A. Additional Simulation Results

A.1. Simultaneous Confidence Interval

Figures A.1 and A.2 display the empirical coverage probability and the average width for the linear regression and logistic regression models under Toeplitz design with $d = 2^3$ and $d = 2^5$. Figures A.3 and A.4 display the empirical coverage probability and the average width for the linear regression and logistic regression models under equi-correlation design with $d \in \{2^1, 2^3, 2^5, 2^7\}$. See Section 4.1 for the full details on the simulation setup. The observations made in Section 4.1 also apply to all the cases here. Moreover, we see that the results for equi-correlation design are similar to those for Toeplitz design.

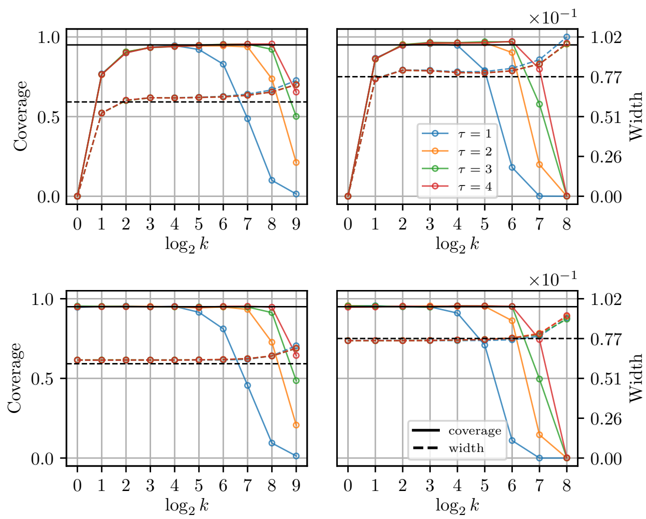

Figure A.1. Empirical coverage probability (**left axis**) and average width (**right axis**) of simultaneous confidence intervals by `k-grad` (**top**) and `n+k-1-grad` (**bottom**) in a linear regression model with Toeplitz design and varying dimension (**left**: $d = 2^3$, **right**: $d = 2^5$). Black solid line represents nominal confidence level (95%) and black dashed line represents oracle width.

A.2. Pointwise Confidence Interval

Figures A.1 and A.2 display the empirical coverage probability and the average width for the linear regression and logistic regression models under Toeplitz design with $d \in \{2^1, 2^3, 2^5, 2^7\}$. The simulation setup is the same as in Section 4.1. All the pointwise confidence intervals are constructed for the second coordinate of θ^*. The algorithm is modified by replacing $\|\cdot\|_\infty$ with $|(\cdot)_2|$ as discussed in Section 2.1. Comparing the results to those in Sections 4.1 and A.1, we see that the performance of `k-grad` and `n+k-1-grad` in constructing pointwise confidence intervals is similar to that in constructing simultaneous confidence intervals. Therefore, the discussions on simultaneous confidence intervals in 4.1 can apply to the cases here.

B. Proofs of Main Results

Proof of Theorem 3.1. By Lemmas F.9 and F.10, we obtain that

$$\left\|\widetilde{\theta} - \widehat{\theta}\right\|_\infty = \left\|\widetilde{\theta}^{(\tau)} - \widehat{\theta}\right\|_\infty \le \left\|\widetilde{\theta}^{(\tau)} - \widehat{\theta}\right\|_2 = O_P\left(\left(\sqrt{\frac{d}{n}}\right)^{\tau+1}\sqrt{\log d}\right), \quad \text{and}$$

$$\left\|\bar{\theta} - \theta^*\right\|_1 = \left\|\widetilde{\theta}^{(\tau-1)} - \theta^*\right\|_1 \le \sqrt{d}\left\|\widetilde{\theta}^{(\tau-1)} - \widehat{\theta}\right\|_2 + \sqrt{d}\left\|\widehat{\theta} - \theta^*\right\|_2 = O_P\left(d\sqrt{\frac{\log d}{N}} + \left(\sqrt{\frac{d}{n}}\right)^{\tau}\sqrt{d\log d}\right),$$

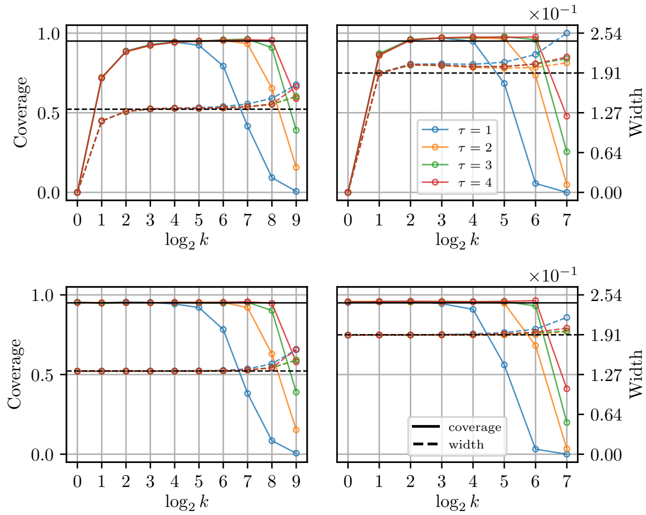

Figure A.2. Empirical coverage probability (**left axis**) and average width (**right axis**) of simultaneous confidence intervals by `k-grad` (**top**) and `n+k-1-grad` (**bottom**) in a logistic regression model with Toeplitz design and varying dimension (**left**: $d = 2^3$, **right**: $d = 2^5$). Black solid line represents nominal confidence level (95%) and black dashed line represents oracle width.

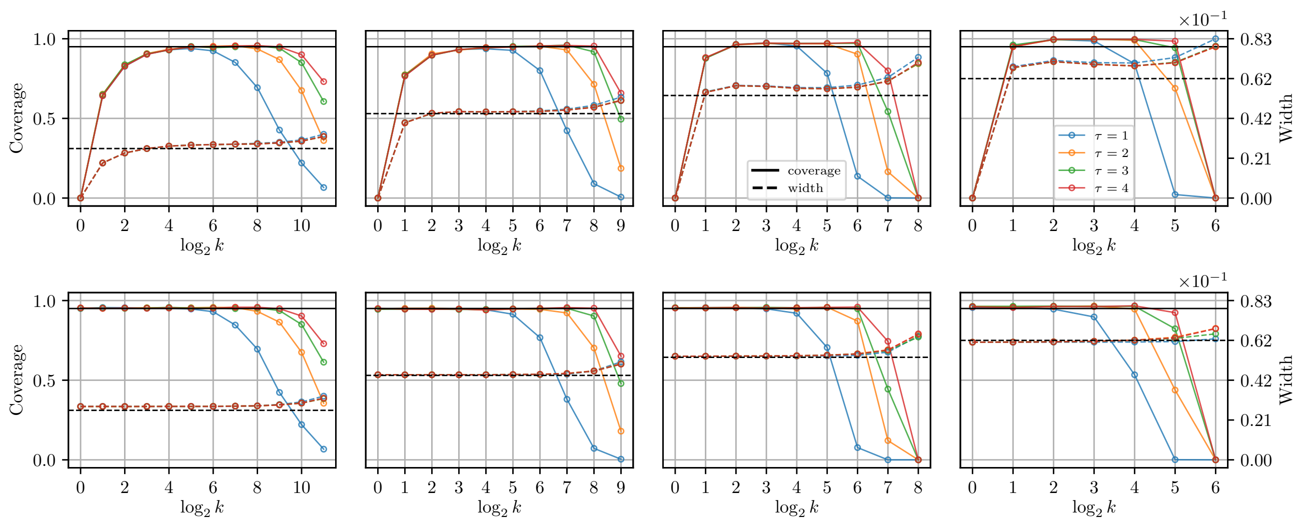

Figure A.3. Empirical coverage probability (**left axis**) and average width (**right axis**) of simultaneous confidence intervals by `k-grad` (**top**) and `n+k-1-grad` (**bottom**) in a linear regression model with equi-correlation design and varying dimension (from **left** to **right**: $d = 2^1, 2^3, 2^5, 2^7$). Black solid line represents nominal confidence level (95%) and black dashed line represents oracle width.

if $N \gtrsim d \log d$ and $n \gtrsim d$. Then, by Lemma C.1, we have $\sup_{\alpha \in (0,1)} |P(T \leq c_{\overline{W}}(\alpha)) - \alpha| = o(1)$ and $\sup_{\alpha \in (0,1)} \left| P(\widehat{T} \leq c_{\overline{W}}(\alpha)) - \alpha \right| = o(1)$, as long as $n \gg d \log^{4+\kappa} d$, $k \gg d^2 \log^{5+\kappa} d$, and

$$\left(\sqrt{\frac{d}{n}} \right)^{\tau+1} \sqrt{\log d} \ll \frac{1}{\sqrt{N} \log^{1/2+\kappa} d}, \quad \text{and}$$

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

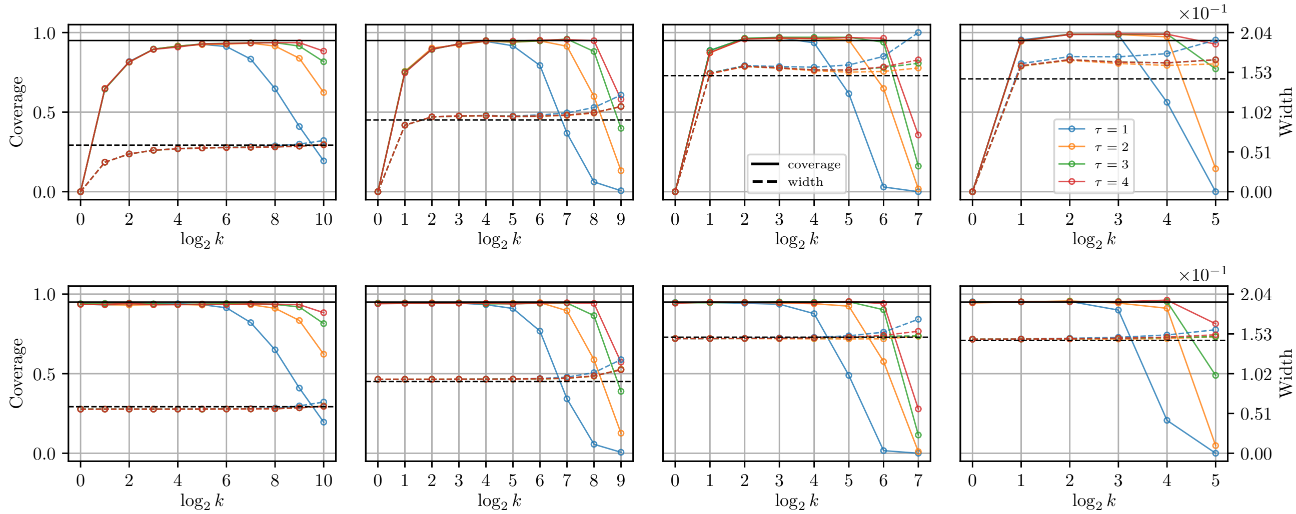

Figure A.4. Empirical coverage probability (**left axis**) and average width (**right axis**) of simultaneous confidence intervals by `k-grad` (**top**) and `n+k-1-grad` (**bottom**) in a logistic regression model with equi-correlation design and varying dimension (from **left** to **right**: $d = 2^1, 2^3, 2^5, 2^7$). Black solid line represents nominal confidence level (95%) and black dashed line represents oracle width.

