum h(X_i)\|\Delta^A\|_2 \right\}
\]
\[
= \{ O_p(n^{-1/2}) + O_p(1)\|\Delta^A\|_2^2 + O_p(1)\|\Delta^A\|_2 \} \left\{ O_p(n^{-1/2}) + O_p(1)\|\Delta^A\|_2 \right\}
\]
\[
= O_p(1)\|\Delta^A\|_2^3 + O_p(1)\|\Delta^A\|_2^2 + O_p(n^{-1/2})\|\Delta^A\|_2 + O_p(n^{-1})
\]
\[
\|n^{-1} \sum I_i(\theta_0) \otimes \{\hat{H}_i - H_i\}\|_2 \leq n^{-1} \sum \|I_i(\theta_0)\|_2 h(X_i)\|\Delta^A\|_2 = O_p(1)\|\Delta^A\|_2
\]
\[
\|n^{-1} \sum \{I_i(\theta^A) - I_i(\theta_0)\} \otimes \{\hat{H}_i - I_0\}\|_2 \leq n^{-1} \sum \|I_i(\theta^A) - I_i(\theta_0)\|_2 \|\hat{H}_i - I_0\|_2
\]
\[
\leq n^{-1} \sum \{h(X_i)\|\Delta^A\|_2^2 + \|H_i\|_2\|\Delta^A\|_2 \} \left\{ h(X_i)\|\Delta^A\|_2 + \|H_i - I_0\|_2 \right\}
\]
\[
= n^{-1} \sum h(X_i)^2\|\Delta^A\|_2^3 + n^{-1} \sum h(X_i)(\|H_i - I_0\|_2 + \|H_i\|_2)\|\Delta^A\|_2^2 +
\]
\[
= O_p(1)\|\Delta^A\|_2^3 + O_p(1)\|\Delta^A\|_2^2 + O_p(1)\|\Delta^A\|_2
\]

Therefore,
\[
\|\hat{Q}_{12} - Q_{12}\|_2 \leq \|\hat{Q}_{12} - n^{-1} \sum I_i \otimes (H_i - I_0)\|_2 + n^{-1} \sum I_i \otimes (H_i - I_0) - Q_{12}\|_2
\]
\[
= O_p(1)\|\Delta^A\|_2^3 + O_p(1)\|\Delta^A\|_2^2 + O_p(1)\|\Delta^A\|_2 + O_p(n^{-1/2})
\]

Let \( E_{12,j} = E [l_j(\theta_0; X_i) \{H(\theta_0; X_i) - I_0\}] \) and \( Q_{12} = (E_{12,1}, \ldots, E_{12,d})^T \). By definition,
\[
(I_{dxd} \otimes u^T) \times Q_{12} \times u = \begin{pmatrix} u^T E_{12,1} u \\ \vdots \\ u^T E_{12,d} u \end{pmatrix}, \forall u \in \mathbb{R}^d.
\]

Then, \( \hat{Q}_{12}(\hat{\Delta}_0^{\otimes 2}) - Q_{12}(\Delta_0^{\otimes 2}) = (\hat{Q}_{12} - Q_{12})(\hat{\Delta}_0^{\otimes 2}) + Q_{12}(\hat{\Delta}_0^{\otimes 2}) - Q_{12}(\Delta_0^{\otimes 2}) \), that is
\[
(I_{dxd} \otimes \hat{\Delta}_0^T) \times \hat{Q}_{12} \times \hat{\Delta}_0 - (I_{dxd} \otimes \Delta_0^T) \times Q_{12} \times \Delta_0
\]
\[
= (I_{dxd} \otimes \hat{\Delta}_0^T) \times (\hat{Q}_{12} - Q_{12}) \times \hat{\Delta}_0 + (I_{dxd} \otimes \hat{\Delta}_0^T) \times Q_{12} \times \hat{\Delta}_0 -
\]
\[
(I_{dxd} \otimes \Delta_0^T) \times Q_{12} \times \hat{\Delta}_0.
\]

39
Note that
\[
\| \hat{Q}_{12} - Q_{12}(\hat{\Delta}_0^0) \|_2 \leq \| \hat{\Delta}_0 \|_2^2 \| \hat{Q}_{12} - Q_{12} \|_2
\]
\[
= O_p(1) \| \Delta^A \|_2^2 \| \hat{\Delta}_0 \|_2^2 + O_p(1) \| \Delta^A \|_2^2 \| \hat{\Delta}_0 \|_2^2 + O_p(n^{-1/2}) \| \hat{\Delta}_0 \|_2^2
\]
Proof of Theorem 3.2

In either case, Lemma 4.1 indicates \( \| \Delta^4 \|_2 = O_P\{ (nm)^{-1/2} + n^{-1} \} \). As a result, Proposition 3.1 indicates

\[
\left\| \hat{H}^{-1} \hat{Q}(\hat{\Delta}_0^{e_2}) - I_0^{-1} Q(\Delta_0^{e_2}) \right\|_2 = O_P(n^{-1}m^{-1/2} + n^{-3/2})
\]
\[
\left\| \hat{Q}_{11}^{-1} \hat{Q}_{12}(\hat{\Delta}_0^{e_2}) - Q_{11}^{-1} Q_{12}(\Delta_0^{e_2}) \right\|_2 = O_P(n^{-1}m^{-1/2} + n^{-3/2})
\]

Theorem 3.1 indicates

\[
\tilde{\theta}_1 - \theta^* = Q_{11}^{-1} Q_{12}(\Delta_0^{e_2}) - 2^{-1} I_0^{-1} Q(\Delta_0^{e_2}) + R_1^*,
\]

with \( \| R_1^* \|_2 = O_P\{ (m^{-1} + n^{-1})(\| \Delta_0 \|_2 + \| \Delta^e \|_2) + m^{-1}n^{-1} \} \). Also, \( \| \Delta_0 \|_2 = O_P(n^{-1/2}) \) and \( \| \Delta^e \| = O_P\{ (nm)^{-1/2} + n^{-1} \} \). Then,

\[
\left\| \tilde{\theta}_{os} - \theta^* \right\| \leq \| R_1^* \|_2 + \left\| \hat{H}^{-1} \hat{Q}(\hat{\Delta}_0^{e_2}) - I_0^{-1} Q(\Delta_0^{e_2}) \right\|_2 + \left\| \hat{Q}_{11}^{-1} \hat{Q}_{12}(\hat{\Delta}_0^{e_2}) - Q_{11}^{-1} Q_{12}(\Delta_0^{e_2}) \right\|_2
\]
\[
= O_P(n^{-3/2} + m^{-1}n^{-1 / 2}).
\]

The second statements can be verified using the similar logic.

Technical Lemmas

Lemmas of multi-node

Lemma 4.1. Under Assumptions 2-3, consider the mean of the M-estimators from \( m \) centers:

\[
\| \bar{\theta} - \theta_0 \|_2 = O_P\{ (nm)^{-1/2} \} + O(n^{-1}).
\]

Proof. Let \( \bar{d}_i = \hat{\theta}_i - \theta_0 \). By Lemma 4.14 \( \| \mathbb{E}(d_i) \|_2 = O(n^{-1}) \) and \( \mathbb{E}(\| d_i \|_2^2) = O(n^{-1}) \).

\[
\mathbb{E}\{ \| \bar{d} - \mathbb{E}(d_i) \|_2^2 \} = m^{-2} \sum_i \sum_{i'} \mathbb{E}\{ (d_i - \mathbb{E}(d_i))^\top (d_i - \mathbb{E}(d_i)) \} = m^{-1} \mathbb{E}(\| d_i - \mathbb{E}(d_i) \|_2^2) = m^{-1} \mathbb{E}(\| d_i \|_2^2) = O((nm)^{-1})
\]
\[
\| \bar{d} \|_2 = O_P\{ (nm)^{-1/2} \} + O(n^{-1}).
\]

\( \square \)
Lemma 4.2. Under Assumptions \[1, 3\]

\[
\left\| m^{-1} \sum d_i \{ l_i(\theta_0) + I_0 d_i \}^\top \right\|_2 = O_P(m^{-1/2} n^{-3/2} + n^{-2}).
\]

Proof. We have the decomposition:

[Equation here]

Lemma 4.14 indicates \( E(\|d_i - d_{0,i}\|_2^2) = O(n^{-2}). \) Under Assumptions \[1, 3\], Lemma 4.18 states

\[ l_i(\theta_0)(d_i - d_{0,i})^\top = W_i + W_i' \]

with \( E(\|W_i\|_2^2) = O(n^{-3}), \) \( E(\|W_i\|_2) = O(n^{-2}), \) and \( E(\|W_i'\|_2) = O(n^{-2}). \) Also, by Lemma 4.15

\[ E \left( \left\| m^{-1} \sum W_i \right\|_2^2 \right) = O(m^{-1} n^{-3} + n^{-4}). \]

Hence

\[
\left\| m^{-1} \sum d_i \{ l_i(\theta_0) + I_0 d_i \}^\top \right\|_2 \leq \left\| m^{-1} \sum W_i \right\|_2 + \left\| m^{-1} \sum W_i' \right\|_2 +
\left\| m^{-1} \sum (d_i - d_{0,i})(d_i - d_{0,i})^\top I_0 \right\|_2
= O_P(m^{-1/2} n^{-3/2} + n^{-2}).
\]

\qed

Lemma 4.3. Under Assumptions \[1, 3\]

\[
\left\| m^{-1} \sum n \hat{d}_i \hat{d}_i^\top - I_0^{-1} Q_{11} I_0^{-1} \right\|_2 = O_P(n^{-1/2} + m^{-1/2}).
\]

Proof. First, we have \( m^{-1} \sum \hat{d}_i \hat{d}_i^\top = m^{-1} \sum d_i d_i^\top + (\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^\top. \) Let \( d_{0,i} = \)
On the other hand, consider \(-I^{-1}_0 l(\theta_0)\), and Lemma \[4.14\] indicates \(\mathbb{E}(\|d_i - d_{0,i}\|_2^2) = O(n^{-2})\) under Assumption \[3\].

\[
\begin{align*}
&d_i d_i^\top \\
= &(d_{0,i} + d_i - d_{0,i})(d_{0,i} + d_i - d_{0,i})^\top \\
= &d_{0,i} d_{0,i} + d_{0,i}(d_i - d_{0,i})^\top + (d_i - d_{0,i}) d_{0,i}^\top + (d_i - d_{0,i})(d_i - d_{0,i})^\top \\
\mathbb{E}(\|d_i d_i^\top - d_{0,i} d_{0,i}^\top\|_2) \\
\leq &2 \left\{ \mathbb{E}(\|d_i - d_{0,i}\|_2^2) \mathbb{E}(\|d_{0,i}\|_2^2) \right\}^{\frac{1}{2}} + \mathbb{E}(\|d_i - d_{0,i}\|_2^2) = O(n^{-\frac{3}{2}}).
\end{align*}
\]

On the other hand, consider \(d_{0,i} d_{0,i}^\top - n^{-1} I_{0}^{-1} Q_{11} I_{0}^{-1}\) where \(Q_{11} = \mathbb{E} \{ l(\theta_0; X_{ij}) l(\theta_0; X_{ij})^\top \}\). Under Assumption \[3\], Lemma \[4.12\] indicates \(\mathbb{E}(\|l_i(\theta_0)\|_2^4) = O(n^{-2})\), and

\[
\mathbb{E}(\|d_{0,i} d_{0,i}^\top - n^{-1} I_{0}^{-1} Q_{11} I_{0}^{-1}\|_2^2) = O(n^{-2}).
\]

