Supplementary Material for

Distributed Algorithms for U-statistics-based Empirical Risk Minimization

Lanjue Chen[1], Alan T.K. Wan[2], Shuyi Zhang[1] and Yong Zhou[1]

[1]*Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education, Academy of Statistics and Interdisciplinary Sciences and School of Statistics, East China Normal University, Shanghai*

[2]*Department of Management Sciences, School of Data Science and Department of Biostatistics, City University of Hong Kong, Kowloon, Hong Kong*

This Supplementary Material contains seven sections. Section 1 contains the properties of the naive estimator. In Section 2, we provide results on the properties of the OS-ERM and SU-ERM estimators when V_0 is degenerate. Proofs of some useful lemmas, and theorems related to the naive estimator, variance estimation and the estimators obtained via the iterative algorithms are contained in Sections 3 – 6. Section 7 provides details on the computation of running times of the various estimators.

1 Properties of the naive estimator

In the following theorem and corollaries, we derive the theoretical properties of the naive estimator $\bar{\theta}_N = K^{-1}\sum_{k=1}^{K}\widehat{\theta}_n^k$, including the upper bound and the optimal convergence rate of its MSE, and the consistency properties. These results on the naive estimator enable a formal comparison with the properties of our proposed estimators.

Theorem S1. *Let Conditions (C1)-(C6) be satisfied. In particular, there exists $\zeta > 0$ such that $p = O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau > 0$. Then the MSE of the naive estimator is bounded above by the right-hand-side of the following inequality:*

$$\mathbb{E}[\|\bar{\theta}_N - \theta^*\|_2^2] \leq \frac{4A}{N} + C\left(\frac{p}{n^2} + \frac{p^2}{n^3}\right)A + C\frac{\eta_{n,K,p}}{\sqrt{n}K}\sqrt{A} + C\left(\frac{p^2\eta_{n,K,p}^2}{n^2} + \frac{p^8}{n^8K} + \frac{p\log n\log\log n}{nN}\right),$$

where $A = \mathbb{E}[\|[\nabla^2\mathcal{L}_0(\theta^*)]^{-1}g(\theta^*;Z)\|_2^2]$ with $g(\theta^*;Z) = \mathbb{E}[\nabla\ell(\theta^*;Z,Z')|Z]$, C is independent of (K,n,N,p), and $\eta_{n,K,p}^2 = p^2/n^2 + p^8/n^8$. If one also assumes $p = o(n)$, then

$$\mathbb{E}[\|\bar{\theta}_N - \theta^*\|_2^2] = \frac{4A}{N} + O\left(\frac{p}{N}\sqrt{\frac{A}{n}} + \frac{p^2}{n^2} + \frac{p\log n\log\log n}{nN}\right).$$

Corollary S1. *Let Conditions (C1)-(C6) be satisfied. The naive estimator $\bar{\theta}_N$ is consistent if $p = o(n)$.*

Corollary S2. *Let Conditions (C1)-(C6) be satisfied. If $pK = O(n)$, the naive estimator $\bar{\theta}_N$ achieves the optimal MSE convergence rate $\mathbb{E}\|\bar{\theta} - \theta^*\|_2^2 = O(p/N)$, i.e., the same MSE convergence*

rate of the global estimator $\widehat{\theta}_N$. If one also assumes $pK = o(n)$, then

$$\mathbb{E}[\|\bar{\theta}_N - \theta^*\|_2^2] = \frac{4A}{N} + o\left(\frac{p}{N}\right),$$

which is the asymptotic upper bound of the MSE of $\widehat{\theta}_N$.

Remark S1. *Corollaries 1 and S2 show that if the initial estimator of the proposed SU-ERM and OS-ERM estimators are chosen such that $\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}((p/n)^{1/2})$, then the MSE of the two proposed estimators and the naive estimator all attain the optimal bound $4A/N$ under the condition of $pK = o(n)$. Clearly, this is relevant only when the number of machines and the dimension of the parameters are small relative to the size of the data on each machine. On the other hand, when the number of machines and the covariate dimension are large, the naive estimator cannot attain the optimal rate and the MSE bound, but the SU-ERM and OS-ERM estimators can still achieve the asymptotic properties of the global estimator $\widehat{\theta}_N$, provided that an initial estimator that satisfies $\|\widehat{\theta}_0 - \theta^*\|_2 = O_{\mathbb{P}}(K^{-1/2})$ is selected. The MSE bound of our proposed estimators depends crucially on the initial estimator $\widehat{\theta}_0$. A superior initial estimator has the effect of improving the estimation efficiency. Thus, the rates of convergence of the SU-ERM and OS-ERM estimators will remain optimal in the face of a large number of machines if an appropriate initial estimator is chosen, e.g., $\widehat{\theta}_0 = K^{-1}\sum_{k=1}^{K}\widehat{\theta}_n^k$.*

2 The case of degenerate V_0

Theorem 4 in the main paper focuses on the case where V_0 is non-degenerate, and shows that $\widetilde{\theta}_N$ and $\widetilde{\theta}_N^Q$ are asymptotically equivalent to $\widehat{\theta}_N$ when $V_0 > 0$. The following analysis shows that when V_0 is degenerate, $\widehat{\theta}_N$ and $\widetilde{\theta}_N$ perform differently. Note that if V_0 is degenerate, then for any $v_0 \in \mathbb{R}^p$ $(v_0 \neq 0)$ such that $v_0^\top \nabla^2\mathcal{L}_0(\theta^*)^{-1}V_0\nabla^2\mathcal{L}_0(\theta^*)^{-1}v_0 = 0$, we have

$$\frac{Nv_0^\top(\widehat{\theta}_N - \theta^*)}{\sqrt{2v_0^\top\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}V_1\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}v_0}} \xrightarrow{\mathcal{D}} \sum_{l=1}^{\infty}\lambda_l(\chi_{1,l}^2 - 1),$$

where $V_1 = \mathbb{E}[\nabla\ell(\theta^*; Z, Z')\nabla\ell^\top(\theta^*; Z, Z')]$, $\{\chi_{1,l}^2\}_{l=1}^{\infty}$ are independent χ_1^2 random variables with one degree of freedom and $\{\lambda_l\}_{l=1}^{\infty}$ are eigenvalues satisfying

$$\mathbb{E}\left[\left|\frac{\sqrt{2}v_0^\top\nabla^2\mathcal{L}_0(\theta^*)^{-1}h(\theta^*; Z, Z')}{\sqrt{v_0^\top\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}V_1\nabla^2\mathcal{L}_0\left(\theta^*\right)^{-1}v_0}} - \sum_{l=1}^{L}\lambda_l h_{1,l}(\theta^*; Z)h_{2,l}(\theta^*; Z')\right|^2\right] \overset{L\to\infty}{\longrightarrow} 0$$

with $\{h_{1,l}(\theta^*; z), h_{2,l}(\theta^*; z)\}_{l=1}^{\infty}$ being eigenfunctions. The following theorem, which nests Theorem 4 as a special case, gives the asymptotic distributions of the SU-ERM and OS-ERM estimators under both the non-degenerate and degenerate cases.