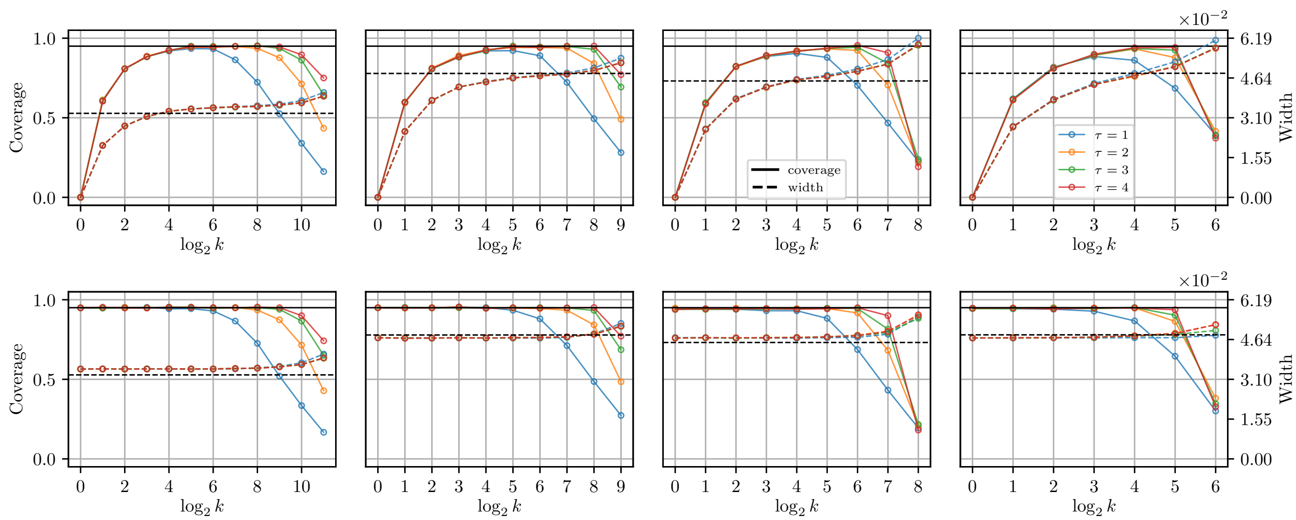

Figure A.5. Empirical coverage probability (**left axis**) and average width (**right axis**) of pointwise confidence intervals by `k-grad` (**top**) and `n+k-1-grad` (**bottom**) in a linear regression model with Toeplitz design and varying dimension (from **left** to **right**: $d = 2^1, 2^3, 2^5, 2^7$). Black solid line represents nominal confidence level (95%) and black dashed line represents oracle width.

We complete the proof by solving these inequalities for τ. $\square$

Proof of Theorem 3.2. By the argument in the proof of Theorem 3.1 with applying Lemma C.2, we have $\sup_{\alpha\in(0,1)} \left|P(T \leq c_{\widetilde{W}}(\alpha)) - \alpha\right| = o(1)$ and $\sup_{\alpha\in(0,1)} \left|P(\widehat{T} \leq c_{\widetilde{W}}(\alpha)) - \alpha\right| = o(1)$, as long as $n \gg d \log^{4+\kappa} d$, $n + k \gg d^2 \log^{5+\kappa} d$, and

$$\left(\sqrt{\frac{d}{n}}\right)^{\tau+1} \sqrt{\log d} \ll \frac{1}{\sqrt{N} \log^{1/2+\kappa} d}, \quad \text{and}$$

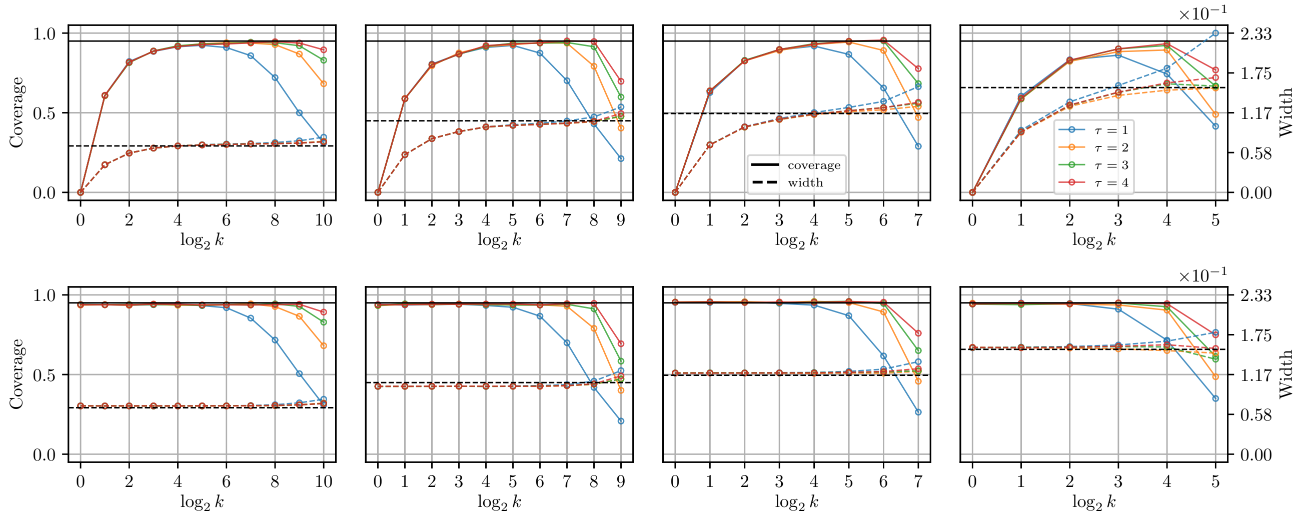

Figure A.6. Empirical coverage probability (**left axis**) and average width (**right axis**) of pointwise confidence intervals by `k-grad` (**top**) and `n+k-1-grad` (**bottom**) in a logistic regression model with Toeplitz design and varying dimension (from **left** to **right**: $d = 2^1, 2^3, 2^5, 2^7$). Black solid line represents nominal confidence level (95%) and black dashed line represents oracle width.

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

We complete the proof by solving these inequalities for τ. □

Proof of Theorem 3.6. By Lemmas F.11 and F.12, we obtain that

$$\left\|\widetilde{\theta} - \widehat{\theta}\right\|_\infty = \left\|\widetilde{\theta}^{(\tau)} - \widehat{\theta}\right\|_\infty \leq \left\|\widetilde{\theta}^{(\tau)} - \widehat{\theta}\right\|_2 = \begin{cases} O_P\left(\frac{1}{d^{3/2}}\left(d^2\sqrt{\frac{\log d}{n}}\right)^{2^\tau}\right), & \tau \leq \tau_0, \\ O_P\left(\frac{1}{d^{3/2}}\left(d^2\sqrt{\frac{\log d}{n}}\right)^{2^{\tau_0}}\left(\sqrt{\frac{d\log d}{n}}\right)^{\tau-\tau_0}\right), & \tau > \tau_0, \end{cases} \quad \text{and} \tag{B.1}$$

$$\begin{aligned} \left\|\bar{\theta} - \theta^*\right\|_1 &= \left\|\widetilde{\theta}^{(\tau-1)} - \theta^*\right\|_1 \leq \sqrt{d}\left\|\widetilde{\theta}^{(\tau-1)} - \theta^*\right\|_2 \leq \sqrt{d}\left\|\widetilde{\theta}^{(\tau-1)} - \widehat{\theta}\right\|_2 + \sqrt{d}\left\|\widehat{\theta} - \theta^*\right\|_2 \\ &= \begin{cases} O_P\left(d\sqrt{\frac{\log d}{N}} + \frac{1}{d}\left(d^2\sqrt{\frac{\log d}{n}}\right)^{2^{\tau-1}}\right), & \tau \leq \tau_0 + 1, \\ O_P\left(d\sqrt{\frac{\log d}{N}} + \frac{1}{d}\left(d^2\sqrt{\frac{\log d}{n}}\right)^{2^{\tau_0}}\left(\sqrt{\frac{d\log d}{n}}\right)^{\tau-\tau_0-1}\right), & \tau > \tau_0 + 1, \end{cases} \end{aligned} \tag{B.2}$$

if $n \gtrsim d^4 \log d$, where τ_0 is the smallest integer t such that

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

that is,

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

Then, by Lemma C.3, we have $\sup_{\alpha\in(0,1)}|P(T\le c_{\overline{W}}(\alpha))-\alpha|=o(1)$ and $\sup_{\alpha\in(0,1)}\left|P(\widehat{T}\le c_{\overline{W}}(\alpha))-\alpha\right|=o(1)$, as long as $n\gg d^4\log d$, $k\gg d^2\log^{5+\kappa}d$, $nk\gg d^5\log^{3+\kappa}d$,

$$\text{RHS of (B.1)}\ll\frac{1}{\sqrt{N}\log^{1/2+\kappa}d},\quad\text{and}\quad\text{RHS of (B.2)}\ll\frac{1}{\sqrt{nd}\log^{1+\kappa}d}.$$

We complete the proof by solving these inequalities for τ.

□

Proof of Theorem 3.7. By the argument in the proof of Theorem 3.1 with applying Lemma C.4, we have $\sup_{\alpha\in(0,1)}\left|P(T\le c_{\widetilde{W}}(\alpha))-\alpha\right|=o(1)$ and $\sup_{\alpha\in(0,1)}\left|P(\widehat{T}\le c_{\widetilde{W}}(\alpha))-\alpha\right|=o(1)$, as long as $n\gg d^4\log d$, $n+k\gg d^2\log^{5+\kappa}d$, $nk\gg d^5\log^{3+\kappa}d$,

$$\text{RHS of (B.1)}\ll\frac{1}{\sqrt{N}\log^{1/2+\kappa}d},\quad\text{and}\quad\text{RHS of (B.2)}\ll\min\left\{\frac{1}{d\log^{11/4+\kappa}d},\frac{1}{\sqrt{d}\log^{1+\kappa}d}\sqrt{\frac{1}{n}+\frac{1}{k}}\right\}.$$

We complete the proof by solving these inequalities for τ.

□

C. Lemmas on Bounding Bootstrap Errors

Lemma C.1 (k-grad)**.** *In linear model, under Assumptions (A1) and (A2), if $n\gg d\log^{4+\kappa}d$, $k\gg d^2\log^{5+\kappa}d$,*

$$\left\|\widetilde{\theta}-\widehat{\theta}\right\|_\infty\ll\frac{1}{\sqrt{N}\log^{1/2+\kappa}d},\quad\textit{and}\quad\left\|\bar{\theta}-\theta^*\right\|_1\ll\min\left\{\frac{1}{d\sqrt{\log k}\log^{2+\kappa}d},\frac{1}{\sqrt{nd}\log^{1+\kappa}d}\right\},$$

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

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

$$\sup_{\alpha\in(0,1)}\left|P(\widehat{T}\le c_{\overline{W}}(\alpha))-\alpha\right|=o(1).\tag{C.2}$$

Proof of Lemma C.1. As noted by (Zhang & Cheng, 2017), since $\|\sqrt{N}(\widetilde{\theta}-\theta^*)\|_\infty=\max_l\sqrt{N}|\widetilde{\theta}_l-\theta^*_l|=\sqrt{N}\max_l\left((\widetilde{\theta}_l-\theta^*_l)\vee(\theta^*_l-\widetilde{\theta}_l)\right)$, the arguments for the bootstrap consistency result with

$$T=\max_l\sqrt{N}(\widetilde{\theta}-\theta^*)_l\quad\text{and}\tag{C.3}$$

$$\widehat{T}=\max_l\sqrt{N}(\widehat{\theta}-\theta^*)_l\tag{C.4}$$

imply the bootstrap consistency result for $T=\|\sqrt{N}(\widetilde{\theta}-\theta^*)\|_\infty$ and $\widehat{T}=\|\sqrt{N}(\widehat{\theta}-\theta^*)\|_\infty$. Hence, from now on, we redefine T and $\widehat{T}$ as (C.3) and (C.4). Define an oracle multiplier bootstrap statistic as

$$W^*:=\max_{1\le l\le d}-\frac{1}{\sqrt{N}}\sum_{i=1}^n\sum_{j=1}^k\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*;Z_{ij})\right)_l\epsilon^*_{ij},\tag{C.5}$$

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

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

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

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

$$\min_l \mathbb{E}\left[\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*;Z)\right)_l^2\right] = \sigma^2\min_l \Theta_{l,l} \geq \sigma^2\lambda_{\min}(\Theta) = \frac{\sigma^2}{\lambda_{\max}(\Sigma)}, \tag{C.6}$$

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

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

Since $\Theta w/\|\Theta w\|_2 \in S^{d-1}$, we have that $(\Theta w/\|\Theta w\|_2)\,x$ is sub-Gaussian with $O(1)$ ψ_2-norm, and hence, $w^\top\Theta x$ is sub-Gaussian with $O(\|\Theta w\|_2) = O(\lambda_{\max}(\Theta)) = O(\lambda_{\min}(\Sigma)^{-1}) = O(1)$ ψ_2-norm, under Assumption (A1). Since e is also sub-Gaussian under Assumption (A2) and is independent of $w^\top\Theta x$, we have that $w^\top\Theta x e$ is sub-exponential with uniformly bounded ψ_1-norm for all $w \in S^{d-1}$, and also, all $\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*;Z)\right)_l$ are sub-exponential with uniformly bounded ψ_1-norm. Combining this with (C.6), we have verified Assumption (E.1) of (Chernozhukov et al., 2013) for $\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*;Z)$.