When it comes to \(m\) centers, by Lemma \[4.16\]

\[
\mathbb{E}\left(\left\|m^{-1} \sum d_{0,i} d_{0,i}^\top - n^{-1} I_{0}^{-1} Q_{11} I_{0}^{-1}\right\|_2^2\right) = O(n^{-2} m^{-1}).
\]

Also, Lemma \[4.1\] states \(\|\hat{\theta} - \theta_0\|_2 = O_p\{(nm)^{-1/2}\} + O(n^{-1})\). All together, we have

\[
\begin{align*}
&\left\|m^{-1} \sum \hat{d}_i \hat{d}_i^\top - n^{-1} I_{0}^{-1} Q_{11} I_{0}^{-1}\right\|_2 \\
\leq &\left\|m^{-1} \sum d_i d_i^\top - n^{-1} I_{0}^{-1} Q_{11} I_{0}^{-1}\right\|_2 + \|\hat{\theta} - \theta_0\|_2 \\
\leq &\left\|m^{-1} \sum (d_i d_i^\top - d_{0,i} d_{0,i}^\top)\right\|_2 + \left\|m^{-1} \sum d_{0,i} d_{0,i}^\top - n^{-1} I_{0}^{-1} Q_{11} I_{0}^{-1}\right\|_2 + \|\hat{\theta} - \theta_0\|_2 \\
= &O_p(n^{-3/2}) + O_p(n^{-1} m^{-1/2}).
\end{align*}
\]

Hence,

\[
\left\|m^{-1} \sum n \hat{d}_i \hat{d}_i^\top - I_{0}^{-1} Q_{11} I_{0}^{-1}\right\|_2 = O_p(n^{-1/2} + m^{-1/2}).
\]

\[\square\]

**Lemma 4.4.** Define \(G_i = \hat{g}_i \hat{g}_i^\top\), \(U_i = -\hat{g}_i d_i^\top\), \(G = m^{-1} \sum G_i\), and \(U = m^{-1} \sum U_i\). Under Assumption \[7\] when \(\theta \in U(\rho)\),

\[
G = n^{-1} Q_{11} + R_G,
\]

43
where \( \| R_G \|_2 = O_P(1) \| \Delta \|_2^2 + O_P(n^{-1/2}) \| \Delta \|_2^2 + O_P(n^{-1}) \| \Delta \|_2 + O_P(n^{-1}m^{-1/2}); \)
\[
U = -n^{-1}Q_{11}I_0^{-1} + RU,
\]
where \( \| R_U \|_2 = O_P(n^{-1/2}) \| \Delta \|_2^2 + O_P(n^{-1}) \| \Delta \|_2 + O_P(n^{-1}m^{-1/2} + n^{-3/2}). \)

Specifically, when \( \| \Delta \|_2 = O_P(n^{-1/2}), \)
\[
\| nG - Q_{11} \|_2 = O_P(\| \Delta \|_2) + O_P(m^{-1/2}),
\]
\[
\| nU + Q_{11}I_0^{-1} \|_2 = O_P(n^{-2} + m^{-1/2}).
\]

Proof. We can expand \( \hat{g}_i \) such that \( \hat{g}_i = l_i(\theta_0) + \tilde{H}_i \Delta - \{ \bar{I}(\theta_0) + \tilde{H} \Delta \}. \) For parsimony, let \( l_i = l_i(\theta_0) \) and \( \bar{I} = \bar{I}(\theta_0). \)

\[
\sum \hat{g}_i \hat{g}_i^\top = \sum l_i l_i^\top + \sum l_i \Delta^\top (\tilde{H}_i - \tilde{H}) - \sum \bar{I} \bar{I}^\top + \sum (\tilde{H}_i - \tilde{H}) \Delta^\top (\tilde{H}_i - \tilde{H}) - \sum (\tilde{H}_i - \tilde{H}) \bar{I}^\top \\
- \sum \bar{I}_i \bar{I}_i^\top \\
= \sum l_i l_i^\top - m \bar{I}^\top + \sum l_i \Delta^\top (\tilde{H}_i - \tilde{H}) + \sum (\tilde{H}_i - \tilde{H}) \Delta l_i^\top + \sum (\tilde{H}_i - \tilde{H}) \Delta^\top (\tilde{H}_i - \tilde{H}).
\]

Under Assumptions \([13]\) by Lemma \([4.5\) and \([4.7\] we have
\[
\| m^{-1} \sum l_i(\theta_0) \Delta^\top (\tilde{H}_i - \tilde{H}) \|_2 = O_P(n^{-1/2}) \| \Delta \|_2^2 + O_P(n^{-1}) \| \Delta \|_2 \\
\| m^{-1} \sum (\tilde{H}_i - \tilde{H}) \Delta^\top (\tilde{H}_i - \tilde{H}) \|_2 = O_P(1) \| \Delta \|_2^4 + O_P(n^{-1}) \| \Delta \|_2^2.
\]
Also, \( \| I \|_2^2 = O_P\{nm\} \) and by Lemma 4.16

\[
\mathbb{E}(\| l_i \|_2^2) = O(n^{-2})
\]

\[
\mathbb{E}(l_i l_i^\top) = n^{-1} \sum_j \sum_{j'} \mathbb{E}\{ l(\theta_0; X_{ij}) l(\theta_0; X_{ij'}) \} = n^{-1} Q_{11}
\]

\[
\mathbb{E}\left( \left\| m^{-1} \sum_i l_i l_i^\top - n^{-1} Q_{11} \right\|_2^2 \right) = O(m^{-1}) \mathbb{E}(\| l_i l_i^\top \|_2^2) = O(n^{-2}m^{-1}),
\]

where \( Q_{11} = \mathbb{E}\{ l(\theta_0; X_{11}) l(\theta_0; X_{11})^\top \} \). Therefore,

\[
G = n^{-1} Q_{11} + R_G,
\]

where

\[
R_G = m^{-1} \sum_i l_i l_i^\top - n^{-1} Q_{11} - \Pi^\top + m^{-1} \sum l_i \Delta^\top (\tilde{H}_i - \tilde{H}) + m^{-1} \sum (\tilde{H}_i - \tilde{H}) \Delta l_i^\top + m^{-1} \sum (\tilde{H}_i - \tilde{H}) \Delta \Delta^\top (\tilde{H}_i - \tilde{H})
\]

\[
\| R_G \|_2 = O_P(1) \| \Delta \|_4^4 + O_P(n^{-2}) \| \Delta \|_2^2 + O_P(n^{-1}) \| \Delta \|_2 + O_P(n^{-1}m^{-\frac{1}{2}}).
\]

Similarly, consider \( U = -n^{-1} \sum \hat{g}_i \hat{d}_i \).

\[
-U = m^{-1} \sum g_i (d_i - \bar{d}) = m^{-1} \sum g_i d_i
\]

\[
= m^{-1} \sum l_i d_i^\top + m^{-1} \sum (\tilde{H}_i - \tilde{H}) \Delta d_i^\top - l(\theta_0) \hat{d}^\top
\]

\[
= m^{-1} \sum l_i d_{0,i}^\top + m^{-1} \sum l_i (d_i - d_{0,i})^\top + m^{-1} \sum (\tilde{H}_i - \tilde{H}) \Delta d_i^\top - l(\theta_0) \hat{d}^\top
\]

\[
= -n^{-1} Q_{11} I_0^{-1} + \left( -m^{-1} \sum l_i l_i^\top + n^{-1} Q_{11} \right) I_0^{-1} + m^{-1} \sum l_i (d_i - d_{0,i})^\top + m^{-1} \sum (\tilde{H}_i - \tilde{H}) \Delta d_i^\top - l(\theta_0) \hat{d}^\top.
\]
Recall that by Lemmas 4.1, 4.6, and 4.14
\[ \| \bar{d} \|_2 = O_P \left\{ (nm)^{-1/2} \right\} + O(n^{-1}) \]
\[ \left\| m^{-1} \sum_i (\tilde{H}_i - \tilde{H}) \Delta d_i^\top \right\|_2 = O_P(n^{-1/2}) \| \Delta \|_2^2 + O_P(n^{-1}) \| \Delta \|_2 \]
\[ \mathbb{E} \left\{ \| l_i(\theta_0) (d_i - d_{0,i})^\top \|_2 \right\} \leq \left[ \mathbb{E} \{ \| l_i(\theta_0) \|^2 \} \mathbb{E}(\| d_i - d_{0,i} \|^2) \right]^{1/2} = O(n^{-3/2}). \]

Together with (12) that
\[ \| m^{-1} \sum_i l_i^\top - n^{-1} Q_{11} \|_2 = O_P(n^{-1} m^{-1/2}), \]
\[ U = -n^{-1} Q_{11} I_0^{-1} + R_U, \] (14)

where
\[ R_U = \left( n^{-1} Q_{11} - m^{-1} \sum_i l_i l_i^\top \right) I_0^{-1} + m^{-1} \sum_i l_i (d_i - d_{0,i})^\top + \]
\[ m^{-1} \sum_i (\tilde{H}_i - \tilde{H}) \Delta d_i^\top - \bar{l}(\theta_0) \bar{d}, \]
\[ \| R_U \|_2 = O_P(n^{-1/2}) \| \Delta \|_2^2 + O_P(n^{-1}) \| \Delta \|_2 + O_P(n^{-1} m^{-1/2} + n^{-3/2}). \]

When \( \theta \in U(\rho) \) and \( \| \Delta \| = O_P(n^{-1/2}) \), (13) and (14) can be simplified:
\[ nG - Q_{11} = O_P(1) \| \Delta \|_2 + O_P(m^{-1/2}), \]
\[ nU + Q_{11} I_0^{-1} = O_P(n^{-1/2} + m^{-1/2}). \]

\[ \square \]

**Lemma 4.5.** Consider \( m^{-1} \sum_i l_i(\theta_0) \Delta^\top (\tilde{H}_i - \tilde{H}) \) with \( \theta \in U(\rho) \).

1. If Assumptions 1-3 hold,
\[ \left\| m^{-1} \sum_i l_i(\theta_0) \Delta^\top (\tilde{H}_i - \tilde{H}) \right\|_2 = O_P(n^{-1/2}) \| \Delta \|_2^2 + O_P(n^{-1}) \| \Delta \|_2; \]

2. If Assumptions 2-4 hold,
\[ m^{-1} \sum_i l_i(\theta_0) \Delta^\top (\tilde{H}_i - \tilde{H}) = n^{-1} (I_{d \times d} \otimes \Delta^\top) \times Q_{12} + R_{22}, \]
where $\|R_{22}\|_2 = O_p(n^{-1/2})\|\Delta\|_2^3 + O_p\{n^{-1} + (nm)^{-1/2}\}\|\Delta\|_2^2 + O_p(n^{-1}m^{-1/2})\|\Delta\|_2$.