Theorem S2. *Let Conditions (C1)-(C6) be satisfied. In particular, there exists $\zeta > 0$ such that $p = O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau > 0$. Assume the initial estimator $\widehat{\theta}_0$ lies in $U(\theta^*, \tilde{\rho})$, where $\tilde{\rho} = \min\{(1-\rho)\lambda_-\delta_\rho/(32\lambda_+), \sqrt{(1-\rho)\lambda_-\delta_\rho/(32M)}\}$ with $\delta_\rho = \min\{\rho, \rho\lambda_-/(4M)\}$, and $\|\widehat{\theta}_0 - \theta^*\|_2 =$*

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

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

where $g(\theta; z) = \mathbb{E}\left[\nabla \ell\left(\theta; Z, Z'\right) | Z = z\right]$, $h(\theta^*; z, z') = \nabla \ell(\theta^*; z, z') - g(\theta^*; z) - g(\theta^*; z')$, $G_{n,K,p}(\widehat{\theta}_0)$ *is a random function of* $\widehat{\theta}_0$ *satisfying* $\|G_{n,K,p}(\widehat{\theta}_0)\|_2 = O_{\mathbb{P}}((p/n)^{1/2}\|\widehat{\theta}_0 - \theta^*\|_2)$, *and* $\|\omega_{n,K,p}\|_2 = o_{\mathbb{P}}(\|\widetilde{\theta}_N - \theta^*\|_2)$. *Furthermore, (i) if* V_0 *is non-degenerate and* $pK = o(n)$, *then* $\widetilde{\theta}_N$ *is asymptotically equivalent to* $\widehat{\theta}_N$, *and for any* $v_0 \in \mathbb{R}^p$ $(v_0 \neq 0)$,

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

(ii) on the other hand, if V_0 *is degenerate and* $pK = o(n^2)$, *then for any* $v_0 \in \mathbb{R}^p$ $(v_0 \neq 0)$ *satisfying* $v_0^\top \nabla^2 \mathcal{L}_0(\theta^*)^{-1} V_0 \nabla^2 \mathcal{L}_0(\theta^*)^{-1} v_0 = 0$, *we have*

$$\frac{N v_0^\top (\widetilde{\theta}_N - \theta^*)}{\sqrt{2K \nabla^2 \mathcal{L}_0\left(\theta^*\right)^{-1} V_1 \nabla^2 \mathcal{L}_0\left(\theta^*\right)^{-1}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

This theorem also holds for the OS-ERM estimator. One can obtain the analogous result for the OS-ERM estimator by replacing $\widetilde{\theta}_N$ *by* $\widetilde{\theta}_N^Q$ *everywhere in the statements of the theorem and replacing* $\widehat{\theta}_0 \in U(\theta^*, \tilde{\rho})$ *by* $\widehat{\theta}_0 \in U(\theta^*, \rho')$, *where* ρ' *is defined in Proposition 1.*

Remark S2. *In Theorem S2, we present the asymptotic normality in the form of* $\sqrt{N} v_0^\top (\widetilde{\theta}_N - \theta^*)$. *We introduce the vector* v_0 *to present the results in a one-dimensional space, because* p, *the parameter dimension, is divergent here, being different from the classical* U*-estimation, where* p *is fixed.*

3 Proofs of Lemmas 1 – 3

Proof of Lemma 1. Under Condition (C6), by applying Theorem 1 in de la Peña (1992) and setting $\Phi(x) = x^{2\nu}$, we obtain

$$\begin{aligned} \mathbb{E}\left[\left\|\nabla \mathcal{L}_n^k\left(\theta^*\right)\right\|_2^{2\nu}\right] &= \mathbb{E}\left\| \frac{1}{n(n-1)} \sum_{1 \le i \neq j \le n} \nabla \ell\left(\theta^*; Z_{k,i}, Z_{k,j}\right) \right\|_2^{2\nu} \\ &\le 8^{2\nu} \mathbb{E}\left\| \frac{1}{n(n-1)} \sum_{1 \le i \neq j \le n} \nabla \ell\left(\theta^*; Z_{k,i}, \widetilde{Z}_{j,k}\right) \right\|_2^{2\nu}, \end{aligned}$$

where $\{\widetilde{Z}_{i,k}\}_{i=1}^n$ is an independent copy of $\{\tilde{Z}_{i,k}\}_{i=1}^n$. From Jensen's inequality, we have

$$\mathbb{E}\left\| \frac{1}{n(n-1)} \sum_{1 \le i \neq j \le n} \nabla \ell\left(\theta^*; Z_{k,i}, \widetilde{Z}_{j,k}\right) \right\|_2^{2\nu} \le \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left\| \frac{1}{n-1} \sum_{j \neq i} \nabla \ell\left(\theta^*; Z_{k,i}, \widetilde{Z}_{j,k}\right) \right\|_2^{2\nu}.$$

Let i be fixed. By the conditional expectation on $Z_{k,i}$, we can write

$$\mathbb{E}\left\|\frac{1}{n-1}\sum_{j\neq i}\nabla\ell\left(\theta^*;Z_{k,i},\widetilde{Z}_{j,k}\right)\right\|_2^{2\nu} = \mathbb{E}\left\{\mathbb{E}\left\|\frac{1}{n-1}\sum_{j\neq i}\nabla\ell\left(\theta^*;Z_{k,i},\widetilde{Z}_{j,k}\right)\right\|_2^{2\nu}\Big|\widetilde{Z}_{j,k}\right\}$$
$$= \mathbb{E}\left\{\mathbb{E}\left\|\frac{1}{n-1}\sum_{j\neq i}\nabla\ell\left(\theta^*;Z_{k,i},z\right)\right\|_2^{2\nu}\Big|_{z=\widetilde{Z}_{j,k}}\right\}.$$

By Condition (C6), $\mathbb{E}\left[\|\sqrt{p}\nabla\ell(\theta;Z,z)\|_2^{16}\right]\leq G^{16}(z)$ for all $z\in\mathcal{Z}$. Then by replacing $\nabla f(\theta^*;X)$ by $\sqrt{p}\nabla\ell(\theta;Z,z)$ in the proof of Lemma 7 of Zhang et al. (2013), we have

$$\mathbb{E}\left\|\frac{\sqrt{p}}{n-1}\sum_{j\neq i}\nabla\ell\left(\theta^*;Z_{k,i},z\right)\right\|_2^{2\nu}\leq C\frac{G^{2\nu}(z)}{n^\nu}$$

for all $z\in\mathcal{Z}$, where C is a constant unrelated to (n,K,N,p,z). The property that C does not depend on any specific z is induced by Theorem 2.1 of de Acosta (1981) and the Burkholder's inequality (Burkholder, 1973). Hence it can be derived that $\mathbb{E}\left\|\frac{1}{n-1}\sum_{j\neq i}\nabla\ell\left(\theta^*;Z_{k,i},\widetilde{Z}_{j,k}\right)\right\|_2^{2\nu}\leq C\frac{p^\nu G^{2\nu}}{n^\nu}$. This proves the first result of the lemma.