Define

$$T_0 := \max_{1\leq l\leq d} -\sqrt{N}\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right)_l, \tag{C.7}$$

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

$$\begin{aligned}\sup_{\alpha\in(0,1)}|P(T \leq c_{W^*}(\alpha)) - \alpha| &\lesssim N^{-c} + v^{1/3}\left(1 \vee \log\frac{d}{v}\right)^{2/3} + P\left(\left\|\widehat{\Omega} - \Omega_0\right\|_{\max} > v\right) \\ &\quad + \zeta\sqrt{1 \vee \log\frac{d}{\zeta}} + P\left(|T - T_0| > \zeta\right),\end{aligned} \tag{C.8}$$

where

$$\begin{aligned}\widehat{\Omega} &:= \mathrm{cov}_\epsilon\left(-\frac{1}{\sqrt{N}}\sum_{i=1}^n\sum_{j=1}^k \nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*;Z_{ij})\epsilon_{ij}^*\right) \\ &= \nabla^2\mathcal{L}^*(\theta^*)^{-1}\left(\frac{1}{N}\sum_{i=1}^n\sum_{j=1}^k \nabla\mathcal{L}(\theta^*;Z_{ij})\nabla\mathcal{L}(\theta^*;Z_{ij})^\top\right)\nabla^2\mathcal{L}^*(\theta^*)^{-1}, \quad \text{and}\end{aligned} \tag{C.9}$$

$$\Omega_0 := \mathrm{cov}\left(-\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*;Z)\right) = \nabla^2\mathcal{L}^*(\theta^*)^{-1}\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\nabla^2\mathcal{L}^*(\theta^*)^{-1}. \tag{C.10}$$

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

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

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

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

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

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

$$\begin{aligned}\overline{\Omega} &:= \operatorname{cov}_\epsilon\left(-\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\widetilde{\Theta}\sqrt{n}\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta})\right)\epsilon_j\right) \\ &= \widetilde{\Theta}\left(\frac{1}{k}\sum_{j=1}^{k} n\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta})\right)\left(\nabla\mathcal{L}_j(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta})\right)^\top\right)\widetilde{\Theta}^\top.\end{aligned} \tag{C.11}$$

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

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

where by (C.8),

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

and then,

$$\begin{aligned}\sup_{\alpha\in(0,1)}|P(T \le c_{\overline{W}}(\alpha)) - \alpha| &\lesssim N^{-c} + v^{1/3}\left(1 \vee \log\frac{d}{v}\right)^{2/3} + P\left(\left\|\widehat{\Omega} - \Omega_0\right\|_{\max} > v\right) \\ &+ \zeta\sqrt{1 \vee \log\frac{d}{\zeta}} + P\left(|T - T_0| > \zeta\right) + u^{1/3}\left(1 \vee \log\frac{d}{u}\right)^{2/3} + P\left(\left\|\overline{\Omega} - \widehat{\Omega}\right\|_{\max} > u\right).\end{aligned} \tag{C.12}$$

Applying Lemmas D.1, E.2, and E.1, we have that there exist some $\zeta, u, v > 0$ such that

$$\zeta\sqrt{1 \vee \log\frac{d}{\zeta}} + P\left(|T - T_0| > \zeta\right) = o(1), \quad \text{and} \tag{C.13}$$

$$u^{1/3}\left(1 \vee \log\frac{d}{u}\right)^{2/3} + P\left(\left\|\overline{\Omega} - \widehat{\Omega}\right\|_{\max} > u\right) = o(1), \quad \text{and} \tag{C.14}$$

$$v^{1/3}\left(1 \vee \log\frac{d}{v}\right)^{2/3} + P\left(\left\|\widehat{\Omega} - \Omega_0\right\|_{\max} > v\right) = o(1), \tag{C.15}$$

and hence, after simplifying the conditions, obtain the first result in the lemma. To obtain the second result, we use Lemma D.2, which yields

$$\xi\sqrt{1 \vee \log\frac{d}{\xi}} + P\left(|\widehat{T} - T_0| > \xi\right) = o(1). \tag{C.16}$$

□

Lemma C.2 (`n+k-1-grad`). *In linear model, under Assumptions (A1) and (A2), if $n \gg d\log^{4+\kappa} d$, $n + k \gg d^2\log^{5+\kappa} d$,*

$$\left\|\widetilde{\theta} - \widehat{\theta}\right\|_\infty \ll \frac{1}{\sqrt{N}\log^{1/2+\kappa} d}, \quad \textit{and} \quad \left\|\bar{\theta} - \theta^*\right\|_1 \ll \min\left\{\frac{1}{d\sqrt{\log((n+k)d)}\log^{2+\kappa} d}, \frac{1}{\sqrt{d}\log^{1+\kappa} d}\sqrt{\frac{1}{n} + \frac{1}{k}}\right\},$$

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

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

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

Proof of Lemma C.2. By the argument in the proof of Lemma C.1, we have that

$$\begin{aligned}\sup_{\alpha\in(0,1)} \left|P(T \le c_{\widetilde{W}}(\alpha)) - \alpha\right| &\lesssim N^{-c} + v^{1/3}\left(1 \vee \log\frac{d}{v}\right)^{2/3} + P\left(\left\|\widehat{\Omega} - \Omega_0\right\|_{\max} > v\right) \\ &+ \zeta\sqrt{1 \vee \log\frac{d}{\zeta}} + P\left(|T - T_0| > \zeta\right) + u^{1/3}\left(1 \vee \log\frac{d}{u}\right)^{2/3} + P\left(\left\|\widetilde{\Omega} - \widehat{\Omega}\right\|_{\max} > u\right),\end{aligned} \tag{C.19}$$

where

$$\begin{aligned}\widetilde{\Omega} &:= \mathrm{cov}_\epsilon\left(-\frac{1}{\sqrt{n+k-1}}\left(\sum_{i=1}^n \widetilde{\Theta}\left(\nabla\mathcal{L}(\bar\theta; Z_{i1}) - \nabla\mathcal{L}_N(\bar\theta)\right)\epsilon_{i1} + \sum_{j=2}^k \widetilde{\Theta}\sqrt{n}\left(\nabla\mathcal{L}_j(\bar\theta) - \nabla\mathcal{L}_N(\bar\theta)\right)\epsilon_j\right)\right) \\ &= \widetilde{\Theta}\frac{1}{n+k-1}\left(\sum_{i=1}^n \left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}_N(\theta)\right)^\top\right. \\ &\quad \left. + \sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)^\top\right)\widetilde{\Theta}^\top,\end{aligned} \tag{C.20}$$

if $N \gtrsim \log^{7+\kappa} d$ for some $\kappa > 0$. Applying Lemmas D.1, E.2, and E.3, we have that there exist some $\zeta, u, v > 0$ such that (C.13),

$$u^{1/3}\left(1 \vee \log\frac{d}{u}\right)^{2/3} + P\left(\left\|\widetilde{\Omega} - \widehat{\Omega}\right\|_{\max} > u\right) = o(1), \tag{C.21}$$

and (C.15) hold, and hence, after simplifying the conditions, obtain the first result in the lemma. To obtain the second result, we use Lemma D.2, which yields (C.16). □

Lemma C.3 (k-grad)**.** *In GLM, under Assumptions (B1)–(B4), if $n \gg d\log^{5+\kappa} d$, $k \gg d^2\log^{5+\kappa} d$, $nk \gg d^5\log^{3+\kappa} d$,*

$$\left\|\widetilde{\theta} - \widehat{\theta}\right\|_\infty \ll \frac{1}{\sqrt{N}\log^{1/2+\kappa} d}, \quad \textit{and} \quad \left\|\bar\theta - \theta^*\right\|_1 \ll \frac{1}{\sqrt{nd}\log^{1+\kappa} d},$$

for some $\kappa > 0$, then we have that (C.1) *and* (C.2) *hold.*

Proof of Lemma C.3. We redefine T and $\widehat{T}$ as (C.3) and (C.4). We define an oracle multiplier bootstrap statistic as in (C.5). Under Assumption (B3),

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

is bounded away from zero. Combining this with Assumption (B4), we have verified Assumption (E.1) of (Chernozhukov et al., 2013) for $\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*; Z)$. Then, we use the same argument as in the proof of Lemma C.1, and obtain (C.12) with

$$\overline{\Omega} := \widetilde{\Theta}(\bar{\theta}) \left(\frac{1}{k} \sum_{j=1}^{k} n \left(\nabla \mathcal{L}_j(\bar{\theta}) - \nabla \mathcal{L}_N(\bar{\theta}) \right) \left(\nabla \mathcal{L}_j(\bar{\theta}) - \nabla \mathcal{L}_N(\bar{\theta}) \right)^\top \right) \widetilde{\Theta}(\bar{\theta})^\top, \tag{C.22}$$

under the condition $\log^7(dN)/N \lesssim N^{-c}$ for some constant $c > 0$, which holds if $N \gtrsim \log^{7+\kappa} d$ for some $\kappa > 0$. Applying Lemmas D.3, E.5, and E.4, we have that there exist some $\zeta, u, v > 0$ such that (C.13), (C.14), and (C.15) hold, and hence, after simplifying the conditions, obtain the first result in the lemma. To obtain the second result, we use Lemma D.4, which yields (C.16). □

Lemma C.4 (`n+k-1-grad`)**.** *In GLM, under Assumptions (B1)–(B4), if $n \gg d \log^{5+\kappa} d$, $n + k \gg d^2 \log^{5+\kappa} d$, $nk \gg d^5 \log^{3+\kappa} d$,*

$$\left\| \widetilde{\theta} - \widehat{\theta} \right\|_\infty \ll \frac{1}{\sqrt{N} \log^{1/2+\kappa} d}, \quad \textit{and}$$

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

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

Proof of Lemma C.4. By the argument in the proof of Lemma C.3, we obtain (C.19) with

$$\begin{aligned} \widetilde{\Omega} := \widetilde{\Theta}(\bar{\theta}) \frac{1}{n+k-1} \Bigg(& \sum_{i=1}^{n} \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta) \right)^\top \\ & + \sum_{j=2}^{k} n \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta) \right)^\top \Bigg) \widetilde{\Theta}(\bar{\theta})^\top, \end{aligned} \tag{C.23}$$

if $N \gtrsim \log^{7+\kappa} d$ for some $\kappa > 0$. Applying Lemmas D.3, E.5, and E.6, we have that there exist some $\zeta, u, v > 0$ such that (C.13), (C.21), and (C.15) hold, and hence, after simplifying the conditions, obtain the first result in the lemma. To obtain the second result, we use Lemma D.4, which yields (C.16). □

D. Lemmas on Bounding Bahadur Representation Errors

For both linear model and GLM, we denote the global design matrix and the local design matrices by $X_N = (X_1^\top, \ldots, X_k^\top)^\top \in \mathbb{R}^{N \times d}$ and $X_j = (x_{1j}, \ldots, x_{nj})^\top \in \mathbb{R}^{n \times d}$ for $j = 1, \ldots, k$. We write each covariate vector as $x_{ij} = (x_{ij,1}, \ldots, x_{ij,d})^\top \in \mathbb{R}^{d \times 1}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$. Also, we denote the global response vector and the local response vectors by $y_N = (y_1^\top, \ldots, y_k^\top)^\top \in \mathbb{R}^{N \times 1}$ and $y_j = (y_{1j}, \ldots, y_{nj}) \in \mathbb{R}^{n \times 1}$ for $j = 1, \ldots, k$. For linear model, we define the global noise vector and the local noise vectors by $e_N = (e_1^\top, \ldots, e_k^\top)^\top \in \mathbb{R}^{N \times 1}$ and $e_j = (e_{1j}, \ldots, e_{nj}) \in \mathbb{R}^{n \times 1}$ for $j = 1, \ldots, k$.