**Proof.** First,

$$m^{-1} \sum_i l_i(\theta_0) \Delta^\top (\tilde{H}_i - \tilde{H}) = m^{-1} \sum_i l_i(\theta_0) \Delta^\top (\tilde{H}_i - I_0) + \tilde{l}(\theta_0) \Delta^\top (I_0 - \tilde{H}).$$

By Lemmas 4.12 and 4.10 for $k \in [1, 4]$

$$\|\tilde{l}(\theta_0)\|_2^k = O_p\{(nm)^{-k/2}\}$$

$$\|\tilde{H} - I_0\|_2^k \leq 2^{k-1} \left\{ nm^{-1} \sum_j h(X_{ij}) \right\}^k \|\Delta\|_2^k + 2^{k-1} \|H - I_0\|_2^k$$

$$= O_p(1)\|\Delta\|_2^k + O_p\{(nm)^{-k/2}\}$$

$$\|\tilde{l}(\theta_0)\Delta^\top (I_0 - \tilde{H})\|_2^k = O_p\{(nm)^{-k/2}\} \|\Delta\|_2^{2k} + O_p\{(nm)^{-k}\} \|\Delta\|_2^k. \quad (15)$$

**The first statement**

On the other hand, under Assumptions 3 for $k' \in [1, 2]$, at the $i$th center

$$\|l_i(\theta_0) \Delta^\top (\tilde{H}_i - I_0)\|_2^{k'} \leq 2^{k'-1} \|l_i(\theta_0) \Delta^\top (\tilde{H}_i - H_i)\|_2^{k'} + 2^{k'-1} \|l_i(\theta_0) \Delta^\top (H_i - I_0)\|_2^{k'}$$

$$\leq 2^{k'-1} \|l_i(\theta_0)\|_2^{k'} \left\{ n^{-1} \sum_j h(X_{ij}) \right\}^{k'} \|\Delta\|_2^{2k} +$$

$$2^{k'-1} \|l_i(\theta_0)\|_2^{k'} \|H_i - I_0\|_2^{k'} \|\Delta\|_2^{k'}.$$

By Jensen’s inequality and Holder’s inequality,

$$\mathbb{E} \left[ \|l_i(\theta_0)\|_2^{k'} \left\{ n^{-1} \sum_j h(X_{ij}) \right\}^{k'} \right] \leq \left[ \mathbb{E} \left\{ \|l_i(\theta_0)\|_2^{2k'} \right\} \mathbb{E} \left\{ h(X_{ij})^{2k} \right\} \right]^{1/2} = O(n^{-k/2})$$

$$\mathbb{E}(\|l_i(\theta_0)\|_2^{k'} \|H_i - I_0\|_2^{k'}) \leq \left\{ \mathbb{E}(\|l_i(\theta_0)\|_2^{2k}) \mathbb{E}(\|H_i - I_0\|_2^{2k}) \right\}^{1/2} = O(n^{-k'}).$$

47
Let \( k' = 1, \)
\[
\|\bar{I}(\theta_0)\Delta^\top(I_0 - \bar{H})\|_2 = O_p\{(nm)^{-1/2}\}\|\Delta\|_2^2 + O_p\{(nm)^{-1}\}\|\Delta\|_2
\]  
(17)
\[
\left\| m^{-1} \sum_i l_i(\theta_0) \Delta^\top(\tilde{H}_i - I_0) \right\|_2 \leq m^{-1} \sum_i \|l_i(\theta_0)\|_2 \left\{ n^{-1} \sum_j h(X_{ij}) \right\} \|\Delta\|_2^2 + m^{-1} \sum_i \|l_i(\theta_0)\|_2 \|\tilde{H}_i - I_0\|_2 \|\Delta\|_2 \\
= O_p(n^{-1/2})\|\Delta\|_2^2 + O_p(n^{-1})\|\Delta\|_2.
\]  
(18)

Hence,
\[
\left\| m^{-1} \sum_i l_i(\theta_0) \Delta^\top(\tilde{H}_i - \bar{H}) \right\|_2 \leq \left\| m^{-1} \sum_i l_i(\theta_0) \Delta^\top(\tilde{H}_i - I_0) \right\|_2 + \|\bar{I}(\theta_0)\Delta^\top(I_0 - \bar{H})\|_2 \\
= O_p(n^{-1/2})\|\Delta\|_2^2 + O_p(n^{-1})\|\Delta\|_2.
\]  
(19)

The second statement

Note that the major contribution in this term comes from \( m^{-1} \sum_i l_i(\theta_0)\Delta^\top(\tilde{H}_i - I_0). \) When Assumption 1 holds, by Lemma 4.10 for \( k' \in [1, 2], \)
\[
(\tilde{H}_i - H_i)\Delta = 2^{-1}Q(\Delta^{\otimes 2}) + R_{1i} + R_{2i},
\]
where \( \|R_{1i}\|_2 \leq (6n)^{-1} \sum_j m(X_{ij})\|\Delta\|_2^3 \) and \( \|R_{2i}\|_2 \leq R_{2i}\|\Delta\|_2 \) with \( \mathbb{E}(R^k_{2i}) = O(n^{-k'/2}). \) Then,
\[
m^{-1} \sum_i l_i(\theta_0)\Delta^\top(\tilde{H}_i - H_i) = \bar{I}(\theta_0)[2^{-1}Q(\Delta^{\otimes 2})]^\top + m^{-1} \sum_i l_i(R_{1i} + R_{2i})^\top \\
\|l_i(R_{1i} + R_{2i})^\top\|_2^{k'} \leq 2^{k'-1}\|l_i\|_2^{k'} \left\{ (6n)^{-1} \sum_j m(X_{ij}) \right\}^{k'} \|\Delta\|_2^{3k'} + 2^{k'-1}\|l_i\|_2^{k'} R_{2i}\|\Delta\|_2^{2k'}.
\]

Additionally, by definition
\[
\|Q(\Delta^{\otimes 2})\|_2 \leq \lambda_3\|\Delta\|_2^2.
\]

48
Let $k' = 1$, then $m^{-1} \sum \| I_i (R_{1i} + R_{2i})^\top \|_2^2 = O_P(n^{-1/2}) \| \Delta \|_2^3 + O_P(n^{-1}) \| \Delta \|_2^2$.

$$m^{-1} \sum \| I_i (\theta_0) \Delta^\top (\tilde{H}_i - H_i) \|_2^2 = O_P(n^{-1/2}) \| \Delta \|_2^3 + O_P\{n^{-1} + (nm)^{-1/2}\} \| \Delta \|_2^2. \quad (20)$$

On the other hand, consider $m^{-1} \sum I_i (\theta_0) \Delta^\top (H_i - I_0)$. By Lemma 4.19 with the existence of the fourth moments of $l_i (\theta_0; X_{ij})$ and $H_i (\theta_0; X_{ij})$,

$$m^{-1} \sum I_i \Delta^\top (H_i - I_0) = (I_{dxd} \otimes \Delta^\top) \times \{ I_i \otimes (H_i - I_0) \} = (I_{dxd} \otimes \Delta^\top) \times (n^{-1} Q_{12} + R_{23}) ,$$

where $\mathbb{E}(\| R_{23} \|_2^2) = O(m^{-1}n^{-2})$.

$$\left\| n^{-1} (I_{dxd} \otimes \Delta^\top) \times R_{23} \right\| \leq \| I_{dxd} \otimes \Delta^\top \|_2 \| R_{23} \|_2 = \| \Delta \|_2 O_P(m^{-1/2}n^{-1}) \quad (21)$$

Combining (17), (20), and (21) leads to

$$m^{-1} \sum I_i (\theta_0) \Delta^\top (\tilde{H}_i - \tilde{H}) = m^{-1} \sum I_i (\theta_0) \Delta^\top (\tilde{H}_i - H_i) + m^{-1} \sum I_i (\theta_0) \Delta^\top (H_i - I_0) +$$

$$\tilde{I}(\theta_0) \Delta^\top (I_0 - \tilde{H}) = n^{-1} (I_{dxd} \otimes \Delta^\top) \times Q_{12} + R_{22},$$

where $\| R_{22} \|_2 = O_P(n^{-1/2}) \| \Delta \|_2^3 + O_P\{n^{-1} + (nm)^{-1/2}\} \| \Delta \|_2^2 + O_P(n^{-1}m^{-1/2}) \| \Delta \|_2$.

Lemma 4.6. Consider $m^{-1} \sum (\tilde{H}_i - \tilde{H}) \Delta d_i^\top$ with $\theta \in U(\rho)$.

1. If Assumptions 1-4 hold,

$$\left\| m^{-1} \sum (\tilde{H}_i - \tilde{H}) \Delta d_i^\top \right\|_2 \leq O_P(n^{-1/2}) \| \Delta \|_2^2 + O_P(n^{-1}) \| \Delta \|_2.$$

2. If Assumptions 2-4 hold,

$$m^{-1} \sum (\tilde{H}_i - \tilde{H}) \Delta d_i^\top = -n^{-1} Q_{12}^\top \times (I_{dxd} \otimes \Delta) \times I_0^{-1} - R_{22}^\top I_0^{-1} + R_{42},$$

where $\| R_{22} \|_2 = O_P(n^{-1/2}) \| \Delta \|_2^3 + O_P\{n^{-1} + (nm)^{-1/2}\} \| \Delta \|_2^2 + O_P(n^{-1}m^{-1/2}) \| \Delta \|_2$,

and $\| R_{42} \|_2 = O_P\{n^{-1} + (nm)^{-1/2}\} \| \Delta \|_2^2 + O_P\{n^{-3/2} + (nm)^{-1}\} \| \Delta \|_2$. 49
Proof. First

\[ m^{-1} \sum_i (\mathbf{H}_i - \tilde{\mathbf{H}}) \Delta \mathbf{d}_i^T = m^{-1} \sum_i (\mathbf{H}_i - \mathbf{I}_0) \Delta \mathbf{d}_i^T + (\mathbf{I}_0 - \tilde{\mathbf{H}}) \Delta \mathbf{d}^T. \]

Consider \((\mathbf{I}_0 - \tilde{\mathbf{H}}) \Delta \mathbf{d}^T\). By (15), \(\| \mathbf{H} - \mathbf{I}_0 \|_2^k = O_P(1)\| \Delta \|_2^k + O_P\{(nm)^{-k/2}\}\). By Lemma 4.14 and 4.1, \(\|\mathbf{E}(\mathbf{d}_i)\|_2 = O(n^{-1}), \mathbb{E}(\|\mathbf{d}_i\|_2^k) = O(n^{-k/2}),\) and \(\|\mathbf{d}\|_2 = O_P\{(nm)^{-1/2}\} + O(n^{-1})\). Hence,

\[ \| (\mathbf{I}_0 - \tilde{\mathbf{H}}) \Delta \mathbf{d}^T \|_2 = O_P\{(nm)^{-1/2} + n^{-1}\} \| \Delta \|_2^2 + O_P\{n^{-3/2}m^{-1/2} + (nm)^{-1}\} \| \Delta \|_2. \] (22)

The first statement

\[ \| (\tilde{\mathbf{H}}_i - \mathbf{I}_0) \Delta \mathbf{d}_i^T \|_2^k \leq 2^{k-1}\| (\tilde{\mathbf{H}}_i - \mathbf{H}_i) \Delta \mathbf{d}_i^T \|_2^k + 2^{k-1}\| (\mathbf{H}_i - \mathbf{I}_0) \Delta \mathbf{d}_i^T \|_2^k \]

\[ \leq 2^{k-1}\|\mathbf{d}_i\|_2^k \left\{ n^{-1} \sum_j h(X_{ij}) \right\}^k \| \Delta \|_2^{2k} + 2^{k-1}\|\mathbf{d}_i\|_2^k \| \mathbf{H}_i - \mathbf{I}_0 \|_2^k \| \Delta \|_2^k. \]