The second claim on the upper bound on $\mathbb{E}\left[\left\|\nabla^2\mathcal{L}_n^k(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\right\|_2^{2\nu}\right]$ can be obtained by using the similar techniques of Lemma 7 of Zhang et al. (2013). In fact, by applying Theorem A.1(2) of Chen et al. (2012), we obtain

$$\mathbb{E}\left[\left\|\nabla^2\mathcal{L}_n^k(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\right\|_2^{2\nu}\right]\leq(5H)^{2\nu}\frac{(\log p)^\nu}{n^\nu}+(4eH)^{2\nu}\frac{(\log p)^{2\nu}}{n^{2\nu-1}},$$

where e is Euler's number. When there exists $\zeta>0$ such that $p=O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau>0$, it can be derived that $\mathbb{E}\left[\left\|\nabla^2\mathcal{L}_n^k(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\right\|_2^{2\nu}\right]\leq C'\frac{(\log p)^\nu H^{2\nu}}{n^\nu}$ holds for $\nu=8$, and hence holds for all $\nu=1,2,\ldots,8$ by using the Jensen's inequality. The claims for $\mathcal{L}_N(\theta)$ can be derived analogously by recognizing that $p=O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ implies $p=O(N^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$. □

Proof of Lemma 2. By applying the union bound, we can write

$$\mathbb{P}\left(\bigcup_{k=0}^K\mathcal{E}_j^c\right)\leq\sum_{k=1}^K\mathbb{P}\left(M_k>2M\right)+\sum_{k=1}^K\mathbb{P}\left(\left\|\nabla^2\mathcal{L}_n^k(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\right\|_2>\frac{\rho\lambda_-}{4}\right)$$
$$+\mathbb{P}\left(M_N>2M\right)+\mathbb{P}\left(\left\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}_0(\theta^*)\right\|_2>\frac{\rho\lambda_-}{2}\right)$$
$$+\mathbb{P}\left(\left\|\nabla\mathcal{L}_N(\theta^*)\right\|_2>\frac{(1-\rho)\lambda_-}{2}\delta_1(\rho,\lambda_-,\lambda_+,M)\right).$$

Applying Lemma 1 and using the properties of the higher moments of U-statistics stated in Theorem 1 of Lee (2019, Chapter 1.5), we can show that there exists some constant C'' independent of (n,K,p,N) such that

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

This completes the proof of the lemma. □

Proof of Lemma 3. Along the lines of the proof of Lemma 6 of Zhang et al. (2013), we first prove that $\mathcal{S}_N(\theta)$ is $(1-\rho)\lambda_-$-strongly convex over the ball $U := \left\{\theta \in \mathbb{R}^p : \|\theta - \widehat{\theta}_N\|_2 < \delta_\rho\right\}$ under the event $\mathcal{E}_0 \cap \mathcal{E}_1$. Since $\nabla^2\mathcal{S}_N(\theta) = \nabla^2\mathcal{L}_n^1(\theta)$, this claim also implies the strong convexity of $\mathcal{L}_n^1(\theta)$ over $\theta \in U$. In fact, for $\forall\gamma \in U$, from the triangle inequality, we have

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

Now, consider the first term on the right hand side in the above expression. Under Condition (C5) and the event $\mathcal{E}_1$, we have

$$\|\nabla^2\mathcal{L}_n^1(\gamma) - \nabla^2\mathcal{L}_n^1(\widehat{\theta}_N)\|_2 \leq 2M\|\gamma - \widehat{\theta}_N\|_2 \leq \frac{\rho\lambda_-}{2}.$$

Similarly, under Condition (C5) and the event $\mathcal{E}_0 \bigcap \mathcal{E}_1$, we can show that

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

Therefore, $\left\|\nabla^2\mathcal{S}_N(\gamma) - \nabla^2\mathcal{L}_0\left(\theta^*\right)\right\|_2 \leq \rho\lambda_-$. As $\nabla^2\mathcal{L}_0\left(\theta^*\right) \succeq \lambda_- I$, we obtain $\nabla^2\mathcal{S}_N(\gamma) \succeq \lambda_- I - \rho\lambda_- I = (1-\rho)\lambda_- I$, implying that $\mathcal{S}_N$ is $(1-\rho)\lambda_-$-strongly convex on the ball U.

Hence, to prove the result, it suffices to show that $\|\nabla\mathcal{S}_N(\widehat{\theta}_N)\|_2 \leq \frac{(1-\rho)\lambda_-\delta_\rho}{2}$, and the rest of the derivations follow straightforwardly from Lemma 6 of Zhang et al. (2013). Note that $\nabla\mathcal{S}_N(\widehat{\theta}_N) = \nabla\mathcal{L}_n^1(\widehat{\theta}_N) - \nabla\mathcal{L}_n^1(\widehat{\theta}_0) + \nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0)$. Denote $H_k' = \int_0^1 \nabla^2\mathcal{L}_n^k(\theta^* + t(\widehat{\theta}_N - \theta^*))dt$ and $H_k'' = \int_0^1 \nabla^2\mathcal{L}_n^k(\theta^* + t(\widehat{\theta}_0 - \theta^*))dt$, $k = 1, \ldots, K$. Then we have

$$\begin{aligned}&\|\nabla\mathcal{L}_n^1(\widehat{\theta}_N) - \nabla\mathcal{L}_n^1(\widehat{\theta}_0)\|_2 \\ \leq&\|\nabla\mathcal{L}_n^1(\widehat{\theta}_N) - \nabla\mathcal{L}_n^1\left(\theta^*\right)\|_2 + \|\nabla\mathcal{L}_n^1(\theta^*) - \nabla\mathcal{L}_n^1(\widehat{\theta}_0)\|_2 \\ =&\|H_1'(\widehat{\theta}_N - \theta^*)\|_2 + \|H_1''(\widehat{\theta}_0 - \theta^*)\|_2 \\ \leq&\|\nabla^2\mathcal{L}_n^1(\theta^*) - \nabla^2\mathcal{L}_0\left(\theta^*\right)\|_2\|\widehat{\theta}_N - \theta^*\|_2 + \|\nabla^2\mathcal{L}_n^1(\theta^*) - \nabla^2\mathcal{L}_0\left(\theta^*\right)\|_2\|\widehat{\theta}_0 - \theta^*\|_2 \\ &+ \|\nabla^2\mathcal{L}_0\left(\theta^*\right)\|_2\|\widehat{\theta}_N - \theta^*\|_2 + \|\nabla^2\mathcal{L}_0\left(\theta^*\right)\|_2\|\widehat{\theta}_0 - \theta^*\|_2 \\ &+ 2M\|\widehat{\theta}_N - \theta^*\|_2^2 + 2M\|\widehat{\theta}_0 - \theta^*\|_2^2, \qquad \text{(S.2)}\end{aligned}$$

where the equality in the third line in (S.2) is obtained by using the integral form of Taylor's expansion, and the last inequality in (S.2) holds under Condition (C5) and the event $\mathcal{E}_0 \cap \mathcal{E}_1$. Under Conditions (C3) and (C5), one can readily extend Lemma 6 of Zhang et al. (2013) to the situation where the empirical risk is expressed in the form of U-statistics. Hence under the event $\mathcal{E}_0$,

$$\|\widehat{\theta}_N - \theta^*\|_2 \leq \frac{2\|\nabla\mathcal{L}_N(\theta^*)\|_2}{(1-\rho)\lambda_-} \leq \delta_1(\rho, \lambda_-, \lambda_+, M).$$

Therefore, under the event $\mathcal{E}_0 \cap \mathcal{E}_1$,

$$\left\|\nabla\mathcal{L}_n^1(\widehat{\theta}_N) - \nabla\mathcal{L}_n^1(\widehat{\theta}_0)\right\|_2 \leq \frac{(1-\rho)\lambda_-\delta_\rho}{4}.$$