Lemma D.1. *T and T_0 are defined as in* (C.3) *and* (C.7) *respectively. In linear model, under Assumptions (A1) and (A2), provided that $\left\| \widetilde{\theta} - \widehat{\theta} \right\|_\infty = O_P(r_{\widetilde{\theta}})$, we have that*

$$|T - T_0| = O_P \left(r_{\widetilde{\theta}} \sqrt{N} + \frac{d\sqrt{\log d}}{\sqrt{N}} \right).$$

Moreover, if $N \gg d^2 \log^{2+\kappa} d$ and

$$\left\| \widetilde{\theta} - \widehat{\theta} \right\|_\infty \ll \frac{1}{\sqrt{N} \log^{1/2+\kappa} d},$$

for some $\kappa > 0$, then there exists some $\zeta > 0$ such that (C.13) *holds.*

Proof of Lemma D.1. First, we note that

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

Now, we bound $\left\|\widehat{\theta}-\theta^*+\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty$. In linear model, we have that $\widehat{\theta}=\left(X_N^\top X_N\right)^{-1}X_N^\top y_N=\theta^*+\left(X_N^\top X_N\right)^{-1}X_N^\top e_N$, and then,

$$\left\|\widehat{\theta}-\theta^*+\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty=\left\|\left(\frac{X_N^\top X_N}{N}\right)^{-1}\frac{X_N^\top e_N}{N}-\Theta\frac{X_N^\top e_N}{N}\right\|_\infty\le\left\|\left(\frac{X_N^\top X_N}{N}\right)^{-1}-\Theta\right\|_\infty\left\|\frac{X_N^\top e_N}{N}\right\|_\infty.$$

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

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

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

$$P\left(\left\|\frac{X_N^\top e_N}{N}\right\|_\infty>\max_l\sqrt{\Sigma_{l,l}}\sigma\left(\frac{\log\frac{2d}{\delta}}{cN}\vee\sqrt{\frac{\log\frac{2d}{\delta}}{cN}}\right)\right)\le\delta. \tag{D.1}$$

Under Assumption (A1), we have that $\max_l\Sigma_{l,l}\le\|\Sigma\|_{\max}=O(1)$, and then,

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

Using the same argument for obtaining (F.3), we have that

$$\left\|\left(\frac{X_N^\top X_N}{N}\right)^{-1}-\Theta\right\|_\infty\le\sqrt{d}\left\|\left(\frac{X_N^\top X_N}{N}\right)^{-1}-\Theta\right\|_2=O_P\left(\frac{d}{\sqrt{N}}\right),$$

and therefore,

$$\left\|\widehat{\theta}-\theta^*+\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty=O_P\left(\frac{d\sqrt{\log d}}{N}\right).$$

Putting together the preceding bounds leads to the first result in the lemma. Choosing

$$\zeta=\left(r_{\widetilde{\theta}}\sqrt{N}+\frac{d\sqrt{\log d}}{\sqrt{N}}\right)^{1-\kappa},$$

with any $\kappa>0$, we deduce that $P\left(|T-T_0|>\zeta\right)=o(1)$. We also have that

$$\zeta\sqrt{1\vee\log\frac{d}{\zeta}},\quad\text{if}\quad\left(r_{\widetilde{\theta}}\sqrt{N}+\frac{d\sqrt{\log d}}{\sqrt{N}}\right)\log^{1/2+\kappa}d=o(1).$$

We complete the proof by simplifying the conditions. □

Lemma D.2. *$\widehat{T}$ and T_0 are defined as in* (C.4) *and* (C.7) *respectively. In linear model, under Assumptions (A1) and (A2), we have that*

$$|\widehat{T}-T_0|=O_P\left(\frac{d\sqrt{\log d}}{\sqrt{N}}\right).$$

Moreover, if $N\gg d^2\log^{2+\kappa}d$ for some $\kappa>0$, then there exists some $\xi>0$ such that (C.16) *holds.*

Proof of Lemma D.2. By the proof of Lemma D.1, we obtain that

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

Choosing

$$\xi = \left(\frac{d\sqrt{\log d}}{\sqrt{N}}\right)^{1-\kappa},$$

with any $\kappa > 0$, we deduce that $P\left(|\widehat{T} - T_0| > \xi\right) = o(1)$. We also have that

$$\xi\sqrt{1 \vee \log\frac{d}{\xi}}, \quad \text{if} \quad \left(\frac{d\sqrt{\log d}}{\sqrt{N}}\right)\log^{1/2+\kappa} d = o(1),$$

which holds if $N \gg d^2 \log^{2+\kappa} d$.

□

Lemma D.3. *T and T_0 are defined as in (C.3) and (C.7) respectively. In GLM, under Assumptions (B1)–(B3), provided that $\left\|\widetilde{\theta} - \widehat{\theta}\right\|_\infty = O_P(r_{\widetilde{\theta}})$ and $N \gtrsim d^4 \log d$, we have that*

$$|T - T_0| = O_P\left(r_{\widetilde{\theta}}\sqrt{N} + \frac{d^{5/2}\log d}{\sqrt{N}}\right).$$

Moreover, if $N \gg d^5 \log^{3+\kappa} d$ and

$$\left\|\widetilde{\theta} - \widehat{\theta}\right\|_\infty \ll \frac{1}{\sqrt{N}\log^{1/2+\kappa} d},$$

for some $\kappa > 0$, then there exists some $\zeta > 0$ such that (C.13) holds.

Proof of Lemma D.3. First, we note that

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

Now, we bound $\left\|\widehat{\theta} - \theta^* + \nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty$. Note by an expression of remainder of the first order Taylor expansion that

$$\begin{aligned}\left\|\widehat{\theta} - \theta^* + \nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty &= \left\|\widehat{\theta} - \theta^* - \Theta(\nabla\mathcal{L}_N(\widehat{\theta}) - \nabla\mathcal{L}_N(\theta^*))\right\|_\infty \\ &= \left\|\widehat{\theta} - \theta^* - \Theta\int_0^1 \nabla^2\mathcal{L}_N(\theta^* + s(\widehat{\theta} - \theta^*))ds(\widehat{\theta} - \theta^*)\right\|_\infty \\ &= \left\|\Theta\int_0^1\left(\nabla^2\mathcal{L}^*(\theta^*) - \nabla^2\mathcal{L}_N(\theta^* + s(\widehat{\theta} - \theta^*))\right)ds(\widehat{\theta} - \theta^*)\right\|_\infty \\ &\leq \left\|\Theta\right\|_\infty\int_0^1\left\|\nabla^2\mathcal{L}^*(\theta^*) - \nabla^2\mathcal{L}_N(\theta^* + s(\widehat{\theta} - \theta^*))\right\|_{\max} ds\left\|\widehat{\theta} - \theta^*\right\|_1.\end{aligned}$$

Under Assumption (B1), we have by an expression of remainder of the first order Taylor expansion that

$$\left|g''(y_{ij}, x_{ij}^\top(\theta^* + s(\widehat{\theta} - \theta^*))) - g''(y_{ij}, x_{ij}^\top\theta^*)\right| = \left|\int_0^1 g'''(y_{ij}, x_{ij}^\top(\theta^* + st(\widehat{\theta} - \theta^*)))dt \cdot tx_{ij}^\top(\widehat{\theta} - \theta^*)\right| \lesssim \left|x_{ij}^\top(\widehat{\theta} - \theta^*)\right|,$$

and then,

$$\begin{aligned}\left\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}_N(\theta^*+s(\widehat{\theta}-\theta^*))\right\|_{\max}&=\left\|\frac{1}{N}\sum_{i=1}^n\sum_{j=1}^k x_{ij}x_{ij}^\top\left(g''(y_{ij},x_{ij}^\top(\theta^*+s(\widehat{\theta}-\theta^*)))-g''(y_{ij},x_{ij}^\top\theta^*)\right)\right\|_{\max}\\
&\le\frac{1}{N}\sum_{i=1}^n\sum_{j=1}^k\left\|x_{ij}x_{ij}^\top\left(g''(y_{ij},x_{ij}^\top(\theta^*+s(\widehat{\theta}-\theta^*)))-g''(y_{ij},x_{ij}^\top\theta^*)\right)\right\|_{\max}\\
&=\frac{1}{N}\sum_{i=1}^n\sum_{j=1}^k\left\|x_{ij}x_{ij}^\top\right\|_{\max}\left|g''(y_{ij},x_{ij}^\top(\theta^*+s(\widehat{\theta}-\theta^*)))-g''(y_{ij},x_{ij}^\top\theta^*)\right|\\
&\lesssim\frac{1}{N}\sum_{i=1}^n\sum_{j=1}^k\|x_{ij}\|_\infty^2\left|x_{ij}^\top(\widehat{\theta}-\theta^*)\right|\le\frac{1}{N}\sum_{i=1}^n\sum_{j=1}^k\|x_{ij}\|_\infty^3\|\widehat{\theta}-\theta^*\|_1\\
&\lesssim\left\|\widehat{\theta}-\theta^*\right\|_1,\end{aligned}\tag{D.2}$$

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

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

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

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

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

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

which implies that

$$\left\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}^*(\theta^*)\right\|_{\max}=O_P\left(\sqrt{\frac{\log d}{N}}\right).\tag{D.3}$$

Then, by the triangle inequality, we have that

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

Note that $\|\Theta\|_\infty\le\sqrt{d}\,\|\Theta\|_2=O\left(\sqrt{d}\right)$. By Lemma F.11, if $N\gtrsim d^4\log d$, we have that

$$\left\|\widehat{\theta}-\theta^*\right\|_1\le\sqrt{d}\left\|\widehat{\theta}-\theta^*\right\|_2=O_P\left(\frac{d\sqrt{\log d}}{\sqrt{N}}\right),$$

and therefore,

$$\left\|\widehat{\theta}-\theta^*+\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty=O_P\left(\frac{d^{5/2}\log d}{N}\right).$$

Putting together the preceding bounds leads to the first result in the lemma. Choosing

$$\zeta = \left(r_{\widetilde{\theta}}\sqrt{N} + \frac{d^{5/2}\log d}{\sqrt{N}} \right)^{1-\kappa},$$

with any $\kappa > 0$, we deduce that $P\left(|T - T_0| > \zeta\right) = o(1)$. We also have that

$$\zeta\sqrt{1 \vee \log\frac{d}{\zeta}}, \quad \text{if} \quad \left(r_{\widetilde{\theta}}\sqrt{N} + \frac{d^{5/2}\log d}{\sqrt{N}} \right) \log^{1/2+\kappa} d = o(1).$$

We complete the proof by simplifying the conditions. □

Lemma D.4. *$\widehat{T}$ and T_0 are defined as in (C.4) and (C.7) respectively. In GLM, under Assumptions (B1)–(B3), provided that $\left\|\widetilde{\theta} - \widehat{\theta}\right\|_\infty = O_P(r_{\widetilde{\theta}})$ and $N \gtrsim d^4 \log d$, we have that*

$$|\widehat{T} - T_0| = O_P\left(r_{\widetilde{\theta}}\sqrt{N} + \frac{d^{5/2}\log d}{\sqrt{N}} \right).$$

Moreover, if $N \gg d^5 \log^{3+\kappa} d$ for some $\kappa > 0$, then there exists some $\xi > 0$ such that (C.16) holds.

Proof of Lemma D.4. By the proof of Lemma D.3, we obtain that if $N \gtrsim d^4 \log d$,

$$\begin{aligned} |\widehat{T} - T_0| &\leq \max_{1\leq l\leq d} \left| \sqrt{N}(\widehat{\theta} - \theta^*)_l + \sqrt{N}\left(\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right)_l \right| \\ &= \sqrt{N}\left\|\widehat{\theta} - \theta^* + \nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}_N(\theta^*)\right\|_\infty = O_P\left(\frac{d^{5/2}\log d}{\sqrt{N}}\right). \end{aligned}$$

Choosing

$$\xi = \left(\frac{d^{5/2}\log d}{\sqrt{N}} \right)^{1-\kappa},$$

with any $\kappa > 0$, we deduce that $P\left(|\widehat{T} - T_0| > \xi\right) = o(1)$. We also have that

$$\xi\sqrt{1 \vee \log\frac{d}{\xi}}, \quad \text{if} \quad \left(\frac{d^{5/2}\log d}{\sqrt{N}} \right) \log^{1/2+\kappa} d = o(1),$$

which holds if $N \gg d^5 \log^{3+\kappa} d$. □

E. Lemmas on Bounding Variance Estimation Errors

Lemma E.1. *$\overline{\Omega}$ and $\widehat{\Omega}$ are defined as in (C.11) and (C.9) respectively. In linear model, under Assumptions (A1) and (A2), provided that $\left\|\bar{\theta} - \theta^*\right\|_1 = O_P(r_{\bar{\theta}})$, $r_{\bar{\theta}}\sqrt{\log(kd)} \lesssim 1$, $n \gtrsim d$, and $k \gtrsim \log^2(dk)\log d$, we have that*

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

Moreover, if $n \gg d\log^{4+\kappa} d$, $k \gg d^2 \log^{5+\kappa} d$, and

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

for some $\kappa > 0$, then there exists some $u > 0$ such that (C.14) holds.