For \(k \in [1, 2]\), using Holder’s inequality and Jensen’s inequality, at the \(i\)th node

\[ \mathbb{E} \left[\|\mathbf{d}_i\|_2^k \left\{ n^{-1} \sum_j h(X_{ij}) \right\}^k \right] \leq \left[ \mathbb{E}(\|\mathbf{d}_i\|_2^{2k}) \mathbb{E}\left\{ h(X_{ij})^{2k} \right\} \right]^{1/2} = O(n^{-k/2}) \]

\[ \mathbb{E}(\|\mathbf{d}_i\|_2^k \| \mathbf{H}_i - \mathbf{I}_0 \|_2^k) \leq \left[ \mathbb{E}(\|\mathbf{d}_i\|_2^{2k}) \mathbb{E}(\| \mathbf{H}_i - \mathbf{I}_0 \|_2^{2k}) \right]^{1/2} = O(n^{-k}). \]
Let $k = 1$. Then,
\[
\left\| m^{-1} \sum_i (\tilde{H}_i - I_0) \Delta d_i^T \right\|_2 \leq m^{-1} \sum_i \|d_i\|_2 \left\{ n^{-1} \sum_j h(X_{ij}) \right\} \|\Delta\|_2 + m^{-1} \sum_i \|d_i\|_2 \|H_i - I_0\|_2 \|\Delta\|_2
\]
\[
= O_p(\sqrt[n-1]{2}) \|\Delta\|_2^2 + O_p(n^{-1}) \|\Delta\|_2
\]
\[
\left\| m^{-1} \sum_i (\tilde{H}_i - \bar{H}) \Delta d_i^T \right\|_2 \leq \left\| m^{-1} \sum_i (\tilde{H}_i - I_0) \Delta d_i^T \right\|_2 + \|I_0 - \bar{H}\| \Delta d^T \|_2
\]
\[
= O_p(n^{-1/2}) \|\Delta\|_2^2 + O_p(n^{-1}) \|\Delta\|_2. \tag{23}
\]

The second statement
Define $d_{0,i} = -I_0^{-1} l_i(\theta_0)$, and
\[
m^{-1} \sum_i (\tilde{H}_i - \bar{H}) \Delta d_i^T = m^{-1} \sum_i (\tilde{H}_i - \bar{H}) \Delta d_{0,i}^T + m^{-1} \sum_i (\bar{H} - I_0) \Delta d_{0,i}^T
\]
\[
+ m^{-1} \sum_i (\bar{H} - I_0) \Delta (d_i - d_{0,i})^T + (I_0 - \bar{H}) \Delta \bar{d}_i^T.
\]

Consider $m^{-1} \sum_i (\tilde{H}_i - \bar{H}) \Delta d_{0,i}^T$. By the second part of Lemma 4.5
\[
m^{-1} \sum_i (\tilde{H}_i - \bar{H}) \Delta d_{0,i}^T = -m^{-1} \sum_i (\tilde{H}_i - \bar{H}) \Delta l_i(\theta_0)^T I_0^{-1}
\]
\[
= -n^{-1} Q_{12}^T \times (I_{dxd} \otimes \Delta) \times I_0^{-1} - R_{22}^T I_0^{-1}, \tag{24}
\]
where $\|R_{22}\|_2 = O_p(n^{-1/2}) \|\Delta\|_2^3 + O_p\{n^{-1} + (nm)^{-1/2}\} \|\Delta\|_2^2 + O_p(n^{-1}m^{-1/2}) \|\Delta\|_2.

Consider $m^{-1} \sum_i (\bar{H} - I_0) \Delta d_{0,i}^T$. Note that by definition, $m^{-1} \sum_i (\bar{H} - I_0) \Delta d_{0,i}^T = -(\bar{H} - I_0) \Delta l(\theta_0)^T I_0^{-1}$. By (15),
\[
\left\| (\bar{H} - I_0) \Delta l(\theta_0)^T I_0^{-1} \right\|_2 \leq \lambda^{-1} \left\| \bar{H} - I_0 \right\|_2 \|l(\theta_0)\|_2 \|\Delta\|_2
\]
\[
= O_p\{nm^{-1/2}\} \|\Delta\|_2^2 + O_p\{nm^{-1}\} \|\Delta\|_2. \tag{25}
\]
Consider $m^{-1} \sum (\tilde{H}_i - I_0) \Delta (d_i - d_{0,i})^\top$.

$$d_i - d_{0,i} = \left( -H_i^{-1} + I_0^{-1} \right) I_i(\theta_0) = I_0^{-1} (H_i - I_0) H_i^{-1} I_i(\theta_0) = I_0^{-1}(I_0 - H_i) d_i$$

$$(\tilde{H}_i - I_0) \Delta (d_i - d_{0,i})^\top = \left( \tilde{H}_i - H_i + H_i - I_0 \right) \Delta d_i^\top (I_0 - H_i) I_0^{-1}$$

$$= (\tilde{H}_i - H_i) \Delta d_i^\top (I_0 - H_i) I_0^{-1} + (\tilde{H}_i - H_i) \Delta d_i^\top (H_i - \overline{H}_i) I_0^{-1}$$

$$(H_i - I_0) \Delta d_i^\top (I_0 - H_i) I_0^{-1} + (H_i - I_0) \Delta d_i^\top (H_i - \overline{H}_i) I_0^{-1}.$$  

Consider each component, under Assumption 3.

$$\left\| (\tilde{H}_i - H_i) \Delta d_i^\top (I_0 - H_i) I_0^{-1} \right\|_2 \leq \left\{ n^{-1} \sum_j h(X_{ij}) \right\} \| d_i \|_2 \| H_i - I_0 \|_2 \lambda^{-1} \| \Delta \|_2$$

$$\left\| (\tilde{H}_i - H_i) \Delta d_i^\top (H_i - \overline{H}_i) I_0^{-1} \right\|_2 \leq \left\{ n^{-1} \sum_j h(X_{ij}) \right\} \| d_i \|_2 \| H_i - H_i \|_2 \lambda^{-1} \| \Delta \|_2$$

$$\left\| (H_i - I_0) \Delta d_i^\top (I_0 - H_i) I_0^{-1} \right\|_2 \leq \| H_i - I_0 \|_2 \| d_i \|_2 \lambda^{-1} \| \Delta \|_2$$

$$\left\| (H_i - I_0) \Delta d_i^\top (H_i - \overline{H}_i) I_0^{-1} \right\|_2 \leq \| H_i - I_0 \|_2 \| d_i \|_2 \lambda^{-1} \| \Delta \|_2.$$
By Lemma 4.12 and Lemma 4.13,

\[
\mathbb{E} \left[ \left\{ n^{-1} \sum_j h(x_{ij}) \right\} \|d_i\|_2 \|H_i - I_0\|_2 \right]
\leq \left( \mathbb{E} \left[ \left\{ n^{-1} \sum_j h(x_{ij}) \right\}^2 \right] \right)^{1/2} \left\{ \mathbb{E} \|d_i\|_4^2 \mathbb{E} \|H_i - I_0\|_4^2 \right\}^{1/4} = O(n^{-1})
\]

\[
\mathbb{E} \left[ \left\{ n^{-1} \sum_j h(x_{ij}) \right\} \|d_i^T(\bar{H}_i - H_i)\|_2 \right]
\leq \left( \mathbb{E} \left[ \left\{ n^{-1} \sum_j h(x_{ij}) \right\}^2 \right] \mathbb{E} \left\{ \|d_i^T(\bar{H}_i - H_i)\|_2^2 \right\} \right)^{1/2} = O(n^{-1})
\]

\[
\mathbb{E} \left\{ \|H_i - I_0\|_2^2 \|d_i\|_2 \right\}
\leq \left\{ \mathbb{E} \|d_i\|_2^2 \mathbb{E} \|H_i - I_0\|_2^2 \right\}^{1/2} = O(n^{-3/2})
\]

\[
\mathbb{E} \left\{ \|H_i - I_0\|_2 \|d_i^T(\bar{H}_i - H_i)\|_2 \right\}
\leq \left[ \mathbb{E} \|H_i - I_0\|_2 \mathbb{E} \left\{ \|d_i^T(\bar{H}_i - H_i)\|_2^2 \right\} \right]^{1/2} = O(n^{-3/2}).
\]

Therefore,

\[
m^{-1} \sum_i (\bar{H}_i - I_0) \Delta (d_i - d_{0,i})^T = O_p(n^{-1}) \|\Delta\|_2^2 + O_p(n^{-3/2}) \|\Delta\|_2.
\] (26)

Combining (22), (24), (25), and (26)

\[
m^{-1} \sum_i (\bar{H}_i - \bar{H}) \Delta d_i^T = -n^{-1}Q_{12}^T \times (I_{d \times d} \otimes \Delta) \times I_0^{-1} - R_{22} I_0^{-1} + R_{42},
\] (27)

where

\[
R_{42} = -(\bar{H} - I_0) \Delta \bar{l}(\theta_0)^T I_0^{-1} + m^{-1} \sum_i (\bar{H}_i - I_0) \Delta (d_i - d_{0,i})^T + (I_0 - \bar{H}) \Delta \bar{d}^T
\]

\[
\|R_{42}\|_2 = O_p\{n^{-1} + (nm)^{-1/2}\} \|\Delta\|_2^2 + O_p\{n^{-3/2} + (nm)^{-1}\} \|\Delta\|_2.
\]
Lemma 4.7. Consider $m^{-1} \sum_i (\tilde{H}_i - \bar{H}) \Delta \Delta^\top (\tilde{H}_i - \bar{H})$ with $\theta \in U(\rho)$. If Assumptions 1-3 hold,

$$
\left\| m^{-1} \sum_i (\tilde{H}_i - \bar{H}) \Delta \Delta^\top (\tilde{H}_i - \bar{H}) \right\|_2 = O_P(1) \| \Delta \|_2^4 + o_P(n^{-1}) \| \Delta \|_2^2.
$$

Proof. By definition $\sum_i \tilde{H}_i = m\bar{H}$.

$$
\sum_i (\tilde{H}_i - \bar{H}) \Delta \Delta^\top (\tilde{H}_i - \bar{H}) = \sum_i \tilde{H}_i \Delta \Delta^\top (\tilde{H}_i - \bar{H}) = \sum_i (\tilde{H}_i - I_0) \Delta \Delta^\top (\tilde{H}_i - \bar{H})
$$

$$
= \sum_i (\tilde{H}_i - I_0) \Delta \Delta^\top (\tilde{H}_i - I_0 + I_0 - \bar{H})
$$

$$
= \sum_i (\tilde{H}_i - I_0) \Delta \Delta^\top (\tilde{H}_i - I_0) + m(\bar{H} - I_0) \Delta \Delta^\top (I_0 - \bar{H})
$$

$$
\| (\tilde{H}_i - I_0) \Delta \|_2 \leq \| \Delta \|_2 \| \tilde{H}_i - H + H_0 - I_0 \|_2
$$

$$
\leq 2^{k-1} \| \Delta \|_2^2 \left\{ n^{-1} \sum_j h(X_{ij}) \right\}^k + 2^{k-1} \| \Delta \|_2^4 \| H_0 - I_0 \|_2^2.
$$

When $k = 2$,

$$
\left\| m^{-1} \sum_i (\tilde{H}_i - \bar{H}) \Delta \Delta^\top (\tilde{H}_i - \bar{H}) \right\|_2 \leq m^{-1} \sum \| (\tilde{H}_i - I_0) \Delta \|_2^2 + \| (\bar{H} - I_0) \Delta \|_2^2
$$

$$
= 2 \| \Delta \|_2^4 \left[ m^{-1} \sum \left\{ n^{-1} \sum_j h(X_{ij}) \right\}^2 + \left\{ (nm)^{-1} \sum \sum h(X_{ij}) \right\}^2 \right] +
$$

$$
2 \| \Delta \|_2^2 \left[ m^{-1} \sum_i \| H_i - I_0 \|_2^2 + \| H - I_0 \|_2^2 \right]
$$

$$
= O_P(1) \| \Delta \|_2^4 + O_P(n^{-1}) \| \Delta \|_2^2.
$$

}$
Lemma 4.8. Under Assumption 1-4, for $\theta \in U(\rho)$,

$$(I_0 - \overline{H}^{(g)})\Delta = -\frac{1}{2}Q(\Delta \otimes^2) + R_H,$$

where $\|R_H\| = O_P(\|\Delta\|_2^3) + O_P(n^{-1/2}\|\Delta\|_2^2) + O_P(n^{-1/2}m^{-1/2})\|\Delta\|_2$.