Similarly, $\nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0) = \nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0) - \nabla\widetilde{\mathcal{L}}_N(\theta^*) + \nabla\widetilde{\mathcal{L}}_N(\theta^*)$. Note that under the event $\bigcap_{k=1}^K \mathcal{E}_k$, we can write

$$\|\nabla\widetilde{\mathcal{L}}_N(\theta^*)\|_2 \leq \frac{1}{K}\sum_{k=1}^K \|\nabla\mathcal{L}_k(\theta^*)\|_2 \leq \frac{(1-\rho)\lambda_-\delta_\rho}{8}.$$

Moreover,

$$\begin{aligned}
\left\|\nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0) - \nabla\widetilde{\mathcal{L}}_N(\theta^*)\right\|_2 &\leq \frac{1}{K}\sum_{k=1}^K \left\|\nabla\mathcal{L}_n^k(\widehat{\theta}_0) - \nabla\mathcal{L}_n^k(\theta^*)\right\|_2 \\
&= \frac{1}{K}\sum_{k=1}^K \|H_k''(\widehat{\theta}_0 - \theta^*)\|_2 \\
&\leq \frac{1}{K}\sum_{k=1}^K \{\|\nabla^2\mathcal{L}_n^k(\theta^*) - \nabla^2\mathcal{L}_0(\theta^*)\|_2\|\widehat{\theta}_0 - \theta^*\|_2 \\
&\quad + \|\nabla^2\mathcal{L}_0(\theta^*)\|_2\|\widehat{\theta}_0 - \theta^*\|_2 + 2M\|\widehat{\theta}_0 - \theta^*\|_2^2\}.
\end{aligned}$$

Now, under the event $\bigcap_{k=1}^K \mathcal{E}_k$,

$$\left\|\nabla\widetilde{\mathcal{L}}_N(\widehat{\theta}_0) - \nabla\widetilde{\mathcal{L}}_N(\theta^*)\right\|_2 \leq \frac{(1-\rho)\lambda_-\delta_\rho}{8}.$$

This proves $\|\nabla\mathcal{S}_N(\widehat{\theta}_N)\|_2 \leq \frac{(1-\rho)\lambda_-\delta_\rho}{2}$ under the event $\bigcap_{k=0}^K \mathcal{E}_k$. □

4 Proofs of Theorems Related to the Naive Estimator

Define the following "good" event:

$$\mathcal{E}_1^* = \left\{M_1 \leq 2M, \left\|\nabla^2\mathcal{L}_n^1(\theta^*) - \nabla^2\mathcal{L}_0(\theta^*)\right\|_2 \leq \frac{\rho\lambda_-}{2}, \left\|\nabla\mathcal{L}_n^1(\theta^*)\right\|_2 \leq \frac{(1-\rho)\lambda_-}{2}\delta_\rho\right\}.$$

In the following, we introduce two lemmas useful for proving Theorem S1.

Lemma S1. *Assume that Conditions (C1)-(C6) hold and there exists $\zeta > 0$ such that $p = O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau > 0$. Then there exists a constant C'' independent of (n, K, N, p) such that*

$$\mathbb{P}(\mathcal{E}_1^*) \geq 1 - C''\frac{p^8}{n^8}.$$

Lemma S2. *Assume that Conditions (C1)-(C6) hold and there exists $\zeta > 0$ such that $p = O(n^{\frac{8}{\tau}(1-\frac{1+\zeta}{8})})$ if $\tau > 0$. Then under the event $\mathcal{E}_1^*$,*

$$\|\widehat{\theta}_n^1 - \theta^*\|_2 \leq \frac{2\|\nabla\mathcal{L}_n^1(\theta^*)\|_2}{(1-\rho)\lambda_-}.$$

Proof of Lemma S1. It can be shown that

$$\mathbb{P}\left((\mathcal{E}_1^*)^c\right) \leq \mathbb{P}\left(M_1 > 2M\right) + \mathbb{P}\left(\left\|\nabla^2 \mathcal{L}_n^1\left(\theta^*\right) - \nabla^2 \mathcal{L}_0\left(\theta^*\right)\right\|_2 > \frac{\rho\lambda_-}{4}\right) + \mathbb{P}\left(\left\|\nabla \mathcal{L}_n^1\left(\theta^*\right)\right\|_2 > \frac{(1-\rho)\lambda_-}{2}\delta_\rho\right).$$

By applying Lemma 1 and using the properties of the higher moments of U-statistics stated in Theorem 1 of Lee (2019, Chapter 1.5), we can show that there exists some constant C'' independent of (n, K, p, N) such that

$$\mathbb{P}\left((\mathcal{E}_1^*)^c\right) \leq C'' \frac{p^8}{n^8}.$$

This completes the proof of the lemma. □

Proof of Lemma S2. Lemma S2 can be obtained along the lines of the proof of Lemma 6 of Zhang et al. (2013), by first proving the strong convexity of $\mathcal{L}_n^1(\theta)$ over the ball $U := \{\theta \in \mathcal{R}^p : \|\theta - \theta^*\| \leq \delta_\rho\}$, followed by using a reduction to absurdity argument to derive the claim. □

Proof of Theorem S1. Since $\bar{\theta}_N = K^{-1}\sum_{k=1}^K \widehat{\theta}_n^k$, we obtain the following bound on the MSE of $\bar{\theta}_N$:

$$\begin{aligned}\mathbb{E}[\|\bar{\theta}_N - \theta^*\|_2^2] &= \frac{1}{K^2}\sum_{k=1}^K \mathbb{E}[\|\widehat{\theta}_n^k - \theta^*\|_2^2] + \frac{1}{K^2}\sum_{k_1 \neq k_2} \mathbb{E}[(\widehat{\theta}_n^{k_1} - \theta^*)'(\widehat{\theta}_n^{k_2} - \theta^*)] \\ &\leq \frac{1}{K}\mathbb{E}[\|\widehat{\theta}_n^1 - \theta^*\|_2^2] + \|\mathbb{E}(\widehat{\theta}_n^1 - \theta^*)\|_2^2.\end{aligned}$$

Below, we derive the bounds for $\mathbb{E}[\|\widehat{\theta}_n^1 - \theta^*\|_2^2]$ and $\|\mathbb{E}(\widehat{\theta}_n^1 - \theta^*)\|_2^2$.