Proof of Lemma E.1. Note by the triangle inequality that

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

where Ω_0 is defined as in (C.10). First, we bound $\left\|\widehat{\Omega}-\Omega_0\right\|_{\max}$. With Assumption (E.1) of (Chernozhukov et al., 2013) verified for $\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*;Z)$ in the proof of Lemma C.1, by the proof of Corollary 3.1 of (Chernozhukov et al., 2013), we have that

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

which implies that

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

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

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

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

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

Under Assumption (A1), by Lemma F.7, if $n \gtrsim d$, we have that $\left\|\widetilde{\Theta}\right\|_\infty = O_P\left(\sqrt{d}\right)$, Then, applying Lemma F.2, we have that

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

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

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

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

and by the triangle inequality,

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

By Lemma F.7, we have that

$$\left\|\left\|\widetilde{\Theta} - \Theta\right\|\right\|_{\max} \leq \max_l \left\|\widetilde{\Theta}_l - \Theta_l\right\|_2 = O_P\left(\sqrt{\frac{d}{n}}\right), \quad \text{and}$$

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

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

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

Putting all the preceding bounds together, we obtain that

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

and finally the first result in the lemma. Choosing

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

with any $\kappa > 0$, we deduce that $P\left(\left\|\left\|\overline{\Omega} - \widehat{\Omega}\right\|\right\|_{\max} > u\right) = o(1)$. We also have that

$$u^{1/3}\left(1 \vee \log\frac{d}{u}\right)^{2/3}, \quad \text{if} \quad \left(d\sqrt{\frac{\log d}{k}} + \frac{d\log^2(dk)\log d}{k} + d\sqrt{\log(kd)}r_{\bar{\theta}} + ndr_{\bar{\theta}}^2 + \sqrt{\frac{d}{n}}\right)\log^{2+\kappa} d = o(1).$$

We complete the proof by simplifying the conditions.

□

Lemma E.2. *$\widehat{\Omega}$ and Ω_0 is defined as in (C.9) and (C.10) respectively. In linear model, under Assumptions (A1) and (A2), we have that*

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

Moreover, if $N \gg \log^{5+\kappa} d$ for some $\kappa > 0$, then there exists some $v > 0$ such that (C.15) holds.

Proof of Lemma E.2. The first result is derived in the proof of Lemma E.1. Choosing

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

with any $\kappa > 0$, we deduce that $P\left(\left\|\left\|\widehat{\Omega} - \Omega_0\right\|\right\|_{\max} > v\right) = o(1)$. We also have that

$$v^{1/3}\left(1 \vee \log\frac{d}{v}\right)^{2/3}, \quad \text{if} \quad \left(\sqrt{\frac{\log d}{N}} + \frac{\log^2(dN)\log d}{N}\right)\log^{2+\kappa} d = o(1),$$

which holds if $N \gg \log^{5+\kappa} d$. □

Lemma E.3. *$\widetilde{\Omega}$ and $\widehat{\Omega}$ are defined as in (C.20) and (C.9) respectively. In linear model, under Assumptions (A1) and (A2), provided that $\left\|\bar{\theta}-\theta^*\right\|_1 = O_P(r_{\bar{\theta}})$, $r_{\bar{\theta}}\sqrt{\log((n+k)d)} \lesssim 1$, and $n \gtrsim d$, we have that*

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

Moreover, if $n \gg d\log^{4+\kappa} d$, $n+k \gg d^2\log^{5+\kappa} d$, and

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

for some $\kappa > 0$, then there exists some $u > 0$ such that (C.21) holds.

Proof of Lemma E.3. Note by the triangle inequality that

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

where Ω_0 is defined as in (C.10). By the proof of Lemma E.1, we have that

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

Next, we bound $\left\|\left\|\widetilde{\Omega}-\Omega_0\right\|\right\|_{\max}$ using the same argument as in the proof of Lemma E.1. By the triangle inequality, we have that

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

We have shown in the proof of Lemma E.1 that

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

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

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

Under Assumption (A1), by Lemma F.7, if $n \gtrsim d$, we have that

$$\left\|\left\|\widetilde{\Theta}\right\|\right\|_{\infty} = O_P\left(\sqrt{d}\right).$$

Then, applying Lemma F.4, we have that

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

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

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

and finally the first result in the lemma. Choosing

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

with any $\kappa > 0$, we deduce that $P\left(\left\|\left\|\widetilde{\Omega}-\widehat{\Omega}\right\|\right\|_{\max} > u\right) = o(1)$. We also have that

$$u^{1/3}\left(1 \vee \log\frac{d}{u}\right)^{2/3}, \quad \text{if}$$

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

We complete the proof by simplifying the conditions.

□

Lemma E.4. *$\overline{\Omega}$ and $\widehat{\Omega}$ are defined as in (C.22) and (C.9) respectively. In GLM, under Assumptions (B1)–(B4), provided that $\left\|\bar{\theta}-\theta^*\right\|_1 = O_P(r_{\bar{\theta}})$, $r_{\bar{\theta}} \lesssim 1$, $n \gtrsim d\log d$, and $k \gtrsim \log d$, we have that*

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

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

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

for some $\kappa > 0$, then there exists some $u > 0$ such that (C.14) holds.

Proof of Lemma E.4. We use the same argument as in the proof of Lemma E.1. Note by the triangle inequality that

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

where Ω_0 is defined as in (C.10). First, we bound $\left\|\left\|\widehat{\Omega}-\Omega_0\right\|\right\|_{\max}$. With Assumption (E.1) of (Chernozhukov et al., 2013) verified for $\nabla^2\mathcal{L}^*(\theta^*)^{-1}\nabla\mathcal{L}(\theta^*;Z)$ in the proof of Lemma C.3, by the proof of Corollary 3.1 of (Chernozhukov et al., 2013), we have that

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

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

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

Note that

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

By the triangle inequality, we have that

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

Note that $\max_l \left\|\Theta_l\right\|_2 \le \left\|\Theta\right\|_2 = O(1)$ under Assumption (B3). By Lemma F.8, provided that $n \gtrsim d\log d$ and $r_{\bar\theta} \lesssim 1$, we have that

$$
I_2 = O_P\left(\frac{d\log d}{n}+r_{\bar\theta}^2+\sqrt{\frac{d\log d}{n}}+r_{\bar\theta}\right) = O_P\left(\sqrt{\frac{d\log d}{n}}+r_{\bar\theta}\right).
$$

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

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

By Lemma F.8, we have that

$$
\left\|\widetilde{\Theta}(\bar\theta)\right\|_\infty = O_P\left(\sqrt{d}\right).
$$

Then, applying Lemma F.5, we obtain that

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

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

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

and finally the first result in the lemma. Choosing

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

with any $\kappa > 0$, we deduce that $P\left(\left\| \overline{\Omega} - \widehat{\Omega} \right\|_{\max} > u \right) = o(1)$. We also have that

$$u^{1/3} \left(1 \vee \log \frac{d}{u} \right)^{2/3}, \quad \text{if} \quad \left(d\sqrt{\frac{\log d}{k}} + d\sqrt{\log d} r_{\bar{\theta}} + ndr_{\bar{\theta}}^2 + \sqrt{\frac{d \log d}{n}} \right) \log^{2+\kappa} d = o(1).$$

We complete the proof by simplifying the conditions.

□

Lemma E.5. *$\widehat{\Omega}$ and Ω_0 is defined as in (C.9) and (C.10) respectively. In GLM, under Assumptions (B3)–(B4), we have that*

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

Moreover, if $N \gg \log^{5+\kappa} d$ for some $\kappa > 0$, then there exists some $v > 0$ such that (C.15) holds.

Proof of Lemma E.5. The first result is derived in the proof of Lemma E.4. Choosing

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

with any $\kappa > 0$, we deduce that $P\left(\left\| \widehat{\Omega} - \Omega_0 \right\|_{\max} > v \right) = o(1)$. We also have that

$$v^{1/3} \left(1 \vee \log \frac{d}{v} \right)^{2/3}, \quad \text{if} \quad \left(\sqrt{\frac{\log d}{N}} + \frac{\log^2(dN) \log d}{N} \right) \log^{2+\kappa} d = o(1),$$

which holds if $N \gg \log^{5+\kappa} d$. □

Lemma E.6. *$\widetilde{\Omega}$ and $\widehat{\Omega}$ are defined as in (C.23) and (C.9) respectively. In GLM, under Assumptions (B1)–(B4), provided that $\left\| \bar{\theta} - \theta^* \right\|_1 = O_P(r_{\bar{\theta}})$, $r_{\bar{\theta}} \lesssim 1$, and $n \gtrsim d \log d$, we have that*

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

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

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

for some $\kappa > 0$, then there exists some $u > 0$ such that (C.21) holds.

Proof of Lemma E.6. Note by the triangle inequality that

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

where Ω_0 is defined as in (C.10). By the proof of Lemma E.4, we have that

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

Next, we bound $\left\|\widetilde{\Omega}-\Omega_0\right\|_{\max}$ using the same argument as in the proof of Lemma E.4. By the triangle inequality, we have that

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

We have shown in the proof of Lemma E.4 that

$$
I_2=O_P\left(\sqrt{\frac{d\log d}{n}}+r_{\bar\theta}\right),
$$

provided that $n\gtrsim d\log d$ and $r_{\bar\theta}\lesssim 1$. To bound $I_1'(\bar\theta)$, we note that

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

By Lemma F.8, we have that

$$
\left\|\widetilde{\Theta}(\bar\theta)\right\|_\infty=O_P\left(\sqrt{d}\right).
$$

Then, applying Lemma F.6, we have that

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

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

Putting all the preceding bounds together, we obtain that

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

and finally the first result in the lemma. Choosing

$$
u=\left(d\sqrt{\frac{\log d}{n+k}}+\frac{n+k\sqrt{\log d}+k^{3/4}\log^{3/4}d}{n+k}dr_{\bar\theta}+\frac{nkd}{n+k}r_{\bar\theta}^2+\sqrt{\frac{d\log d}{n}}\right)^{1-\kappa},
$$

with any $\kappa > 0$, we deduce that $P\left(\left\|\widetilde{\Omega} - \widehat{\Omega}\right\|_{\max} > u\right) = o(1)$. We also have that

$$u^{1/3}\left(1 \vee \log \frac{d}{u}\right)^{2/3}, \quad \text{if} \quad \left(d\sqrt{\frac{\log d}{n+k}} + \frac{n + k\sqrt{\log d} + k^{3/4}\log^{3/4} d}{n+k} dr_{\bar{\theta}} + \frac{nkd}{n+k} r_{\bar{\theta}}^2 + \sqrt{\frac{d \log d}{n}}\right) \log^{2+\kappa} d = o(1).$$

We complete the proof by simplifying the conditions. □

F. Technical Lemmas

Lemma F.1. *For any θ, we have that*

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

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

$U_2 := \left\|\frac{1}{k}\sum_{j=1}^{k} n\nabla\mathcal{L}_j(\theta^*)\nabla\mathcal{L}_j(\theta^*)^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*; Z)\nabla\mathcal{L}(\theta^*; Z)^\top\right]\right\|_{\max}$, *and* $U_3(\theta) := n\left\|\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right\|_\infty^2$.

Proof of Lemma F.1. We write $\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}^*(\theta)$ as $(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)) + (\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta))$, and have that

$$\begin{aligned}
&\sum_{j=1}^{k} n\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top \\
&= \sum_{j=1}^{k} n\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)^\top + nk\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top \\
&\quad + n\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)\sum_{j=1}^{k}\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)^\top + n\sum_{j=1}^{k}\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top \\
&= \sum_{j=1}^{k} n\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)^\top + nk\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top,
\end{aligned}$$

where we use $\nabla\mathcal{L}_N(\theta) = \frac{1}{k}\sum_{j=1}^{k}\nabla\mathcal{L}_j(\theta)$ in the last equality. Then, we have that

$$\begin{aligned}
&\sum_{j=1}^{k} n\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)^\top \\
&= \sum_{j=1}^{k} n\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top - nk\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top,
\end{aligned}$$

and by the triangle inequality,

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

By the fact that $\left\|aa^\top\right\|_{\max} = \|a\|_\infty^2$ for any vector a, we have that $\left\|\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top\right\|_{\max} = n^{-1}U_3(\theta)$. □

Lemma F.2. *In linear model, under Assumptions (A1) and (A2), provided that $\left\|\bar{\theta}-\theta^*\right\|_1 = O_P(r_{\bar{\theta}})$, we have that*

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

Proof of Lemma F.2. By Lemma F.1, it suffices to bound $U_1(\bar{\theta})$, U_2, and $U_3(\bar{\theta})$. We begin by bounding U_2. In linear model, we have that

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

Note that

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

is bounded away from zero, under Assumptions (A1) and (A2). Also, using same argument for obtaining (D.1), we have that

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

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

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

which implies by Markov's inequality that

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

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

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

Under Assumption (A1), each $x_{ij,l}$ is sub-Gaussian, and therefore, the product $x_{ij,l}x_{ij,l'}$ of any two is sub-exponential. By Bernstein's inequality, we have that for any $\delta \in (0,1)$,