Proof. Note that

$$(I_0 - \overline{H}^{(g)})\Delta = (I_0 - H^{(g)}_0)\Delta + (H^{(g)}_0 - \overline{H}^{(g)})\Delta.$$  

The first component, by Lemma 4.12, $\|H^{(g)}_0 - \overline{H}^{(g)}\|_2 = O_P(n^{-1/2}m^{-1/2})\|\Delta\|_2$. Lemma 4.10 indicates $(H^{(g)}_0 - \overline{H}^{(g)})\Delta = 2^{-1}Q(\Delta \otimes^2) + R_{H1}$, where $\|R_{H1}\| = O_P(n^{-1/2}\|\Delta\|_2^3) + O_P(\|\Delta\|_2^2)$. Therefore,

$$(I_0 - \overline{H}^{(g)})\Delta = -\frac{1}{2}Q(\Delta \otimes^2) + R_H,$$

where $\|R_H\| = O_P(\|\Delta\|_2^3) + O_P(n^{-1/2}\|\Delta\|_2^2) + O_P(n^{-1/2}m^{-1/2})\|\Delta\|_2$.  

Lemmas of single-node

In this section, we present lemmas that apply to single-node or non-distributed data. The index of nodes is therefore dropped for parsimony. First, define the M-estimator

$$\hat{\theta} = \arg\min_{\theta \in \Theta} n^{-1} \sum L(\theta; X_i),$$

and $d = \hat{\theta} - \theta_0$. Additionally,

$$l_i(\theta) = \nabla L(\theta; X_i), \quad H_i(\theta) = \nabla^2 L(\theta; X_i)$$

$$l_i = \nabla L(\theta_0; X_i), \quad H_i = \nabla^2 L(\theta_0; X_i), \quad I_0 = \mathbb{E}(H_i)$$

$$\overline{H}_i = \int_0^1 \nabla^2 L(\theta_0 + t(\hat{\theta} - \theta_0); X_i) \, dt$$

$$\bar{l} = n^{-1} \sum \nabla L(\theta_0; X_i), \quad \bar{H} = n^{-1} \sum \nabla^2 L(\theta_0; X_i), \quad \overline{H} = n^{-1} \sum \overline{H}_i.$$  

55
Lemma 4.9. Define
\[ G_j = \nabla^2 \left\{ \frac{\partial}{\partial \theta_j} L(\theta; X) \right\}, j = 1, \ldots, d. \]

Under Assumption 1-3, when \( G_j(\theta; X) \) is continuous with respect to \( \theta \), for \( j = 1, \ldots, d \)
\[ \max_j \| G_j(\theta; X) \|_2 \leq h(X), \forall \theta \in \mathcal{U}(\rho). \]

Proof. Recall the definition
\[ \nabla^3 L(\theta; X_i)(u) = \begin{cases} u^\top G_1(\theta; X_i) \\ \vdots \\ u^\top G_d(\theta; X_i) \end{cases}. \]

Assumption 3 indicates that \( \forall \theta', \theta \in \mathcal{U}(\rho), \| \nabla^2 L(\theta'; X_i) - \nabla^2 L(\theta; X_i) \|_2 \leq h(X_i) \| d \|_2 \), where \( d = \theta' - \theta \). By Assumption 3
\[ \| \nabla^2 L(\theta'; X_i) - \nabla^2 L(\theta; X_i) \|_2 \leq h(X_i) \| d \|_2 \]
\[ \left\| d^\top \int_0^1 G_j(\theta + t d; X_i) \, dt \right\|_2 \leq h(X_i) \| d \|_2. \] (28)

Assumption 4 states that \( \| G_j(\theta; X_i) \|_2 \) is continuous with respect to \( \theta \). Consequently, \( \| G_j(\theta; X_i) \|_2 \leq h(X_i) \) for \( j = 1, \ldots, d \), almost everywhere.

To see this, suppose there exists \( \theta \in \mathcal{U}(\rho) \) and \( \| G_j(\theta; X) \|_2 - h(X) > 2\delta > 0 \). By continuity, there exists a small ball around \( \theta \) such that
\[ U(\delta) = \left\{ \theta' \in \mathcal{U}(\rho); \| G_j(\theta'; X) - G_j(\theta; X) \|_2 \leq \delta \right\}. \]

Since \( d = \theta' - \theta \) can be in any direction, we can choose \( \theta' \in U(\delta) \) such that
\[ \| d \|_2^{-1} \left\| d^\top G_j(\theta; X) \right\|_2 = \| G_j(\theta; X) \|_2. \]
Then by the triangular inequality
\[ \|d\|_2^{-1} \left\| d^\top \int_0^1 G_j(\theta + td; X) \, dt \right\|_2 \]
\[ \geq \|d\|_2^{-1} \left\| d^\top G_j(\theta; X) \right\|_2 - \|d\|_2^{-1} \left\| d^\top \int_0^1 \{G_j(\theta; X) - G_j(\theta + td; X)\} \, dt \right\|_2 \]
\[ \geq \{h(X) + 2\delta\} - \delta > h(X), \]
which contradicts (28).

Lemma 4.10. Let \( H = n^{-1} \sum_i \int_0^1 \nabla^2 L(\theta_0 + t(\theta - \theta_0); X_i) \, dt \). Under Assumption 1-4, for all \( \theta \in U(\rho) \) and \( \Delta = \theta - \theta_0 \),
\[ \|H - H\|_2 \leq \left\{ n^{-1} \sum_i h(X_i) \right\} \|\Delta\|, \]
\[ (H - H)\Delta = (2n)^{-1} \sum_i \nabla^3 L(\theta_0; X_i)(\Delta^{\otimes 2}) + R_1, \]
\[ (H - H)\Delta = 2^{-1} Q(\Delta^{\otimes 2}) + R_1 + R_2, \]
where \( \|R_1\|_2 \leq (6n)^{-1} \sum m(X_i)\|\Delta\|_2^2 \), and \( \|R_2\|_2 \leq R_2\|\Delta\|_2^2 \) with \( \mathbb{E}(R_2^k) = O(n^{-k/2}) \) and
\[ k = \max \left\{ j \geq 2; \mathbb{E}\{h(X_1)^j\} < \infty \right\}. \]

Proof. The first equation is a direct result from Assumption 3. For parsimony, we drop the center index \( i \) in the proof and consider any \( \theta \in U(\rho) \). Let \( \Delta = \theta - \theta_0 \). Under Assumption 1, recall that \( H = n^{-1} \sum_i \int_0^1 \nabla^2 L(\theta_0 + t(\theta - \theta_0); X_i) \, dt \), then
\[ l(\theta) = l(\theta_0) + H\Delta = l(\theta_0) + H\Delta + n^{-1} \sum_i \int_0^1 (1 - t)\nabla^3 L(\theta_0 + t(\theta - \theta_0); X_i)(\Delta^{\otimes 2}) \, dt \]
\[ = l(\theta_0) + H\Delta + (2n)^{-1} \sum_i \nabla^3 L(\theta_0; X_i)(\Delta^{\otimes 2}) + R_1, \]
where

\[
R_1 = n^{-1} \sum_i \int_0^1 (1 - t) \left[ \nabla^3 L\{\theta_0 + t(\theta - \theta_0); X_i\} - \nabla^3 L(\theta_0; X_i) \right] (\Delta \otimes^2) \, dt
\]

\[
\|R_1\|_2 \leq (6n)^{-1} \sum_i m(X_i) \|\Delta\|_2^3.
\]

Further decomposition gives

\[
n^{-1} \sum_i \nabla^3 L(\theta_0; X_i)(\Delta \otimes^2) = Q(\Delta \otimes^2) + R_2
\]

\[
R_2 = \left[ n^{-1} \sum_i \nabla^3 L(\theta_0; X_i) - Q \right] (\Delta \otimes^2)
\]

\[
= \left( \Delta^\top \left[ n^{-1} \sum G_1(\theta_0; X_i) - \mathbb{E}\{G_1(\theta_0; X_i)\} \right] \Delta \right)
\]

\[
\quad \quad \quad \cdots
\]

\[
= \left( \Delta^\top \left[ n^{-1} \sum G_d(\theta_0; X_i) - \mathbb{E}\{G_d(\theta_0; X_i)\} \right] \Delta \right).
\]

For \( j = 1, \ldots, d \), consider \( n^{-1} \sum G_j(\theta_0; X_i) \). By definition, \( G_j(\theta_0; X_i) \) is Hermitian, and Lemma 4.9 indicates \( \|G_j(\theta_0; X_i)\|_2 \leq h(X_i) \). Let \( Z_{j,i} = G_j(\theta_0; X_i) - \mathbb{E}\{G_j(\theta_0; X_i)\} \), and Lemma 4.16 indicates

\[
\mathbb{E} \left( \left\| n^{-1} \sum_i Z_{j,i} \right\|^k \right) = O(n^{-k/2}),
\]

for some \( k \geq 2 \) and \( \mathbb{E}(\|Z_{j,i}\|^k) \) exists. In our case, we have

\[
\mathbb{E} \left( \|Z_{j,i}\|^k \right) \leq 2^{k-1} \mathbb{E}\{h(X_i)^k\} + 2^{k-1} \lambda_h^k.
\]

Hence, the maximum of \( k \) depends on moments of \( h(X_i) \). By definition,

\[
\|R_2\|_2 = \left\| \left( \begin{array}{c}
\Delta^\top n^{-1} \sum_i Z_{1,i} \Delta \\
\vdots \\
\Delta^\top n^{-1} \sum_i Z_{d,i} \Delta
\end{array} \right) \right\|_2 \leq \|\Delta\|_2^2 \left( \sum_j \left\| n^{-1} \sum_i Z_{j,i} \right\|_2^2 \right)^{1/2}
\]

\[
\mathbb{E} \left\{ \left( \sum_j \left\| n^{-1} \sum_i Z_{j,i} \right\|_2^2 \right)^{k/2} \right\} \leq d^{k/2-1} \sum_j \mathbb{E} \left( \left\| n^{-1} \sum_i Z_{j,i} \right\|_2^k \right) = O(n^{-k/2}).
\]
Let $R_1 = \left( \sum_j n^{-1} \sum_i |Z_{j,i}|^2 \right)^{1/2}$, and we complete the proof.