Let us first consider the bound of $\mathbb{E}[\|\widehat{\theta}_n^1 - \theta^*\|_2^2]$. It is straightforward to see that

$$\begin{aligned}0 = \nabla \mathcal{L}_n^1(\widehat{\theta}_n^1) &= \nabla \mathcal{L}_n^1(\theta^*) + \nabla^2 \mathcal{L}_n^1(\theta')(\widehat{\theta}_n^1 - \theta^*) \\ &= \nabla \mathcal{L}_n^1(\theta^*) + (\nabla^2 \mathcal{L}_n^1(\theta') - \nabla^2 \mathcal{L}_n^1(\theta^*))(\widehat{\theta}_n^1 - \theta^*) \\ &\quad + (\nabla^2 \mathcal{L}_n^1(\theta^*) - \nabla^2 \mathcal{L}_0(\theta^*))(\widehat{\theta}_n^1 - \theta^*) + \nabla^2 \mathcal{L}_0(\theta^*)(\widehat{\theta}_n^1 - \theta^*),\end{aligned}$$

where $\theta' = \kappa\theta^* + (1-\kappa)\widehat{\theta}_n^1$ for some $\kappa \in [0,1]$. Multiplying $[\nabla^2 \mathcal{L}_0(\theta^*)]^{-1}$ to both sides of the above equation, we have

$$\begin{aligned}\widehat{\theta}_n^1 - \theta^* = &-[\nabla^2 \mathcal{L}_0(\theta^*)]^{-1}\nabla \mathcal{L}_n^1(\theta^*) + [\nabla^2 \mathcal{L}_0(\theta^*)]^{-1}(\nabla^2 \mathcal{L}_n^1(\theta^*) - \nabla^2 \mathcal{L}_n^1(\theta'))(\widehat{\theta}_n^1 - \theta^*) \\ &+ [\nabla^2 \mathcal{L}_0(\theta^*)]^{-1}(\nabla^2 \mathcal{L}_0(\theta^*) - \nabla^2 \mathcal{L}_n^1(\theta^*))(\widehat{\theta}_n^1 - \theta^*). \qquad \text{(S.3)}\end{aligned}$$

Under Condition (C2), there exists $C_0 > 0$ such that $\|\widehat{\theta}_n^1 - \theta^*\|_2 \leq C_0$. Hence we have

$$\begin{aligned}\mathbb{E}[\|\widehat{\theta}_n^1 - \theta^*\|_2^2] \leq\ & C_0\mathbb{P}((\mathcal{E}_1^*)^c) + \mathbb{E}[\|[\nabla^2 \mathcal{L}_0(\theta^*)]^{-1}\nabla \mathcal{L}_n^1(\theta^*)\|_2^2] \\ &+ 2\|[\nabla^2 \mathcal{L}_0(\theta^*)]^{-1}\|_2^2 \mathbb{E}[\|\nabla^2 \mathcal{L}_n^1(\theta^*) - \nabla^2 \mathcal{L}_n^1(\theta')\|_2^2 \|\widehat{\theta}_n^1 - \theta^*\|_2^2 I(\mathcal{E}_1^*)] \\ &+ 2\|[\nabla^2 \mathcal{L}_0(\theta^*)]^{-1}\|_2^2 \mathbb{E}[\|\nabla^2 \mathcal{L}_0(\theta^*) - \nabla^2 \mathcal{L}_n^1(\theta^*)\|_2^2 \|\widehat{\theta}_n^1 - \theta^*\|_2^2 I(\mathcal{E}_1^*)] \\ &+ 2\|[\nabla^2 \mathcal{L}_0(\theta^*)]^{-1}\|_2 \sqrt{\mathbb{E}[\|\nabla^2 \mathcal{L}_n^1(\theta^*) - \nabla^2 \mathcal{L}_n^1(\theta')\|_2^2 \|\widehat{\theta}_n^1 - \theta^*\|_2^2 I(\mathcal{E}_1^*)]} \\ &\quad \times \sqrt{\mathbb{E}[\|[\nabla^2 \mathcal{L}_0(\theta^*)]^{-1}\nabla \mathcal{L}_n^1(\theta^*)\|_2^2]}\end{aligned}$$

$$+ 2\|[\nabla^2\mathcal{L}_0(\theta^*)]^{-1}\|_2\sqrt{\mathbb{E}[\|\nabla^2\mathcal{L}_0(\theta^*) - \nabla^2\mathcal{L}_n^1(\theta^*)\|_2^2\|\widehat{\theta}_n^1 - \theta^*\|_2^2 I(\mathcal{E}_1^*)]}$$
$$\times \sqrt{\mathbb{E}[\|[\nabla^2\mathcal{L}_0(\theta^*)]^{-1}\nabla\mathcal{L}_n^1(\theta^*)\|_2^2]}. \tag{S.4}$$

Assuming that Condition (C5) holds, it can be shown that under the event $\mathcal{E}_1^*$,

$$\|\nabla^2\mathcal{L}_n^1(\theta^*) - \nabla^2\mathcal{L}_n^1(\theta')\|_2^2 \leq \frac{1}{\mathscr{C}_n^2}\sum_{1\leq i<j\leq n} M^2(Z_{i,1}, Z_{j,1})\|\widehat{\theta}_n^1 - \theta^*\|_2^2.$$

Hence, by the Cauchy-Schwarz inequality and Lemma S2, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla^2\mathcal{L}_n^1(\theta^*) - \nabla^2\mathcal{L}_n^1(\theta')\|_2^2\|\widehat{\theta}_n^1 - \theta^*\|_2^2 I(\mathcal{E}_1^*)] \\
&\leq \frac{1}{\mathscr{C}_n^2}\sum_{1\leq i<j\leq n}\mathbb{E}[M^2(Z_{i,1}, Z_{j,1})\|\widehat{\theta}_n^1 - \theta^*\|_2^4 I(\mathcal{E}_1^*)] \\
&\leq \frac{1}{\mathscr{C}_n^2}\sum_{1\leq i<j\leq n}\sqrt{\mathbb{E}[M^4(Z_{i,1}, Z_{j,1})]\mathbb{E}[\|\widehat{\theta}_n^1 - \theta^*\|_2^8 I(\mathcal{E}_1^*)]} \\
&\leq \frac{16}{(1-\rho)^4\lambda_-^4\mathscr{C}_n^2}\sum_{1\leq i<j\leq n}\sqrt{\mathbb{E}[M^4(Z_{i,1}, Z_{j,1})]\mathbb{E}\|\nabla\mathcal{L}_n^1(\theta^*)\|_2^8}.
\end{aligned}$$

Then by Condition (C5) and Lemma 1, we can write

$$\mathbb{E}[\|\nabla^2\mathcal{L}_n^1(\theta^*) - \nabla^2\mathcal{L}_n^1(\theta')\|_2^2\|\widehat{\theta}_n^1 - \theta^*\|_2^2 I(\mathcal{E}_1^*)] = O(p^2/n^2).$$

Analogously, it can be shown that

$$\mathbb{E}[\|\nabla^2\mathcal{L}_0(\theta^*) - \nabla^2\mathcal{L}_n^1(\theta^*)\|_2^2\|\widehat{\theta}_n^1 - \theta^*\|_2^2 I(\mathcal{E}_1^*)] = O(p\log p/n^2).$$

Note that Condition (C3) implies $\|[\nabla^2\mathcal{L}_0(\theta^*)]^{-1}\|_2 \leq \lambda_-$, and from Lemma S1, we have $\mathbb{P}((\mathcal{E}_1^*)^c) = O(p^8/n^8)$. Furthermore, by generalizing the properties of U-estimation (Bose and Chatterjee, 2018) to allow the dimension p to diverge, we obtain

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

Using the above results in (S.4), we obtain

$$\mathbb{E}[\|\widehat{\theta}_n^1 - \theta^*\|_2^2] \leq \frac{4A}{n} + C\left(\eta_{n,K,p}^2 + \frac{\eta_{n,K,p}}{\sqrt{n}}\sqrt{A} + \frac{p\log n\log\log n}{n^2}\right), \tag{S.5}$$

where $A = \mathbb{E}[\|[\nabla^2\mathcal{L}_0(\theta^*)]^{-1}g(\theta^*; Z)\|_2^2]$ and $\eta_{n,K,p}^2 = p^2/n^2 + p^8/n^8$.