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

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

$$P\left(\left\|\frac{X_N^\top X_N}{N} - \Sigma\right\|_{\max} > \|\Sigma\|_{\max} \left(\frac{\log\frac{2d^2}{\delta}}{cN} \vee \sqrt{\frac{\log\frac{2d^2}{\delta}}{cN}}\right)\right) \leq \delta. \tag{F.1}$$

Similarly, we have that

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

By (F.1) and (D.1), we have that

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

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

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

and then,

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

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

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

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

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

and obtain that

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

By the triangle inequality, we have that

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

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

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

Then, we have that

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

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

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

which implies by the union bound that

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

Putting all the preceding bounds together, we obtain that

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

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

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

and finally the bound in the lemma. □

Lemma F.3. *For any θ, we have that*

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

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

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

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

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

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

Proof of Lemma F.3. We write $\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta)$ as $(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta)) + (\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta))$ and $\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}^*(\theta)$ as $(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta)) + (\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta))$, and have that

$$\begin{aligned}
&\sum_{i=1}^{n} \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) \right) \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) \right)^\top \\
&= \sum_{i=1}^{n} \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta) \right)^\top + n \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right) \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right)^\top \\
&\quad + \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right) \sum_{i=1}^{n} \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta) \right)^\top + \sum_{i=1}^{n} \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right)^\top \\
&= \sum_{i=1}^{n} \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta) \right)^\top + n \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right) \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right)^\top \\
&\quad + n \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right) \left(\nabla \mathcal{L}_1(\theta) - \nabla \mathcal{L}_N(\theta) \right)^\top + n \left(\nabla \mathcal{L}_1(\theta) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right)^\top ,
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{j=2}^{k} n \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}^*(\theta) \right) \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}^*(\theta) \right)^\top \\
&= \sum_{j=2}^{k} n \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta) \right)^\top + n(k-1) \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right) \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right)^\top \\
&\quad + n \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right) \sum_{j=2}^{k} \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta) \right)^\top + n \sum_{j=2}^{k} \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right)^\top .
\end{aligned}$$

Adding up the two preceding equations, we obtain that

$$\begin{aligned}
&\sum_{i=1}^{n} \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) \right) \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}^*(\theta) \right)^\top + \sum_{j=2}^{k} n \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}^*(\theta) \right) \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}^*(\theta) \right)^\top \\
&= \sum_{i=1}^{n} \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta) \right)^\top + \sum_{j=2}^{k} n \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta) \right)^\top \\
&\quad + nk \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right) \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right)^\top \\
&\quad + n \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right) \sum_{j=1}^{k} \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta) \right)^\top + n \sum_{j=1}^{k} \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right)^\top \\
&= \sum_{i=1}^{n} \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}(\theta; Z_{i1}) - \nabla \mathcal{L}_N(\theta) \right)^\top + \sum_{j=2}^{k} n \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta) \right) \left(\nabla \mathcal{L}_j(\theta) - \nabla \mathcal{L}_N(\theta) \right)^\top \\
&\quad + nk \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right) \left(\nabla \mathcal{L}_N(\theta) - \nabla \mathcal{L}^*(\theta) \right)^\top ,
\end{aligned}$$

where we use $\nabla\mathcal{L}_N(\theta) = \frac{1}{k}\sum_{j=1}^k \nabla\mathcal{L}_j(\theta)$ in the last equality. Then, we have that

$$\begin{aligned}
&\sum_{i=1}^n \left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}_N(\theta)\right)^\top + \sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)^\top \\
&= \sum_{i=1}^n \left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}^*(\theta)\right)^\top + \sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top \\
&\quad - nk\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top,
\end{aligned}$$

and by the triangle inequality,

$$\begin{aligned}
&\left\|\left\|\frac{1}{n+k-1}\left(\sum_{i=1}^n \left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}_N(\theta)\right)^\top\right.\right.\right. \\
&\quad \left.\left.\left.+ \sum_{j=2}^k n\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)\left(\nabla\mathcal{L}_j(\theta) - \nabla\mathcal{L}_N(\theta)\right)^\top\right) - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*; Z)\nabla\mathcal{L}(\theta^*; Z)^\top\right]\right\|\right\|_{\max} \\
&\le \frac{n}{n+k-1}\left\|\left\|\sum_{i=1}^n \left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}(\theta; Z_{i1}) - \nabla\mathcal{L}^*(\theta)\right)^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*; Z)\nabla\mathcal{L}(\theta^*; Z)^\top\right]\right\|\right\|_{\max} \\
&\quad + \frac{k-1}{n+k-1}\left\|\left\|\sum_{j=2}^k \left(\sqrt{n}\nabla\mathcal{L}_j(\theta) - \sqrt{n}\nabla\mathcal{L}^*(\theta)\right)\left(\sqrt{n}\nabla\mathcal{L}_j(\theta) - \sqrt{n}\nabla\mathcal{L}^*(\theta)\right)^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*; Z)\nabla\mathcal{L}(\theta^*; Z)^\top\right]\right\|\right\|_{\max} \\
&\quad + \frac{nk}{n+k-1}\left\|\left\|\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top\right\|\right\|_{\max} \\
&:= A(\theta) + B(\theta) + \frac{nk}{n+k-1}\left\|\left\|\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top\right\|\right\|_{\max}.
\end{aligned}$$

By the fact that $\left\|\left\|aa^\top\right\|\right\|_{\max} = \|a\|_\infty^2$ for any vector a, we have that $\left\|\left\|\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)\left(\nabla\mathcal{L}_N(\theta) - \nabla\mathcal{L}^*(\theta)\right)^\top\right\|\right\|_{\max} = (n+k-1)(nk)^{-1}V_3(\theta)$. We apply the triangle inequality to further decompose $A(\theta)$ and $B(\theta)$ and obtain that $B(\theta) \le V_1(\theta) + V_2$ and $A(\theta) \le V_1'(\theta) + V_2'$.

$\square$

Lemma F.4. *In linear model, under Assumptions (A1) and (A2), provided that $\left\|\bar{\theta} - \theta^*\right\|_1 = O_P(r_{\bar{\theta}})$, we have that*

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

Proof of Lemma F.4. By Lemma F.3, it suffices to bound $V_1(\bar\theta)$, $V_1'(\bar\theta)$, V_2, V_2', and $V_3(\bar\theta)$. By the proof of Lemma F.2, we have that under Assumptions (A1) and (A2), assuming that $\left\|\bar\theta-\theta^*\right\|_1 = O_P(r_{\bar\theta})$,

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

$$V_2 = \frac{k-1}{n+k-1} O_P\left(\sqrt{\frac{\log d}{k}}+\frac{\log^2(dk)\log d}{k}\right) = O_P\left(\frac{\sqrt{k\log d}}{n+k}+\frac{\log^2(dk)\log d}{n+k}\right), \quad \text{and}$$
$$V_3(\bar\theta) = \frac{nk}{n+k-1} O_P\left(\left(1+\sqrt{\frac{\log d}{N}}\right) r_{\bar\theta}^2 + \frac{\log d}{N}\right) = O_P\left(\left(1+\sqrt{\frac{\log d}{N}}\right)\frac{nk}{n+k} r_{\bar\theta}^2 + \frac{\log d}{n+k}\right).$$

It remains to bound $V_1'(\bar\theta)$ and V_2'.To bound V_2', we have that in linear model, under Assumptions (A1) and (A2),

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

Note that $\mathbb{E}\left[(x_{i1}e_{i1})_l^2\right] = \sigma^2\Sigma_{l,l}$ is bounded away from zero, and also that $(x_{i1}e_{i1})_l$ is sub-exponential with O(1) ψ_1-norm for each (i,l). Then, by the proof of Corollary 3.1 of (Chernozhukov et al., 2013), we have that

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

which implies by Markov's inequality that

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

Lastly, we bound $V_1'(\bar\theta)$ using the same argument as in bounding $U_1(\bar\theta)$ in the proof of Lemma F.2. We write $\nabla\mathcal{L}(\theta;Z_{i1}) - \nabla\mathcal{L}^*(\theta)$ as $(\nabla\mathcal{L}(\theta;Z_{i1}) - \nabla\mathcal{L}^*(\theta) - \nabla\mathcal{L}(\theta^*;Z_{i1})) + \nabla\mathcal{L}(\theta^*;Z_{i1})$, and obtain by the triangle inequality that

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

Applying Cauchy-Schwarz inequality, we obtain that

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

By the triangle inequality, we have that

$$\begin{aligned}
&\left\|\!\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla\mathcal{L}(\theta^*;Z_{i1})\nabla\mathcal{L}(\theta^*;Z_{i1})^\top\right\|\!\right\|_{\max} \\
&\le \left\|\!\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla\mathcal{L}(\theta^*;Z_{i1})\nabla\mathcal{L}(\theta^*;Z_{i1})^\top - \mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|\!\right\|_{\max} + \left\|\!\left\|\mathbb{E}\left[\nabla\mathcal{L}(\theta^*;Z)\nabla\mathcal{L}(\theta^*;Z)^\top\right]\right\|\!\right\|_{\max} \\
&= \frac{n+k-1}{n}V_2' + \sigma^2\left\|\!\left\|\Sigma\right\|\!\right\|_{\max} = O_P\left(1+\sqrt{\frac{\log d}{n}}+\frac{\log^2(dn)\log d}{n}\right).
\end{aligned}$$

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

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

Then, we have by the triangle inequality that

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

Similarly to obtaining (F.2), we have that

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

which implies by the union bound that

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

Putting all the preceding bounds together, we obtain that

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

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

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

and finally the bound in the lemma. □

Lemma F.5. *In GLM, under Assumptions (B1)–(B3), provided that* $\left\|\bar\theta - \theta^*\right\|_1 = O_P(r_{\bar\theta})$, *we have that*

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

Proof of Lemma F.5. By Lemma F.1, it suffices to bound $U_1(\bar\theta)$, U_2, and $U_3(\bar\theta)$. We begin by bounding U_2. Note that $\nabla\mathcal{L}_N(\theta^*) = \sum_{i=1}^n\sum_{j=1}^k g'(y_{ij}, x_{ij}^\top\theta^*)x_{ij}/N$ and $g'(y_{ij}, x_{ij}^\top\theta^*)x_{ij,l} = O(1)$ for each $l = 1,\ldots,d$ under Assumptions (B1) and (B2). Then, by Hoeffding's inequality, we have that for any $t > 0$,

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

that is, $\sqrt{n}\nabla\mathcal{L}_j(\theta^*)_l$ is sub-Gaussian with $O(1)$ ψ_2-norm. Therefore, $n\nabla\mathcal{L}_j(\theta^*)_l\nabla\mathcal{L}_j(\theta^*)_{l'}$ is sub-exponential with $O(1)$ ψ_1-norm. Note that $\mathbb{E}[n\nabla\mathcal{L}_j(\theta^*)_l\nabla\mathcal{L}_j(\theta^*)_{l'}] = \mathbb{E}[\nabla\mathcal{L}(\theta^*;Z)_l\nabla\mathcal{L}(\theta^*;Z)_{l'}]$. Then, we apply Bernstein's inequality and obtain that for any $\delta \in (0,1)$,

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

which implies by the union bound that

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

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

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

By an expression of remainder of the first order Taylor expansion, we have that

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

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

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

Note that for any θ,

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

Therefore, $\left\|\nabla\mathcal{L}^*(\bar\theta)\right\|_\infty \lesssim \left\|\bar\theta-\theta^*\right\|_\infty$. By (F.5), we have that

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

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

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

and then,

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

Lastly, we bound $U_1(\bar\theta)$. As in the proof of Lemma F.2, we have that

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

and

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

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

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

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

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

and

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

and then

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

In a similar way, we have that

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

and then,

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

Then, we have by the triangle inequality that

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

Using the argument for obtaining (D.3), we have that

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

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

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

Hence, we have that

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

Putting all the preceding bounds together, we obtain that

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

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

and finally the bound in the lemma. □

Lemma F.6. *In GLM, under Assumptions (B1)–(B3), provided that $\left\|\bar\theta - \theta^*\right\|_1 = O_P(r_{\bar\theta})$, we have that*

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

Proof of Lemma F.6. By Lemma F.3, it suffices to bound $V_1(\bar\theta)$, $V_1'(\bar\theta)$, V_2, V_2', and $V_3(\bar\theta)$. By the proof of Lemma F.5, we have that under Assumptions (B1)–(B3), assuming that $\left\|\bar\theta - \theta^*\right\|_1 = O_P(r_{\bar\theta})$,