Lemma 4.11. Let $\delta_\rho = \min \{ \rho, \rho \lambda_-/(4 \lambda_h) \}$, and define the following four events:

\[
\mathcal{E}_1 = \left\{ n^{-1} \sum h(X_i) \leq 2 \lambda_h \right\}
\]
\[
\mathcal{E}_2 = \left\{ \| H - I_0 \|_2 \leq \frac{\rho \lambda_2}{2} \right\}
\]
\[
\mathcal{E}_3 = \left\{ \| l(\theta_0) \|_2 \leq \frac{(1 - \rho) \lambda_- \delta_\rho}{2} \right\}
\]
\[
\mathcal{E}_4 = \left\{ n^{-1} \sum m(X_i) \leq 2 \lambda_m \right\}.
\]

Let $\mathcal{E}_0 = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ and $\tilde{\mathcal{E}}_0 = \mathcal{E}_0 \cap \mathcal{E}_4$.

(1) Suppose for some $k \geq 2$

\[
\mathbb{E}\{\| \nabla L(\theta_0; X_i) \|_k^k \} \leq G^k \text{ and } \mathbb{E}\{\| \nabla^2 L(\theta_0; X_i) - \mathbb{E}\{\| \nabla^2 L(\theta_0; X_i) \| \} \|_2 \} \leq H^k
\]

\[
E\{h(X_i)^k\} \leq \lambda_h^k \text{ and } E\left[|h(X_i) - \mathbb{E}\{h(X_i)\}|^k\right] \leq \lambda_h^k.
\]

Then,

\[
\mathbb{P}(\mathcal{E}_0^c) = O(n^{-k/2}).
\]

Specifically, under $\mathcal{E}_0$,

\[
\|d\|_2 \leq \frac{2 \|l(\theta_0)\|_2}{(1 - \rho) \lambda_-}, \text{ and } H(\theta') \succeq (1 - \rho) \lambda_- I, \forall \theta' \in U(\rho),
\]

and $\|H\|_2 \geq (1 - \rho) \lambda_-$. 

(2) With the additional assumption $\max(E\{m(X_i)^k\}, E\left[|m(X_i) - \mathbb{E}\{m(X_i)\}|^k\right]) \leq \lambda_m^k$,

\[
\mathbb{P}(\tilde{\mathcal{E}}_0^c) = O(n^{-k/2}).
\]

Proof. By Lemma 6 and Lemma 7 in [Zhang et al. (2013)], $\mathbb{P}(\mathcal{E}_0^c) = O(n^{-k/2})$. The probability
of $\mathcal{E}_4$ can be derived in a similar way. By Lemma 4.16,

$$P(\mathcal{E}_4) = \mathbb{E}\left[\mathbb{1}\left\{n^{-1}\sum m(X_i) > 2\lambda_m\right\}\right]$$

$$\leq \mathbb{E}\left[\mathbb{1}\left\{n^{-1}\sum m(X_i) - \mathbb{E}\{m(X_i)\} > \lambda_m\right\}\right]$$

$$\leq \lambda_m^{-k}\mathbb{E}\left[n^{-1}\sum m(X_i) - \mathbb{E}\{m(X_i)\}\right]$$

$$= O(n^{-k/2}).$$

Hence, $P(\mathcal{E}_6) \leq P(\mathcal{E}_5) + P(\mathcal{E}_4) = O(n^{-2}).$

\[
E(\|\bar{l}(\theta_0)\|_2^k) \leq O(n^{-k/2})
\]

$$E(\|H - I_0\|_2^k) \leq O(n^{-k/2})$$

$$E(\|d\|_2^k) = O(n^{-k/2}).$$

See Theorem 1 and Lemmas 7-9 given in [Zhang et al. 2013].

\textbf{Lemma 4.13.} Under the conditions of Lemma 4.11 (1),

$$E\left\{\|(H - \bar{H})d\|_2^{k/2}\right\} = O(n^{-k/2}).$$
Proof.

\[
\| (\mathbf{H} - \overline{\mathbf{H}}) \mathbf{d} \|_2 \leq 1(\mathcal{E}_0)\| (\mathbf{H} - \overline{\mathbf{H}}) \mathbf{d} \|_2 + 1(\mathcal{E}_0^c)\| (\mathbf{H} - \mathbf{I}_0) \mathbf{d} \|_2 + 1(\mathcal{E}_0^c)\| \mathbf{I}_0 \mathbf{d} \|_2 + 1(\mathcal{E}_0^c)\| \mathbf{I} \|_2
\]

\[
\mathbb{E} \left\{ 1(\mathcal{E}_0)\| (\mathbf{H} - \overline{\mathbf{H}}) \mathbf{d} \|_2^{k/2} \right\} \leq (2\lambda_0)^{k/2}\mathbb{E} \left( \| \mathbf{d} \|_2^k \right) = O(n^{-k/2})
\]

\[
\mathbb{E} \left\{ 1(\mathcal{E}_0^c)\| (\mathbf{H} - \mathbf{I}_0) \mathbf{d} \|_2^{k/2} \right\} \leq \mathbb{E} \left( \| \mathbf{H} - \mathbf{I}_0 \|_2^{k/2} \| \mathbf{d} \|_2^{k/2} \right) \leq \left\{ \mathbb{E} \left( \| \mathbf{H} - \mathbf{I}_0 \|_2^k \right) \mathbb{E} \left( \| \mathbf{d} \|_2^k \right) \right\}^{1/2} = O(n^{-k/2})
\]

\[
\mathbb{E} \left\{ 1(\mathcal{E}_0^c)\| \mathbf{I}_0 \mathbf{d} \|_2^{k/2} \right\} \leq \mathbb{E} \left\{ 1(\mathcal{E}_0^c)\| \mathbf{I} \|_2^{k/2} \right\} \leq \left\{ \mathbb{P}(\mathcal{E}_0^c)\mathbb{E} \left( \| \mathbf{d} \|_2^k \right) \right\}^{1/2} = O(n^{-k/2})
\]

\[
\mathbb{E} \left( \| (\mathbf{H} - \overline{\mathbf{H}}) \mathbf{d} \|_2^{k/2} \right) \leq \left\{ 4^{k/2-1}\mathbb{E} \left\{ 1(\mathcal{E}_0)\| (\mathbf{H} - \overline{\mathbf{H}}) \mathbf{d} \|_2^{k/2} \right\} + 4^{k/2-1}\mathbb{E} \left\{ 1(\mathcal{E}_0^c)\| (\mathbf{H} - \overline{\mathbf{H}}) \mathbf{d} \|_2^{k/2} \right\} + 4^{k/2-1}\mathbb{E} \left\{ 1(\mathcal{E}_0^c)\| \mathbf{I}_0 \mathbf{d} \|_2^{k/2} \right\} \right\}^{1/2} = O(n^{-k/2}).
\]

\[\square\]

**Lemma 4.14.** Under the conditions of Lemma [4.11](1), \(\mathbb{E}(\| \mathbf{d} - \mathbf{d}_0 \|_2^{k/2}) = O(n^{-k/2}).\) Specifically, \(\| \mathbb{E}(\mathbf{d}) \|_2 = O(n^{-1}).\)

**Proof.** Let \(\mathbf{d}_0 = -\mathbf{I}_0^{-1}\mathbf{l}(\theta_0).\) Note that \(\hat{\theta} \) may not be in \(U(\rho).\)

\[
\mathbf{d} - \mathbf{d}_0 = \mathbf{I}_0^{-1}(\mathbf{I}_0 - \overline{\mathbf{H}})\mathbf{d} = \mathbf{I}_0^{-1}(\mathbf{I}_0 - \mathbf{H})\mathbf{d} + \mathbf{I}_0^{-1}(\mathbf{H} - \overline{\mathbf{H}})\mathbf{d}.
\]

By Lemma [4.13], \(\mathbb{E} \left\{ \| (\mathbf{H} - \overline{\mathbf{H}}) \mathbf{d} \|_2^{k/2} \right\} = O(n^{-k/2}),\) and

\[
\mathbb{E}(\| (\mathbf{H} - \mathbf{I}_0) \mathbf{d} \|_2^{k/2}) \leq \left\{ \mathbb{E} \left( \| \mathbf{H} - \mathbf{I}_0 \|_2^k \| \mathbf{d} \|_2^k \right) \right\}^{1/2} = O(n^{-k/2})
\]

\[
\mathbb{E} \left( \| \mathbf{d} - \mathbf{d}_0 \|_2^{k/2} \right) \leq 2^{k/2-1}\lambda_{-k/2}^n \mathbb{E} \left( \| (\mathbf{I}_0 - \mathbf{H}) \mathbf{d} \|_2^{k/2} \right) + 2^{k/2-1}\lambda_{-k/2}^n \mathbb{E} \left( \| (\mathbf{H} - \overline{\mathbf{H}}) \mathbf{d} \|_2^{k/2} \right) = O(n^{-k/2}).
\]

\[\square\]
Lemma 4.15. Given independent and identically distributed random matrices \( W_i \) with \( \mathbb{E}(\|W_i\|_2^2) < \infty \) and rank \( r > 0 \),

\[
\mathbb{E} \left( \left\| n^{-1} \sum W_i \right\|_2^2 \right) \leq n^{-1} r \mathbb{E}(\|W_1\|_2^2) + d \mathbb{E}(W_1)^2.
\]

Proof. Recall that \( \|W_i\|_2^2 \leq \|W_i\|_F^2 \leq r \|W_i\|_2^2 \).

\[
\mathbb{E} \left( \left\| \sum W_i \right\|_F^2 \right) = \mathbb{E} \left[ \text{Tr} \left\{ \left( \sum W_i \right) \left( \sum W_i \right)^\top \right\} \right]
= \sum_i \mathbb{E}(\|W_i\|_F^2) + \sum_{i \neq j} \text{Tr} \left\{ \mathbb{E}(W_i) \mathbb{E}(W_j^\top) \right\}
= n \mathbb{E}(\|W_1\|_F^2) + n(n - 1) \mathbb{E}(W_1)^2.
\]

\[
\mathbb{E} \left( \left\| n^{-1} \sum W_i \right\|_2^2 \right) \leq \mathbb{E} \left( \left\| n^{-1} \sum W_i \right\|_F^2 \right) \leq n^{-1} r \mathbb{E}(\|W_1\|_2^2) + d \mathbb{E}(W_1)^2.
\]

Higher order convergence can be controlled by the Marcinkiewicz–Zygmund inequality. \( \square \)

Lemma 4.16. Let \( \{Z_i\}_i^n \) be independent and identically distributed Hermitian matrices with \( \mathbb{E}(Z_i) = 0 \) and \( \mathbb{E}(\|Z_i\|_2^4) \leq \zeta^k \) for some positive numbers \( \zeta > 0 \) and \( k \geq 2 \).

\[
\mathbb{E} \left( \left\| n^{-1} \sum Z_i \right\|_2^k \right) \leq \left[ n^{-\frac{k}{2}} 4^{k-1} \{ e(k + 2 \log d) \}^k + n^{-k+1} 2^{3k-1} \{ e(k + 2 \log d) \}^k \right] \zeta^k.
\]

Proof. Let \( \{\varepsilon_i\}_i^m \) be independent and identically distributed Rademacher random variables.