Next, we derive the upper bound of $\|\mathbb{E}(\widehat{\theta}_n^1 - \theta^*)\|_2^2$. Let $Q_1 = [\nabla^2\mathcal{L}_0(\theta^*)]^{-1}\nabla\mathcal{L}_n^1(\theta^*)$, $Q_2 = [\nabla^2\mathcal{L}_0(\theta^*)]^{-1}(\nabla^2\mathcal{L}_n^1(\theta^*) - \nabla^2\mathcal{L}_n^1(\theta'))$ and $Q_3 = [\nabla^2\mathcal{L}_0(\theta^*)]^{-1}(\nabla^2\mathcal{L}_0(\theta^*) - \nabla^2\mathcal{L}_n^1(\theta^*))$. From the expansion in (S.3), we have

$$\begin{aligned}
\widehat{\theta}_n^1 - \theta^* &= -Q_1 + (Q_2 + Q_3)(\widehat{\theta}_n^1 - \theta^*) \\
&= -Q_1 + (Q_2 + Q_3)(-Q_1 + (Q_2 + Q_3)(\widehat{\theta}_n^1 - \theta^*)) \\
&= -Q_1 + (Q_2 + Q_3)^2(\widehat{\theta}_n^1 - \theta^*) - (Q_2 + Q_3)Q_1.
\end{aligned}$$

Denote $\mathbb{E}^*(\xi) = \mathbb{E}[\xi I(\mathcal{E}_1^*)]$ for any random vector ξ. Note that $\mathbb{E}Q_1 = 0$. By Lemma S1,

$$\begin{aligned}
\|\mathbb{E}(\widehat{\theta}_n^1 - \theta^*)\|_2^2 &\le 2\|\mathbb{E}^*[(Q_2+Q_3)^2(\widehat{\theta}_n^1 - \theta^*)]\|_2^2 + 2\|\mathbb{E}^*[(Q_2+Q_3)Q_1]\|_2^2 + C_0^2\mathbb{P}((\mathcal{E}_1^*)^c) \\
&\le 2\mathbb{E}^*[\|Q_2+Q_3\|_2^4]\mathbb{E}^*[\|\widehat{\theta}_n^1 - \theta^*\|_2^2] + 2\mathbb{E}^*[\|Q_2+Q_3\|_2^2]\mathbb{E}[\|Q_1\|_2^2] + O(p^8/n^8) \\
&\le 16(\mathbb{E}^*[\|Q_2\|_2^4] + \mathbb{E}^*[\|Q_3\|_2^4])\mathbb{E}^*[\|\widehat{\theta}_n^1 - \theta^*\|_2^2] \\
&\quad + 4(\mathbb{E}^*[\|Q_2\|_2^2] + \mathbb{E}^*[\|Q_3\|_2^2])\mathbb{E}^*[\|Q_1\|_2^2] + O(p^8/n^8). \qquad \text{(S.6)}
\end{aligned}$$

Under Conditions (C3) and (C5), for $\nu = 1$ or 2,

$$\begin{aligned}
\mathbb{E}^*[\|Q_2\|_2^{2\nu}] &\le \lambda_-^{-2\nu}\mathbb{E}^*\|\nabla^2\mathcal{L}_n^1(\theta^*) - \nabla^2\mathcal{L}_n^1(\theta')\|_2^{2\nu} \\
&\le \frac{\lambda_-^{-2\nu}}{\mathscr{C}_n^2}\sum_{1\le i<j\le n}\mathbb{E}^*[M^{2\nu}(Z_{i,1},Z_{j,1})\|\widehat{\theta}_n^1 - \theta^*\|_2^{2\nu}] \\
&\le \frac{\lambda_-^{-2\nu}}{\mathscr{C}_n^2}\sum_{1\le i<j\le n}\sqrt{\mathbb{E}[M^{4\nu}(Z_{i,1},Z_{j,1})]\mathbb{E}^*[\|\widehat{\theta}_n^1 - \theta^*\|_2^{4\nu}]} \\
&\le \frac{4M^{2\nu}}{(1-\rho)^2\lambda_-^{2\nu+2}}\sqrt{\mathbb{E}[\|\nabla\mathcal{L}_n^1(\theta^*)\|_2^{4\nu}]} \\
&= O(p^\nu/n^\nu),
\end{aligned}$$

where the last equation holds by virtue of Lemma 1, which also guarantees, for $\nu = 1$ or 2, that

$$\mathbb{E}^*[\|Q_3\|_2^{2\nu}] \le \lambda_-^{-2\nu}\mathbb{E}\|\nabla^2\mathcal{L}_0(\theta^*) - \nabla^2\mathcal{L}_n^1(\theta^*)\|_2^{2\nu} = O((\log p)^\nu/n^\nu).$$

It can be shown from (S.5) that

$$\mathbb{E}[\|\widehat{\theta}_n^1 - \theta^*\|_2^2] = O\left(\frac{A}{n} + \eta_{n,K,p}^2\right).$$

Plugging the above results into (S.6), we obtain

$$\|\mathbb{E}(\widehat{\theta}_n^1 - \theta^*)\|_2^2 \le C\left(\frac{p}{n^2} + \frac{p^2}{n^3}\right)A + C\frac{p^2\eta_{n,K,p}^2}{n^2}.$$

Combining the bounds of $\mathbb{E}\|\widehat{\theta}_n^1 - \theta^*\|_2^2$ and $\|\mathbb{E}(\widehat{\theta}_n^1 - \theta^*)\|_2^2$ leads to

$$\begin{aligned}
\mathbb{E}[\|\bar{\theta}_N - \theta^*\|_2^2] \le \frac{4A}{N} &+ C\left(\frac{p}{n^2} + \frac{p^2}{n^3}\right)A + C\frac{\eta_{n,K,p}}{\sqrt{n}K}\sqrt{A} \\
&+ C\left(\frac{p^2\eta_{n,K,p}^2}{n^2} + \frac{p^8}{n^8K} + \frac{p\log n\log\log n}{nN}\right).
\end{aligned}$$

This completes the proof. □

5 Proofs of Theorems Related to Variance Estimation

Here, we prove the properties on the surrogate U-statistics $\nabla^2\widetilde{\mathcal{L}}_N(\widetilde{\theta}_N)$ and $\widehat{V}_{N,K}(\widetilde{\theta}_N)$ stated in the main body of the paper.