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

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

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

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

To bound V_2', we note that each $\nabla\mathcal{L}(\theta^*; Z_{i1})_l \nabla\mathcal{L}(\theta^*; Z_{i1})_{l'} = g'(y_{i1}, x_{i1}^\top\theta^*)^2 x_{i1,l} x_{i1,l'}$ is bounded under Assumptions (B1) and (B2). Applying Hoeffding's inequality, we obtain that for any $\delta \in (0,1)$,

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

which implies by the union bound that

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

Lastly, we bound $V_1'(\bar\theta)$. As in the proof of Lemma F.4, we have that

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

and

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

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

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

It remains to bound $V_{11}'(\bar\theta)$. Using the same argument for analyzing $\nabla\mathcal{L}_j(\bar\theta) - \nabla\mathcal{L}^*(\bar\theta) - \nabla\mathcal{L}_j(\theta^*)$ in the proof of Lemma F.5, we obtain that

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

and

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

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

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

and

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

and hence,

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

Putting all the preceding bounds together, we obtain that

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

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

and finally the bound in the lemma. □

Lemma F.7. *In linear model, under Assumption (A1), if $n \gtrsim d$, we have that*

$$\left\|\left\|\widetilde{\Theta}\right\|\right\|_\infty = O_P\left(\sqrt{d}\right) \quad \textit{and} \quad \max_l \left\|\widetilde{\Theta}_l - \Theta_l\right\|_2 = O_P\left(\sqrt{\frac{d}{n}}\right).$$

Proof of Lemma F.7. $\widetilde{\Theta}$ is simply the inverse of $X_1^\top X_1/n$. We use the fact that for any matrix $A, B \in \mathbb{R}^{d\times d}$, $\left\|\left\|A^{-1} - B^{-1}\right\|\right\|_2 \le \left\|\left\|B^{-1}\right\|\right\|_2^2 \left\|\left\|A - B\right\|\right\|_2$, and obtain that

$$\left\|\left\|\widetilde{\Theta} - \Theta\right\|\right\|_2 = \left\|\left\|\left(\frac{X_1^\top X_1}{n}\right)^{-1} - \Sigma^{-1}\right\|\right\|_2 \le \left\|\left\|\Sigma^{-1}\right\|\right\|_2^2 \left\|\left\|\frac{X_1^\top X_1}{n} - \Sigma\right\|\right\|_2.$$

Since the design matrix is sub-Gaussian and $\left\|\left\|\Sigma\right\|\right\|_2 = O(1)$, by Proposition 2.1 of (Vershynin, 2012), we have that if $n \gtrsim d$,

$$\left\|\left\|\frac{X_1^\top X_1}{n} - \Sigma\right\|\right\|_2 = O_P\left(\sqrt{\frac{d}{n}}\right).$$

Also note that $\left\|\left\|\Sigma^{-1}\right\|\right\|_2 = O(1)$, and then, we have that

$$\max_l \left\|\widetilde{\Theta}_l - \Theta_l\right\|_2 \le \left\|\left\|\widetilde{\Theta} - \Theta\right\|\right\|_2 = O_P\left(\sqrt{\frac{d}{n}}\right), \quad \text{and} \tag{F.3}$$

$$\left\|\left\|\widetilde{\Theta} - \Theta\right\|\right\|_\infty \le \sqrt{d}\left\|\left\|\widetilde{\Theta} - \Theta\right\|\right\|_2 = O_P\left(\frac{d}{\sqrt{n}}\right).$$

Note that $\left\|\left\|\Theta\right\|\right\|_\infty \le \sqrt{d}\left\|\left\|\Theta\right\|\right\|_2 = \sqrt{d}\left\|\left\|\Sigma^{-1}\right\|\right\|_2 = O\left(\sqrt{d}\right)$. By the triangle inequality, we have that

$$\left\|\left\|\widetilde{\Theta}\right\|\right\|_\infty \le \left\|\left\|\widetilde{\Theta} - \Theta\right\|\right\|_\infty + \left\|\left\|\Theta\right\|\right\|_\infty = O_P\left(\frac{d}{\sqrt{n}}\right) + O\left(\sqrt{d}\right) = O_P\left(\sqrt{d}\right).$$

□

Lemma F.8. *In GLM, under Assumptions (B1)–(B3), if $n \gtrsim d\log d$ and $r_{\bar{\theta}} \lesssim 1$, we have that*

$$\left\|\left\|\widetilde{\Theta}(\bar{\theta})\right\|\right\|_\infty = O_P\left(\sqrt{d}\right) \quad \textit{and} \quad \max_l \left\|\widetilde{\Theta}(\bar{\theta})_l - \Theta_l\right\|_2 = O_P\left(\sqrt{\frac{d\log d}{n}} + r_{\bar{\theta}}\right).$$

Proof of Lemma F.8. $\widetilde{\Theta}(\bar{\theta})$ is simply the inverse of $\nabla^2\mathcal{L}_1(\bar{\theta})$. Then, we have that

$$\left\|\widetilde{\Theta}(\bar{\theta})-\Theta\right\|_2 = \left\|\nabla^2\mathcal{L}_1(\bar{\theta})^{-1}-\nabla^2\mathcal{L}^*(\theta^*)^{-1}\right\|_2 \leq \left\|\nabla^2\mathcal{L}^*(\theta^*)^{-1}\right\|_2^2 \left\|\nabla^2\mathcal{L}_1(\bar{\theta})-\nabla^2\mathcal{L}^*(\theta^*)\right\|_2.$$

Note that

$$\left\|\nabla^2\mathcal{L}_1(\theta^*)-\nabla^2\mathcal{L}^*(\theta^*)\right\|_2 = \left\|\frac{1}{n}\sum_{i=1}^n g''(y_{ij},x_{ij}^\top\theta^*)x_{ij}x_{ij}^\top - \mathbb{E}[g''(y,x^\top\theta^*)xx^\top]\right\|_2,$$

$$\left\|\sqrt{g''(y_{ij},x_{ij}^\top\theta^*)}x_{ij}\right\|_2 = \sqrt{d}\left\|\sqrt{g''(y_{ij},x_{ij}^\top\theta^*)}x_{ij}\right\|_\infty = O(\sqrt{d}),$$

and $\left\|\nabla^2\mathcal{L}^*(\theta^*)\right\|_2 = O(1)$. By Section 1.6.3 of (Tropp et al., 2015), we have that if $n \gtrsim d\log d$,

$$\mathbb{E}\left[\left\|\nabla^2\mathcal{L}_1(\theta^*)-\nabla^2\mathcal{L}^*(\theta^*)\right\|_2\right] \lesssim \sqrt{\frac{d\log d}{n}},$$

which implies that

$$\left\|\nabla^2\mathcal{L}_1(\theta^*)-\nabla^2\mathcal{L}^*(\theta^*)\right\|_2 = O_P\left(\sqrt{\frac{d\log d}{n}}\right).$$

Also note that

$$\left\|\nabla^2\mathcal{L}_1(\bar{\theta})-\nabla^2\mathcal{L}_1(\theta^*)\right\|_2 = \left\|\frac{1}{n}\sum_{i=1}^n \left(g''(y_{ij},x_{ij}^\top\bar{\theta})-g''(y_{ij},x_{ij}^\top\theta^*)\right)x_{ij}x_{ij}^\top\right\|_2 \lesssim \left\|\bar{\theta}-\theta^*\right\|_1.$$

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

$$\left\|\nabla^2\mathcal{L}_1(\bar{\theta})-\nabla^2\mathcal{L}^*(\theta^*)\right\|_2 \leq \left\|\nabla^2\mathcal{L}_1(\bar{\theta})-\nabla^2\mathcal{L}_1(\theta^*)\right\|_2 + \left\|\nabla^2\mathcal{L}_1(\theta^*)-\nabla^2\mathcal{L}^*(\theta^*)\right\|_2 = O_P\left(\sqrt{\frac{d\log d}{n}}+r_{\bar{\theta}}\right).$$

Since $\left\|\nabla^2\mathcal{L}^*(\theta^*)^{-1}\right\|_2 = O(1)$, we have that

$$\max_l \left\|\widetilde{\Theta}(\bar{\theta})_l-\Theta_l\right\|_2 \leq \left\|\widetilde{\Theta}(\bar{\theta})-\Theta\right\|_2 = O_P\left(\sqrt{\frac{d\log d}{n}}+r_{\bar{\theta}}\right), \quad \text{and}$$

$$\left\|\widetilde{\Theta}(\bar{\theta})-\Theta\right\|_\infty \leq \sqrt{d}\left\|\widetilde{\Theta}(\bar{\theta})-\Theta\right\|_2 = O_P\left(d\sqrt{\frac{\log d}{n}}+\sqrt{d}r_{\bar{\theta}}\right).$$

Note that $\|\Theta\|_\infty \leq \sqrt{d}\,\|\Theta\|_2 = \sqrt{d}\left\|\nabla^2\mathcal{L}^*(\theta^*)^{-1}\right\|_2 = O\left(\sqrt{d}\right)$. By the triangle inequality, if $r_{\bar{\theta}} \lesssim 1$, we have that

$$\left\|\widetilde{\Theta}(\bar{\theta})\right\|_\infty \leq \left\|\widetilde{\Theta}(\bar{\theta})-\Theta\right\|_\infty + \|\Theta\|_\infty = O_P\left(d\sqrt{\frac{\log d}{n}}+\sqrt{d}r_{\bar{\theta}}\right) + O\left(\sqrt{d}\right) = O_P\left(\sqrt{d}\right).$$

□

Lemma F.9. *In linear model, under Assumptions (A1) and (A2), if $N \gtrsim d$, then we have that*

$$\left\|\widehat{\theta}-\theta^*\right\|_2 \lesssim \sqrt{\frac{d\log\frac{d}{\delta}}{N}}+\frac{d\log\frac{d}{\delta}}{N},$$

with probability at least $1-\delta$, for any δ such that $e^{-N} \lesssim \delta < 1$.

Proof of Lemma F.9. Note that

$$\left\|\widehat{\theta}-\theta^*\right\|_2 = \left\|\left(X_N^\top X_N\right)^{-1} X_N^\top y_N - \theta^*\right\|_2 = \left\|\left(X_N^\top X_N\right)^{-1} X_N^\top e_N\right\|_2 \leq \left\|\left(\frac{X_N^\top X_N}{N}\right)^{-1}\right\|_2 \left\|\frac{X_N^\top e_N}{N}\right\|_2.$$

By (D.1), we have with probability at least $1-\delta$ that

$$\left\|\frac{X_N^\top e_N}{N}\right\|_2 \leq \sqrt{d}\left\|\frac{X_N^\top e_N}{N}\right\|_\infty \lesssim \sqrt{\frac{d\log\frac{d}{\delta}}{N}} + \frac{d\log\frac{d}{\delta}}{N}.$$

By Proposition 2.1 of (Vershynin, 2012), if $n \gtrsim d$, we have with probability at least $1-\delta$ that

$$\left\|\frac{X_N^\top X_N}{N} - \Sigma\right\|_2 \lesssim \sqrt{\frac{d+\log\frac{1}{\delta}}{N}} + \frac{d+\log\frac{1}{\delta}}{N},$$

and then, by the triangle inequality,

$$\begin{aligned}\left\|\left(\frac{X_N^\top X_N}{N}\right)^{-1}\right\|_2 &\leq \left\|\left(\frac{X_N^\top X_N}{N}\right)^{-1} - \Theta\right\|_2 + \|\Theta\|_2 \leq \|\Theta\|_2^2 \left\|\frac{X_N^\top X_N}{N} - \Sigma\right\|_2 + \|\Theta\|_2 \\ &\lesssim \sqrt{\frac{d+\log\frac{1}{\delta}}{N}} + \frac{d+\log\frac{1}{\delta}}{N} + 1 \lesssim 1, \end{aligned} \tag{F.4}$$

provided that $N \gtrsim d + \log(1/\delta)$. Finally, by the union bound, we have with probability at least $1-2\delta$ that

$$\left\|\widehat{\theta}-\theta^*\right\|_2 \lesssim \sqrt{\frac{d\log\frac{d}{\delta}}{N}} + \frac{d\log\frac{d}{\delta}}{N}.$$

□

Lemma F.10. *In linear model, under Assumptions (A1) and (A2), if $n \gtrsim d$, then we have that for any $t \geq 1$,*

$$\left\|\widetilde{\theta}^{(t)} - \widehat{\theta}\right\|_2 \lesssim \sqrt{\frac{d+\log\frac{1}{\delta}}{n}} \left\|\widetilde{\theta}^{(t-1)} - \widehat{\theta}\right\|_2,$$

with probability at least $1-\delta$, for any δ such that $e^{-n} \lesssim \delta < 1$, where $\widetilde{\theta}^{(t)}$ is the t-step CSL estimator defined in Algorithm 2.