\[
\mathbb{E} \left( \left\| \sum Z_i \right\|_2^k \right) = \mathbb{E} \left( \left\| \sum Z_i - \mathbb{E} \left( \sum Z_i \right) \right\|_2^k \right) = \mathbb{E} \left( \left\| \sum (Z_i - \mathbb{E}(Z_i)) \right\|_2^k \right)
\leq \mathbb{E} \left( \left\| \sum (Z_i - \mathbb{E}(Z_i)) \right\|_2^k \mid Z_i \right)
\mathbb{E} \left( \left\| \sum Z_i \right\|_2^k \right) \leq \mathbb{E} \left( \left\| \sum (Z_i - \mathbb{E}(Z_i)) \right\|_2^k \right).
\]
Note that each element of $Z_i - Z'_i$ is symmetrically distributed. Then,

$$
\mathbb{E}\left(\left\| \sum Z_i \right\|_2^k\right) \leq \mathbb{E}\left(\left\| \sum (Z_i - Z'_i) \right\|_2^k\right) = \mathbb{E}\left(\left\| \sum \varepsilon_i (Z_i - Z'_i) \right\|_2^k \mid \varepsilon_i\right)
$$

$$
= \mathbb{E}\left(\left\| \sum \varepsilon_i (Z_i - Z'_i) \right\|_2^k\right) \leq 2^{k-1} \mathbb{E}\left(\left\| \sum \varepsilon_i Z_i \right\|_2^k\right).
$$

By Theorem A.1 (2) in Chen et al. (2012), since $\varepsilon_i Z_i$ are independent and identically distributed symmetrically distributed Hermitian matrices,

$$
\left\{ \mathbb{E}\left(\left\| n^{-1} \varepsilon_i Z_i \right\|_2^k\right) \right\}^{1/k} \leq (ec)^{1/2} \left\| \left\{ \mathbb{E}\left( n^{-2} Z_i^2 \right) \right\}^{1/2} \right\|_2 + 2ec \left\{ \mathbb{E}\left( \max_i \left\| n^{-1} Z_i \right\|_2^k \right) \right\}^{1/k}
$$

$$
\leq n^{-1/2} (ec)^{1/2} \left\| \mathbb{E}(Z_i^2) \right\|_2^{1/2} + n^{-1/2} ec \left\{ \mathbb{E}(\left\| Z_i \right\|_2^k) \right\}^{1/k}
$$

$$
\leq n^{-1/2} (ec)^{1/2} \left\| \mathbb{E}(Z_i^2) \right\|_2^{1/2} + n^{-1/2}+ \frac{1}{2} ec \left\{ \mathbb{E}(\left\| Z_i \right\|_2^k) \right\}^{1/k},
$$

where $c = k + 2 \log d$. Therefore,

$$
\mathbb{E}\left(\left\| n^{-1} \sum Z_i \right\|_2^k\right) \leq 2^{k-1} \mathbb{E}\left(\left\| n^{-1} \varepsilon_i Z_i \right\|_2^k\right)
$$

$$
\leq n^{-k/2} 4^{k-1} (ec)^{k/2} \left\| \mathbb{E}(Z_i^2) \right\|_2^{k/2} + n^{-k+1} 2^{3k-1} (ec)^k \left\{ \mathbb{E}(\left\| Z_i \right\|_2^k) \right\}
$$

$$
\leq \left\{ n^{-k/2} 4^{k-1} (ec)^{k/2} + n^{-k+1} 2^{3k-1} (ec)^k \right\} \zeta^k.
$$

\[\square\]

**Lemma 4.17.** Under Assumptions 3-4,

$$
l(\theta_0) d^\top (H - H) I_0^{-1} = W + W',
$$

where $W = 2^{-1} l(\theta_0) \{ Q(d_0^{\otimes 2}) \}^\top I_0^{-1}$, with $\mathbb{E}(W) = O(n^{-2})$, $\mathbb{E}(\|W\|_2) = O(n^{-3/2})$, and $\mathbb{E}(\|W'\|_2) = O(n^{-2})$.

Additionally, when Assumption 3 holds, $\mathbb{E}(\|W\|_2^2) = O(n^{-3})$. 

63
Proof. Recall that by Assumption 3 and Lemma 4.11 (2), $\mathbb{P}(\tilde{E}_0^c) = O(n^{-2})$.

$$
\mathbb{E}\{1(\tilde{E}_0^c)\|l(\theta_0)\d^\top (\H - \H)I_0^{-1}\|_2\} \leq \lambda^{-1} \left[\mathbb{P}(\tilde{E}_0^c)\mathbb{E}\{\|l(\theta_0)\|_2^4\}\right]^{1/4} \left[\mathbb{E}\left\{\|\d^\top (\H - \H)\|_2^2\right\}\right]^{1/2}
= O(n^{-2}).
$$

(30)

Under $\tilde{E}_0$, we have $\hat{\theta} \in U(\rho)$. Lemma 4.10 indicates

$$(\H - \H)\d = 2^{-1}Q(\d^\otimes 2) + R_{13} + R_{23},$$

where $\|R_{13}\|_2 \leq (6n)^{-1} \sum m(X_{ij})\|\d\|_2^3$, and $\|R_{23}\|_2 \leq R_2\|\d\|_2^2$ with $\mathbb{E}(R_2^k) = O(n^{-k/2})$.

$$
l(\theta_0)\d^\top (\H - \H)I_0^{-1} = 2^{-1}l(\theta_0)\{Q(\d_0^\otimes 2)\}^\top I_0^{-1} + 2^{-1}l(\theta_0)\{Q(\d_0^\otimes 2) - Q(\d_0^\otimes 2)\}^\top I_0^{-1} + l(\theta_0)R_{13}I_0^{-1} + l(\theta_0)R_{23}I_0^{-1}.
$$

(31)  (32)  (33)

Also, under $\tilde{E}_0$, $\|\d\|_2 \leq 2\|l(\theta_0)\|_2\{(1 - \rho)\lambda_\_\}^{-1}$ and $\|\H - I_0\|_2 \leq \rho\lambda_\_/2$.

First term (31)

Consider $l(\theta_0)\{Q(\d_0^\otimes 2)\}^\top I_0^{-1}$.

Note that this random element does not depends on $\tilde{E}_0$, instead, on $l(\theta_0; X_1)$ and $h(X_1)$.

Lemma 4.9 gives for $j = 1, \ldots, d$, $\|G_j(\theta_0; X_i)\|_2 \leq h(X_i)$ and $\|\mathbb{E}\{G_j(\theta_0; X_i)\}\|_2 \leq \lambda_h$. Then,

$$
\mathbb{E}\left\|\left[\d_0^\top \mathbb{E}\{G_1(\theta_0; X_i)\} \d_0 \right] : \cdots : \left[\d_0^\top \mathbb{E}\{G_d(\theta_0; X_i)\} \d_0 \right]\right\|_1
\leq d\lambda^{-3}\|l(\theta_0)\|_2^3\lambda_h
$$

(34)

$$
\mathbb{E}\left\|\left[I_0^{-1}l(\theta_0)\{Q(\d_0^\otimes 2)\}\right]_2\right\|_2
\leq d\lambda^{-3}\mathbb{E}\{\|l(\theta_0)\|_2^3\lambda_h\} = O(n^{-3/2}).
$$

(35)
Note that $\mathbb{E}(d_0) = 0$. Let $d_{0,i} = -I_0^{-1}l(\theta_0; X_i)$ and $d_0 = n^{-1} \sum d_{0,i}$.

$$-I_0^{-1}l(\theta_0) \{ Q(d_0^{\otimes 2}) \}^T = n^{-3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n Q(d_{0,i} \otimes d_{0,j} \otimes d_{0,k})$$

$$\mathbb{E} \left[ -I_0^{-1}l(\theta_0) \{ Q(d_0^{\otimes 2}) \}^T \right] = n^{-3} \sum_i \mathbb{E} \{ Q(d_{0,i} \otimes d_{0,i} \otimes d_{0,i}) \}$$

$$= n^{-2} \mathbb{E} \{ Q(d_{0,1} \otimes d_{0,1} \otimes d_{0,1}) \}.$$

Hence, its expectation can be controlled.

$$\left\| \mathbb{E} \left[ I_0^{-1}l(\theta_0) \{ Q(d_0^{\otimes 2}) \}^T \right] \right\|_2 \leq n^{-2} \mathbb{E} \{ d\lambda_h \|l(\theta_0; X_1)\|_2^3 \| I_0^{-1} \|_2^3 \} = O(n^{-2}). \quad (36)$$

Let $W = 2^{-1}l(\theta_0) [ Q(d_0^{\otimes 2}) ]^T I_0^{-1}$. Therefore, $\|\mathbb{E}(W)\|_2 = O(n^{-2})$ and $\mathbb{E}(\|W\|_2) = O(n^{-3/2})$.

When Assumption 3 holds, by definition (34)

$$\mathbb{E} \left( \|W\|_2^2 \right) \leq d^2 \lambda^{-4} \lambda_h^2 \mathbb{E} \left\{ \|l(\theta_0)\|_2^6 \right\} = O(n^{-3}).$$

On the other hand, by (34),

$$\mathbb{E} \left\{ \|l(\tilde{\theta}_0^c) W\|_2 \right\} \leq d \lambda h^{-3} \mathbb{E} \left\{ \|l(\tilde{\theta}_0^c) \|_2^3 \lambda_h \right\} \leq d \lambda h^{-3} \lambda_h \left\{ \mathbb{P}(\tilde{\theta}_0^c) \right\}^{1/4} \left\{ \mathbb{E} (\|l\|_2^4) \right\}^{3/4} = O(n^{-2}).$$

Also note that together with (30),

$$\mathbb{E} \left\{ \|l(\tilde{\theta}_0^c) W\|_2 \right\} \leq \mathbb{E} \left\{ \|l(\tilde{\theta}_0^c) \|_2^4 (H - H) I^{-1} \|_2 \right\} + \mathbb{E} \left\{ \|l(\tilde{\theta}_0^c) W\|_2 \right\} = O(n^{-2}). \quad (37)$$

We are going to show that $\mathbb{E} \left\{ \|l(\tilde{\theta}_0^c) W\|_2 \right\} = O(n^{-2}).$

**Second term** (32)

By definition

$$\nabla^3 \mathbb{E} \left\{ L(\theta_0; X_i)(d^{\otimes 2}) \right\} - \nabla^3 \mathbb{E} \left\{ L(\theta_0; X_i)(d^{\otimes 2}) \right\} = \begin{bmatrix} (d + d_0)^T \mathbb{E} \{ G_1(\theta_0; X_i) \} (d - d_0) \\ \vdots \\ (d + d_0)^T \mathbb{E} \{ G_d(\theta_0; X_i) \} (d - d_0) \end{bmatrix}.$$
Therefore, for $j = 1, \ldots, d$, under $\tilde{\mathcal{E}}_0$

$$\|(d + d_0)^\top \mathbb{E}\{G_j(\theta_0; X_j)\} (d - d_0)\| 
\leq \|d\|_2 + \|d_0\|_2 \lambda_h \lambda^{-1}_- (\|H - H\|_2 + \|H - I_0\|_2) \|d\|_2 
\leq \frac{2 \lambda_h [2\{\lambda_-(1 - \rho)\}^{-1} + \lambda^{-1}_-]}{\lambda^2 (1 - \rho)} \|l(\theta_0)\|_2^2 \left\{ \frac{4 \lambda_h}{(1 - \rho) \lambda_-} \|l(\theta_0)\| + \|H - I_0\|_2 \right\} 
\leq \frac{8 \lambda_h^2 [2\{\lambda_-(1 - \rho)\}^{-1} + \lambda^{-1}_-]}{\lambda^3 (1 - \rho)^2} \|l(\theta_0)\|_2^3 + \frac{2 \lambda_h [2\{\lambda_-(1 - \rho)\}^{-1} + \lambda^{-1}_-]}{\lambda^2 (1 - \rho)} \|l(\theta_0)\|_2^2 \|H - I_0\|_2.$$