Proof of Theorem 5. Let us first consider those of $\nabla^2\widetilde{\mathcal{L}}_N(\widetilde{\theta}_N)$. Under Condition (C5), it can be shown that

$$\begin{aligned}\mathbb{E}\|\nabla^2\mathcal{L}_n^k(\widetilde{\theta}_N) - \nabla^2\mathcal{L}_0(\theta^*)\|^2 &\le \frac{2}{\mathscr{C}_n^2}\sum_{i<j}\mathbb{E}[M^2(Z_{i,k}, Z_{j,k})\|\widetilde{\theta}_N - \theta^*\|^2] + 2\mathbb{E}\|\nabla^2\mathcal{L}_n^k(\theta^*) - \nabla^2\mathcal{L}_0(\theta^*)\|^2 \\ &\le 2\sqrt{\mathbb{E}[M^4(Z_{i,k}, Z_{j,k})]\mathbb{E}[\|\widetilde{\theta}_N - \theta^*\|^4]} + 2\mathbb{E}\|\nabla^2\mathcal{L}_n^k(\theta^*) - \nabla^2\mathcal{L}_0(\theta^*)\|^2 \\ &\le 2M^2\sqrt{\mathbb{E}[\|\widetilde{\theta}_N - \theta^*\|^4]} + 2\mathbb{E}\|\nabla^2\mathcal{L}_n^k(\theta^*) - \nabla^2\mathcal{L}_0(\theta^*)\|^2.\end{aligned}$$

By Lemma 1, we have $\mathbb{E}\|\nabla^2\mathcal{L}_n^k(\theta^*) - \nabla^2\mathcal{L}_0(\theta^*)\|^2 \le C\log p/n$. Furthermore, similar to the proofs of Theorem 1, we can write

$$\begin{aligned}\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|^4 \le C\sqrt{\mathbb{E}\|\widehat{\theta}_0 - \theta^*\|^8}\max\left\{\frac{p^2}{N^2}, \frac{\log^2 p}{n^2}, \sqrt{\mathbb{E}\|\widehat{\theta}_0 - \theta^*\|^8}\right\} \\ + C\frac{p^2}{N^2}\max\left\{\frac{p^2}{N^2}, \frac{\log^2 p}{n^2}\right\} + C\frac{p^2}{N^2} + C\frac{K(\log p)^8}{n^8}.\end{aligned}$$

If $\mathbb{E}\|\widehat{\theta}_0 - \theta^*\|^8 = O(\alpha_{n,K,p}^8)$, where $\alpha_{n,K,p} = \sqrt{1/K}$ or $\sqrt{p/n}$, then

$$\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|^4 \le C\alpha_{n,K,p}^8 + C\frac{p^2}{N^2} + C\frac{K(\log p)^8}{n^8}.$$

Hence we have

$$\mathbb{E}\|\nabla^2\mathcal{L}_n^k(\widetilde{\theta}_N) - \nabla^2\mathcal{L}_0(\theta^*)\|^2 = O\left(\alpha_{n,K,p}^4 + \frac{p}{N} + \frac{\sqrt{K}(\log p)^4}{n^4} + \frac{\log p}{n}\right),$$

which yields

$$\mathbb{E}\|\nabla^2\widetilde{\mathcal{L}}_N(\widetilde{\theta}_N) - \nabla^2\mathcal{L}_0(\theta^*)\|^2 = O\left(\frac{\alpha_{n,K,p}^4}{K} + \frac{p}{KN} + \frac{(\log p)^4}{n^4\sqrt{K}} + \frac{\log p}{N}\right).$$

When $\alpha_{n,K,p} = \sqrt{1/K}$ or $\sqrt{p/n}$, the statistic $\nabla^2\widetilde{\mathcal{L}}_N(\widetilde{\theta}_N)$ is a consistent estimator of $\nabla^2\mathcal{L}_0(\theta^*)$.

Next, we turn to the properties of $\widehat{V}_{N,K}(\widetilde{\theta}_N)$. Let

$$\widetilde{V}_{n,k}(\widetilde{\theta}_N, \theta^*) = \frac{1}{n^3}\sum_i\sum_{j\neq i}\sum_{l\neq i,j}\nabla\ell(\widetilde{\theta}_N; Z_{i,k}, Z_{j,k})\nabla\ell(\theta^*; Z_{i,k}, Z_{l,k})^\top.$$

Then we have

$$\begin{aligned}\mathbb{E}[\|\widehat{V}_{n,k}(\widetilde{\theta}_N) - V_0\|^2] &\le 2\mathbb{E}[\|\widehat{V}_{n,k}(\widetilde{\theta}_N) - \widetilde{V}_{n,k}(\widetilde{\theta}_N, \theta^*)\|^2] + 2\mathbb{E}[\|\widetilde{V}_{n,k}(\widetilde{\theta}_N, \theta^*) - V_0\|^2] \\ &=: 2(\mathrm{I}_{n,k} + \mathrm{II}_{n,k}).\end{aligned}$$

To derive the upper bound of $\mathrm{I}_{n,k}$, note that by Lemmas 2 and 3 and the Jensen's inequality, we have, under Condition (C5),

$$\mathrm{I}_{n,k} = \mathbb{E}\left\|\frac{1}{n^3}\sum_i\sum_{j\neq i}\sum_{l\neq i,j}\nabla\ell(\widetilde{\theta}_N; Z_{i,k}, Z_{j,k})\{\nabla\ell(\widetilde{\theta}_N; Z_{i,k}, Z_{l,k}) - \nabla\ell(\theta^*; Z_{i,k}, Z_{l,k})\}^T\right\|^2$$

$$
\begin{aligned}
&\leq \mathbb{E}\left\|\nabla\ell(\widetilde{\theta}_N; Z_{i,k}, Z_{j,k})\{\nabla\ell(\widetilde{\theta}_N; Z_{i,k}, Z_{l,k}) - \nabla\ell(\theta^*; Z_{i,k}, Z_{l,k})\}^T\right\|^2 \\
&\leq \mathbb{E}\|\nabla\ell(\widetilde{\theta}_N; Z_{i,k}, Z_{j,k})\|^2\|\nabla\ell(\widetilde{\theta}_N; Z_{i,k}, Z_{l,k}) - \nabla\ell(\theta^*; Z_{i,k}, Z_{l,k})\|^2 \\
&\leq 2\mathbb{E}\|\nabla\ell(\widetilde{\theta}_N; Z_{i,k}, Z_{j,k}) - \nabla\ell(\theta^*; Z_{i,k}, Z_{j,k})\|^2\|\nabla\ell(\widetilde{\theta}_N; Z_{i,k}, Z_{l,k}) - \nabla\ell(\theta^*; Z_{i,k}, Z_{l,k})\|^2 \\
&\quad + 2\mathbb{E}\|\nabla\ell(\theta^*; Z_{i,k}, Z_{j,k})\|^2\|\nabla\ell(\widetilde{\theta}_N; Z_{i,k}, Z_{l,k}) - \nabla\ell(\theta^*; Z_{i,k}, Z_{l,k})\|^2 \\
&\leq C\mathbb{E}\|\nabla\ell(\theta^*; Z_{i,k}, Z_{j,k})\|^2\|\nabla\ell(\widetilde{\theta}_N; Z_{i,k}, Z_{l,k}) - \nabla\ell(\theta^*; Z_{i,k}, Z_{l,k})\|^2 \\
&\leq CpG^2\mathbb{E}\|\nabla\ell(\widetilde{\theta}_N; Z_{i,k}, Z_{l,k}) - \nabla\ell(\theta^*; Z_{i,k}, Z_{l,k})\|^2 \\
&= CpG^2\mathbb{E}\left\|\int_0^1 \nabla^2\ell(\theta^* + t(\widetilde{\theta}_N - \theta^*); Z_{i,k}, Z_{l,k})dt(\widetilde{\theta}_N - \theta^*)\right\|^2 \\
&\leq CpG^2\mathbb{E}[M^2(Z_{i,k}, Z_{l,k})\|\widetilde{\theta}_N - \theta^*\|^4] + CpG^2\mathbb{E}\left\|\nabla^2\ell(\theta^*; Z_{i,k}, Z_{l,k})(\widetilde{\theta}_N - \theta^*)\right\|^2 + C\frac{K(\log p)^8}{n^8} \\
&\leq CpG^2M^2\sqrt{\mathbb{E}[\|\widetilde{\theta}_N - \theta^*\|^8]} + CpG^2H^2\sqrt{\mathbb{E}[\|\widetilde{\theta}_N - \theta^*\|^4]} + C\frac{K(\log p)^8}{n^8}.
\end{aligned}
$$