Consequently,

$$\mathbb{E}\left( \|l(\theta_0) [Q(d^{\otimes 2}) - Q(d_0^{\otimes 2})]^\top I_0^{-1}\|_2 \right) 
\leq \mathbb{E}\left( \|l(\theta_0)\|_2 \left\| [Q(d^{\otimes 2}) - Q(d_0^{\otimes 2})]^\top I_0^{-1}\|_1 \right\| 
\leq \mathbb{E}\left\| [Q(d_0^{\otimes 2})]^\top I_0^{-1}\|_1 \right\| + \mathbb{E}\left\| [Q(d_0^{\otimes 2})]^\top I_0^{-1}\|_1 \right\| + \mathbb{E} \|l(\theta_0)\|_2^3 \|I_0 - H\|_2 
= O(n^{-2}). \tag{38}$$

The last term $\left(33\right)$

$$\mathbb{E}\left\{ \|l(\tilde{\mathcal{E}}_0)\| \|l(\theta_0)R_{13}I_0^{-1}\|_2 \right\} \leq \lambda^{-1}_- 3^{-1} \lambda_m \mathbb{E}\left\{ \|l(\tilde{\mathcal{E}}_0)\| \|d\|_2 \right\} 
\leq 8 \lambda^{-1}_- 3^{-1} \lambda_m \mathbb{E}\left( \|l(\theta_0)\|_2 \right\}^4 \left\{ (1 - \rho) \lambda_- \right\}^{-3} = O(n^{-2}), \tag{39}$$

$$\mathbb{E}\left\{ \|l(\tilde{\mathcal{E}}_0)\| \|l(\theta_0)R_{23}I_0^{-1}\|_2 \right\} \leq \lambda^{-1}_- \mathbb{E}\left\{ \|l(\tilde{\mathcal{E}}_0)\| \|d\|_2^3 \right\} 
\leq 4 \lambda^{-1}_- \mathbb{E}\left( \|l(\theta_0)\|^3 \|R_2\| \right) \left\{ (1 - \rho) \lambda_- \right\}^{-2} = O(n^{-2}). \tag{40}$$

With $\left(38\right)$, $\left(39\right)$, $\left(40\right)$, and $\left(37\right)$

$$\mathbb{E}\left\{ \|l(\tilde{\mathcal{E}}_0)W\|_2 \right\} = O(n^{-2})$$

$$\mathbb{E}\left\{ \|W\|_2 \right\} \leq \mathbb{E}\left\{ \|l(\tilde{\mathcal{E}}_0)W\|_2 \right\} + \mathbb{E}\left\{ \|l(\tilde{\mathcal{E}}_0)W\|_2 \right\} = O(n^{-2}).$$

$\square$
Lemma 4.18. Under Assumptions 1-3, for each node, let \( W_1 = l(\theta_0) d_0^\top (I_0 - H) I_0^{-1} \) and \( W_2 = 2^{-1} l(\theta_0) \{ \nabla^3 Q(d_0^{\otimes 2}) \}^\top I_0^{-1} \). Then,

\[
\begin{align*}
l(\theta_0)(d - d_0)^\top &= W_1 - W_2 + W' \\
\mathbb{E} (\| W_1 \|_2) &= O(n^{-3/2}), \| \mathbb{E}(W_1) \|_2 = O(n^{-2}) \\
\mathbb{E} (\| W_2 \|_2) &= O(n^{-3/2}), \| \mathbb{E}(W_2) \|_2 = O(n^{-2}) \\
\mathbb{E} (\| W' \|_2) &= O(n^{-2}).
\end{align*}
\]

Additionally, when Assumption 3 holds, \( \mathbb{E} (\| W_1 \|_2^2) = O(n^{-3}) \) and \( \mathbb{E} (\| W_2 \|_2^2) = O(n^{-3}) \).

Proof. Recall that \( d - d_0 = I_0^{-1} (I_0 - H) d \),

\[
l(\theta_0)(d - d_0)^\top = l(\theta_0) d^\top (I_0 - H) I_0^{-1} \\
= l(\theta_0) d_0^\top (I_0 - H) I_0^{-1} + \\
l(\theta_0)(d - d_0)^\top (I_0 - H) I_0^{-1} + \\
l(\theta_0) d^\top (H - H) I_0^{-1}.
\]

Consider \( l(\theta_0)d_0^\top (I_0 - H) I_0^{-1} \):

\[
\begin{align*}
\mathbb{E} \left\{ \| l(\theta_0)d_0^\top (I_0 - H) I_0^{-1} \|_2 \right\} &\leq \lambda^{-2} \mathbb{E} \left\{ \| l_1 \|_2^2 \| I_0 - H \|_2 \right\} = O(n^{-3/2}) \\
\mathbb{E} \left\{ l(\theta_0)d_0^\top (I_0 - H) \right\} &\leq n^{-3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \mathbb{E} \left\{ l_i l_j^\top I_0^{-1} (H_k - I_0) \right\} \\
&= n^{-2} \mathbb{E} \left\{ l_i l_j^\top I_0^{-1} (H_k - I_0) \right\} \\
\| \mathbb{E} \left\{ l(\theta_0)d_0^\top (I_0 - H) \right\} \|_2^2 &\leq O(n^{-2}).
\end{align*}
\]

Let \( W_1 = l(\theta_0)d_0^\top (I_0 - H) I_0^{-1} \). Then, \( \mathbb{E}(\| W_1 \|_2) = O(n^{-3/2}) \) and \( \| \mathbb{E}(W_1) \|_2 = O(n^{-2}) \). When Assumption 3 holds,

\[
\mathbb{E} (\| W_1 \|_2^2) \leq \lambda^{-4} \mathbb{E}\left\{ \| l \|_2^4 \| I_0 - H \|_2 \right\} \leq \lambda^{-4} \left\{ \mathbb{E} (\| l \|_2^6) \right\}^{2/3} \left\{ \mathbb{E} (\| I_0 - H \|_2^6) \right\}^{1/3} = O(n^{-3}).
\]
Consider $l(\theta_0)(d - d_0)\top(I_0 - H)I_0^{-1}$. By Lemma 4.14

$$\mathbb{E}\{||l(\theta_0)(d - d_0)\top(H - I_0)I_0^{-1}||_2\} \leq \lambda^{-1} \mathbb{E}\{||l(\theta_0)||_2 \cdot ||d - d_0||_2 \cdot ||H - I_0||_2\}$$

$$\leq \lambda^{-1} \{\mathbb{E}(||l(\theta_0)||_2^4)\mathbb{E}(||H - I_0||_2^4)\}^{1/4} \{\mathbb{E}(||d - d_0||_2^4)\}^{1/2} = O(n^{-2}).$$

Consider $l(\theta_0)d\top(H - \bar{H})I_0^{-1}$. By Lemma 4.17

$$l(\theta_0)d\top(H - \bar{H})I_0^{-1} = W_2 + W'_2,$$

where $W_2 = 2^{-1}l(\theta_0)\{Q(d_0^{\otimes 2})\}^\top I_0^{-1}$, with $\mathbb{E}(W_2)_2 = O(n^{-2})$, $\mathbb{E}(W_2)_2 = O(n^{-3/2})$, and $\mathbb{E}(W'_2)_2 = O(n^{-2})$. The proof is complete by putting them together.

Lemma 4.19.

1. In a single-node, when the $k$th ($k \geq 2$) moments of $l(\theta_0; X_i)$ and $H(\theta_0; X_i)$ exist, then,

$$\mathbb{E}\{||l \otimes (H - I_0)||_2^{k/2}\} = O(n^{-k/2})$$

$$\mathbb{E}\{l \otimes (H - I_0)\} = n^{-1}Q_{12},\text{ where } Q_{12} = \mathbb{E}[l(\theta_0; X_1) \otimes \{H(\theta_0; X_1) - I_0\}].$$

2. When $m$ independent and identically distributed copies $l_i$ and $H_i$ exist, and $k \geq 4$,

$$m^{-1} \sum l_i \otimes (H_i - I_0) = n^{-1}Q_{12} + R, \text{ with } \mathbb{E}(R)_2^{k/2} = O(m^{-k/4}n^{-k/2}),$$

where $Q_{12} = \mathbb{E}[l(\theta_0; X_1) \otimes \{H(\theta_0; X_1) - I_0\}].$

Proof. For the first statement, note that $\mathbb{E}\{l(\theta_0; X_i)\} = 0$ and $\mathbb{E}\{H(\theta_0; X_i) - I_0\} = 0$. Let $W = l \otimes (H - I_0) = \{l_1(H - I_0), \ldots, l_d(H - I_0)\}^\top$.

$$W^\top W = \{l \otimes (H - I_0)\}^\top \times \{l \otimes (H - I_0)\} = (l^\top l) \otimes (H - I_0)^2$$

$$||l \otimes (H - I_0)||_2 = ||W^\top W||_2^{1/2} = ||l||_2 \cdot ||H - I_0||_2$$

$$\mathbb{E}(||W||_2^{k/2}) = \mathbb{E}(||l||_2^{k/2}||H - I_0||_2^{k/2}) \leq \{\mathbb{E}(||l||_2^{k/2})\mathbb{E}(||H - I_0||_2^{k/2})\}^{1/2} = O(n^{-k/2})$$

$$\mathbb{E}(W) = n^{-2} \sum \mathbb{E}\{l_i \otimes (H_i - I_0)\} = n^{-1}Q_{12}.$$
For the second statement, denote the $j$th element of $W_i$ and $Q_{12}$ by $W_{i,j} = l_j(H_i - I_0)$ and $E_{12,j} = \mathbb{E}(W_{1,j})$, and $W_{i,j}$ is Hermitian. Let $Z_{i,j} = W_{i,j} - n^{-1}E_{12,j}$, and

$$\mathbb{E} \left( \|Z_{i,j}\|^{k/2} \right) \leq 2^{k/2-1} \mathbb{E} \left( \|W_{i,j}\|^{k/2} \right) + 2^{k/2-1} \left( n^{-1} \|E_{12,j}\|^{k/2} \right) = O(n^{-k/2}).$$

By Lemma 4.16,

$$\mathbb{E} \left( \left\| \sum_i m^{-1}Z_{i,j} \right\|^{k/2} \right) = O(m^{-k/4}n^{-k/2})$$

$$\mathbb{E} \left( \left\| m^{-1} \sum_i W_i - n^{-1}Q_{12} \right\|^{k/2} \right)$$

$$= \mathbb{E} \left\{ \left\| \left( m^{-1} \sum_i W_i - n^{-1}Q_{12} \right)^\top \left( m^{-1} \sum_i W_i - n^{-1}Q_{12} \right) \right\|^{k/4} \right\}$$

$$= \mathbb{E} \left\{ \left\| \sum_{j=1}^d \left( m^{-1} \sum_i Z_{i,j} \right)^\top \left( m^{-1} \sum_i Z_{i,j} \right) \right\|^{k/4} \right\}$$

$$\leq d^{k/4-1} \sum_{j=1}^d \mathbb{E} \left( \left\| m^{-1} \sum_i Z_{i,j} \right\|^{k/2} \right) = O(m^{-k/4}n^{-k/2}).$$

\[\Box\]