Analogous to the proof of Theorem 1, it can be shown that

$$
\begin{aligned}
\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|^8 \leq C\sqrt{\mathbb{E}\|\widehat{\theta}_0 - \theta^*\|^{16}}\max\left\{\frac{p^4}{N^4}, \frac{\log^4 p}{n^4}, \sqrt{\mathbb{E}\|\widehat{\theta}_0 - \theta^*\|^{16}}\right\} \\
+ C\frac{p^4}{N^4}\max\left\{\frac{p^4}{N^4}, \frac{\log^4 p}{n^4}\right\} + C\frac{p^4}{N^4} + C\frac{K(\log p)^8}{n^8}.
\end{aligned}
$$

If $\mathbb{E}\|\widehat{\theta}_0 - \theta^*\|^{16} = O(\alpha_{n,K,p}^{16})$, where $\alpha_{n,K,p} = \sqrt{1/K}$ or $\sqrt{p/n}$, then

$$
\mathbb{E}\|\widetilde{\theta}_N - \theta^*\|^8 \leq C\alpha_{n,K,p}^{16} + C\frac{p^4}{N^4} + C\frac{K(\log p)^8}{n^8}.
$$

Hence we have

$$
\mathrm{I}_{n,k} = O\left(\alpha_{n,K,p}^4 + \alpha_{n,K,p}^8 + \frac{p}{N} + \frac{\sqrt{K}(\log p)^4}{n^4} + \frac{K(\log p)^8}{n^8}\right).
$$

Analogously, we can show that $\mathrm{II}_{n,k}$ has the same rate of convergence as $\mathrm{I}_{n,k}$. Hence

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

In particular, when $\alpha_{n,K,p} = \sqrt{1/K}$ or $\sqrt{p/n}$, the statistic $\widehat{V}_{n,k}(\widetilde{\theta}_N)$ is a consistent estimator of $\nabla^2\mathcal{L}_0(\theta^*)$. This completes the proof. □

6 Proofs of Theorems Related to the Iterative Algorithm

Proof of Theorem 6. From Proposition 2, as $n \to \infty$, the following inequality holds with probability approaching unity:

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

$$
\begin{aligned}
&+ C_1 \left(\|\widehat{\theta}_0 - \theta^*\|_2 + \left\| \nabla^2 \mathcal{L}_N(\theta^*) - \nabla^2 \widetilde{\mathcal{L}}_N(\theta^*) \right\|_2 \right) \|\widehat{\theta}_0 - \theta^*\|_2 \\
&+ \left\| \nabla \mathcal{L}_N(\theta^*) - \nabla \widetilde{\mathcal{L}}_N(\theta^*) \right\|_2 \\
=:& R_1 + R_2 + R_3,
\end{aligned}
$$

where

$$
\begin{aligned}
R_1 &= O_{\mathbb{P}}(\alpha_{n,K,p} + (p/N)^{1/2} + (\log p/n)^{1/2}) \|\widehat{\theta}_0 - \widehat{\theta}_N\|_2, \\
R_2 &= O_{\mathbb{P}}(\alpha_{n,K,p} + \sqrt{\log p/N}) \|\widehat{\theta}_0 - \widehat{\theta}_N\|_2 + O_{\mathbb{P}}(\sqrt{p/N} \max\{\alpha_{n,K,p}, \sqrt{\log p/N}\}) \quad \text{and} \\
R_3 &= O_{\mathbb{P}}(\sqrt{p/(nN)}).
\end{aligned}
$$

Therefore,

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

Repeating this process, we obtain

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

This completes the proof. □

7 Details on Running Time Computation

In this section, we provide details on the method of calculating running times in the simulation studies in Section 7.

- Naive estimator:

  (i) For the s-th simulation sample, record the running time for computing $\widehat{\theta}_n^k = \arg\min_\theta \mathcal{L}_n^k(\theta)$ on the k-th machine, denoted by $\mathscr{T}_{s,k}^0$.

  (ii) The time for computing the naive estimator is given by

$$
\mathscr{T}_{\text{naive}} = \frac{1}{SK} \sum_{s=1}^{S} \sum_{k=1}^{K} \mathscr{T}_{s,k}^0.
$$

- SU-ERM and OS-ERM estimators:

  (i) For the s-th simulation sample, record the time for computing the initial estimator, denoted by $\mathscr{T}_s^{\text{initial}}$. If the naive estimator is taken to be the initial estimator, then

$$
\mathscr{T}_s^{\text{initial}} = \frac{1}{K} \sum_{k=1}^{K} \mathscr{T}_{s,k}^0.
$$

  (ii) On the k-th machine, record the running time for computing $\nabla \mathcal{L}_n^k(\widehat{\theta}_0)$, denoted by $\mathscr{T}_{s,k}^1$,

and obtain the average of $\mathscr{T}_{s,k}^1$'s across the K machines:

$$\mathscr{T}_s^{\text{grad}} = \frac{1}{K}\sum_{k=1}^{K} \mathscr{T}_{s,k}^1.$$

(iii) On the first machine, record the time required for solving the optimization problem $\widetilde{\theta}_N = \arg\min_\theta \mathcal{S}_N(\theta)$ (in the case of the SU-ERM estimator) or executing the equation $\widetilde{\theta}_N^Q = \widehat{\theta}_0 - \nabla^2 \mathcal{L}_n^1(\widehat{\theta}_0)^{-1} \nabla \widetilde{\mathcal{L}}_N(\widehat{\theta}_0)$ (in the case of the OS-ERM estimator). Denote the time required as $\mathscr{T}_s^{\text{agg}}$.

(iv) The time required for computing the SU-ERM estimator or the OS-ERM estimators is

$$\mathscr{T}_{\text{SU/OS}} = \frac{1}{S}\sum_{s=1}^{S} (\mathscr{T}_s^{\text{initial}} + \mathscr{T}_s^{\text{grad}} + \mathscr{T}_s^{\text{agg}}).$$

References

Bose, A. and S. Chatterjee (2018). *U-Statistics, M_m-Estimators and Resampling*. Springer: Singapore.

Burkholder, D. L. (1973). Distribution function inequalities for martingales. *The Annals of Probability 1*, 19–42.

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

de Acosta, A. (1981). Inequalities for B-valued random vectors with applications to the strong law of large numbers. *The Annals of Probability 9*, 157–161.

de la Peña, V. H. (1992). Decoupling and Khintchine's inequalities for U-statistics. *The Annals of Probability*, 1877–1892.

Lee, A. J. (2019). *U-Statistics: Theory and Practice*. Routledge: Boca Raton.

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