Proof of Lemma F.10. Note that

$$\begin{aligned}\left\|\widetilde{\theta}^{(t)} - \widehat{\theta}\right\|_2 &= \left\|\widetilde{\theta}^{(t-1)} - \nabla^2\mathcal{L}_1(\widetilde{\theta}^{(t-1)})^{-1}\nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) - \widehat{\theta}\right\|_2 \\ &= \left\|\widetilde{\theta}^{(t-1)} - \left(\frac{X_1^\top X_1}{n}\right)^{-1} \frac{X_N^\top\left(X_N\widetilde{\theta}^{(t-1)}) - y_N\right)}{N} - \left(\frac{X_N^\top X_N}{N}\right)^{-1}\frac{X_N^\top y_N}{N}\right\|_2 \\ &= \left\|\widetilde{\theta}^{(t-1)} - \left(\frac{X_1^\top X_1}{n}\right)^{-1} \frac{X_N^\top\left(X_N\widetilde{\theta}^{(t-1)}) - y_N\right)}{N} - \widetilde{\theta}^{(t-1)} + \left(\frac{X_N^\top X_N}{N}\right)^{-1}\frac{X_N^\top\left(X_N\widetilde{\theta}^{(t-1)}) - y_N\right)}{N}\right\|_2 \\ &\leq \left\|\left(\frac{X_1^\top X_1}{n}\right)^{-1} - \left(\frac{X_N^\top X_N}{N}\right)^{-1}\right\|_2 \left\|\frac{X_N^\top X_N\left(\widetilde{\theta}^{(t-1)} - \widehat{\theta}\right)}{N}\right\|_2 \\ &\leq \left\|\left(\frac{X_1^\top X_1}{n}\right)^{-1} - \left(\frac{X_N^\top X_N}{N}\right)^{-1}\right\|_2 \left\|\frac{X_N^\top X_N}{N}\right\|_2 \left\|\widetilde{\theta}^{(t-1)} - \widehat{\theta}\right\|_2.\end{aligned}$$

By (F.4) with triangle inequality and the union bound, we have with probability at least $1-\delta$ that

$$\begin{aligned}\left\|\left(\frac{X_1^\top X_1}{n}\right)^{-1}-\left(\frac{X_N^\top X_N}{N}\right)^{-1}\right\|_2 &\le \left\|\left(\frac{X_1^\top X_1}{n}\right)^{-1}-\Theta\right\|_2+\left\|\left(\frac{X_N^\top X_N}{N}\right)^{-1}-\Theta\right\|_2\\ &\lesssim \sqrt{\frac{d+\log\frac{1}{\delta}}{n}}+\frac{d+\log\frac{1}{\delta}}{n}+\sqrt{\frac{d+\log\frac{1}{\delta}}{N}}+\frac{d+\log\frac{1}{\delta}}{N}\\ &\lesssim \sqrt{\frac{d+\log\frac{1}{\delta}}{n}}+\frac{d+\log\frac{1}{\delta}}{n},\quad \text{and}\end{aligned}$$

$$\left\|\frac{X_N^\top X_N}{N}\right\|_2\le\left\|\frac{X_N^\top X_N}{N}-\Sigma\right\|_2+\|\Sigma\|_2\le\sqrt{\frac{d+\log\frac{1}{\delta}}{N}}+\frac{d+\log\frac{1}{\delta}}{N}+1.$$

Provided that $d+\log\frac{1}{\delta}\lesssim n$, we obtain the bound in the lemma. □

Lemma F.11. *In GLM, under Assumptions (B1)–(B3), if $N\gtrsim d^4\log d$, then we have that*

$$\left\|\widehat{\theta}-\theta^*\right\|_2\lesssim\sqrt{\frac{d\log\frac{d}{\delta}}{N}},$$

with probability at least $1-\delta$, for any δ such that $e^{-N/d^4}\lesssim\delta<1$.

Proof of Lemma F.11. We use the argument in the proof of Lemma 6 of (Zhang et al., 2012). By Theorem 1.6.2 of (Tropp et al., 2015), we have with probability at least $1-\delta$ that

$$\left\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}^*(\theta^*)\right\|_2\le C\sqrt{\frac{d\log\frac{d}{\delta}}{N}}+C\frac{d\log\frac{d}{\delta}}{N},$$

for some constant $C>0$. By (D.2), for any θ, we have that

$$\left\|\nabla^2\mathcal{L}_N(\theta)-\nabla^2\mathcal{L}_N(\theta^*)\right\|_2\le d\left\|\nabla^2\mathcal{L}_N(\theta)-\nabla^2\mathcal{L}_N(\theta^*)\right\|_{\max}\le Cd\left\|\theta-\theta^*\right\|_1\le Cd^{3/2}\left\|\theta-\theta^*\right\|_2.$$

Let $\rho=(4C\mu d^{3/2})^{-1}$ and assume $4C\mu\sqrt{d\log(d/\delta)/N}\le 1$ and $4C\mu d\log(d/\delta)/N\le 1$. Then, for any $\theta\in U:=\{\theta:\|\theta-\theta^*\|_2\le\rho\}$, we have by the triangle inequality that

$$\left\|\nabla^2\mathcal{L}_N(\theta)-\nabla^2\mathcal{L}^*(\theta^*)\right\|_2\le\left\|\nabla^2\mathcal{L}_N(\theta)-\nabla^2\mathcal{L}_N(\theta^*)\right\|_2+\left\|\nabla^2\mathcal{L}_N(\theta^*)-\nabla^2\mathcal{L}^*(\theta^*)\right\|_2\le(2\mu)^{-1}.$$

Since $\lambda_{\min}(\nabla^2\mathcal{L}^*(\theta^*))\ge\mu^{-1}$, we have $\lambda_{\min}(\nabla^2\mathcal{L}_N(\theta))\ge(2\mu)^{-1}$ for any $\theta\in U$. Then, for any $\theta'\in\mathbb{R}^d$, we have that

$$\mathcal{L}_N(\theta')\ge\mathcal{L}_N(\theta^*)+\nabla\mathcal{L}_N(\theta^*)^\top(\theta'-\theta^*)+(4\mu)^{-1}\min\left\{\|\theta'-\theta^*\|_2^2,\rho^2\right\},$$

and then,

$$\begin{aligned}\min\left\{\|\theta'-\theta^*\|_2^2,\rho^2\right\}&\le 4\mu(\mathcal{L}_N(\theta')-\mathcal{L}_N(\theta^*)-\nabla\mathcal{L}_N(\theta^*)^\top(\theta'-\theta^*))\\ &\le 4\mu(\mathcal{L}_N(\theta')-\mathcal{L}_N(\theta^*)+\|\nabla\mathcal{L}_N(\theta^*)\|_2\,\|\theta'-\theta^*\|_2).\end{aligned}$$

Dividing both sides by $\|\theta'-\theta^*\|_2$ and then setting $\theta'=\kappa\widehat{\theta}+(1-\kappa)\theta^*$ for any $\kappa\in[0,1]$, we have

$$\min\left\{\kappa\left\|\widehat{\theta}-\theta^*\right\|_2,\frac{\rho^2}{\kappa\left\|\widehat{\theta}-\theta^*\right\|_2}\right\}\le\frac{4\mu\left(\mathcal{L}_N(\kappa\widehat{\theta}+(1-\kappa)\theta^*)-\mathcal{L}_N(\theta^*)\right)}{\kappa\left\|\widehat{\theta}-\theta^*\right\|_2}+4\mu\|\nabla\mathcal{L}_N(\theta^*)\|_2<4\mu\|\nabla\mathcal{L}_N(\theta^*)\|_2,$$

where we use that $\mathcal{L}_N(\kappa\widehat{\theta} + (1-\kappa)\theta^*) < \mathcal{L}_N(\theta^*)$ for any $\kappa \in (0,1)$ since $\mathcal{L}_N$ is strongly convex at θ^* and $\widehat{\theta}$ minimizes $\mathcal{L}_N$. Note that $\nabla\mathcal{L}_N(\theta^*) = \sum_{i=1}^n \sum_{j=1}^k g'(y_{ij}, x_{ij}^\top\theta^*)x_{ij}/N$ and $g'(y_{ij}, x_{ij}^\top\theta^*)x_{ij,l} = O(1)$ for each $l = 1, \dots, d$ under Assumptions (B1) and (B2). Then, by Hoeffding's inequality, we have that

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

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

$$\|\nabla\mathcal{L}_N(\theta^*)\|_\infty \le \sqrt{\frac{c\log\frac{2d}{\delta}}{N}}. \tag{F.5}$$

Then, we have with probability at least $1-\delta$ that

$$\|\nabla\mathcal{L}_N(\theta^*)\|_2 \le \sqrt{d}\,\|\nabla\mathcal{L}_N(\theta^*)\|_\infty \le C\sqrt{\frac{d\log\frac{d}{\delta}}{N}},$$

and by the union bound, with probability at least $1-2\delta$,

$$\min\left\{\kappa\left\|\widehat{\theta}-\theta^*\right\|_2, \frac{\rho^2}{\kappa\left\|\widehat{\theta}-\theta^*\right\|_2}\right\} < 4C\mu\sqrt{\frac{d\log\frac{d}{\delta}}{N}} \le \rho,$$

provided that $4C\mu\sqrt{d\log(d/\delta)/N} \le \rho$. Since this holds for any $\kappa \in (0,1)$, if $\left\|\widehat{\theta}-\theta^*\right\|_2 > \rho$, we may set $\kappa = \rho/\left\|\widehat{\theta}-\theta^*\right\|_2 < 1$, and find that

$$\min\left\{\kappa\left\|\widehat{\theta}-\theta^*\right\|_2, \frac{\rho^2}{\kappa\left\|\widehat{\theta}-\theta^*\right\|_2}\right\} = \rho,$$

which would yield a contradiction. Thus, we have $\left\|\widehat{\theta}-\theta^*\right\|_2 \le \rho$, that is, $\widehat{\theta} \in U$. Furthermore, we have that

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

and thus,

$$\left\|\widehat{\theta}-\theta^*\right\|_2 \le 4\mu\|\nabla\mathcal{L}_N(\theta^*)\|_2 \le 4C\mu\sqrt{\frac{d\log\frac{d}{\delta}}{N}},$$

with probability at least $1-2\delta$, provided that $4C\mu\sqrt{d\log(d/\delta)/N} \le 1$, $4C\mu d\log(d/\delta)/N \le 1$, and $4C\mu\sqrt{d\log(d/\delta)/N} \le \rho$, which hold if $\delta \gtrsim e^{-N/d^4}$ and $N \gtrsim d^4\log d$. □

Lemma F.12. *In GLM, under Assumptions (B1)–(B3), if $n \gtrsim d^4\log d$, then we have that for any $t \ge 1$,*

$$\left\|\widetilde{\theta}^{(t)}-\widehat{\theta}\right\|_2 \lesssim \left(\sqrt{\frac{d\log\frac{d}{\delta}}{n}} + d^{3/2}\left\|\widetilde{\theta}^{(t-1)}-\widehat{\theta}\right\|_2\right)\left\|\widetilde{\theta}^{(t-1)}-\widehat{\theta}\right\|_2,$$

with probability at least $1-\delta$, for any δ such that $e^{-n/d^4} \lesssim \delta < 1$, where $\widetilde{\theta}^{(t)}$ is the t-step CSL estimator defined in Algorithm 2.

Proof of Lemma F.12. We use the argument in the proof of Theorem 3 of (Jordan et al., 2019). Note by the triangle inequality that

$$\begin{aligned}
&\left\|\widetilde{\theta}^{(t)}-\widehat{\theta}\right\|_2 \\
&= \left\|\widetilde{\theta}^{(t-1)} - \nabla^2\mathcal{L}_1(\widetilde{\theta}^{(t-1)})^{-1}\nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) - \widehat{\theta}\right\|_2 \\
&\le \left\|\widetilde{\theta}^{(t-1)} - \nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1}\nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) - \widehat{\theta}\right\|_2 + \left\|\left(\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1} - \nabla^2\mathcal{L}_1(\widetilde{\theta}^{(t-1)})^{-1}\right)\nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)})\right\|_2.
\end{aligned}$$

To bound the first term on the right hand side, we have that

$$\begin{aligned}
&\left\|\widetilde{\theta}^{(t-1)} - \nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1}\nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) - \widehat{\theta}\right\|_2 \\
&= \left\|\widetilde{\theta}^{(t-1)} - \widehat{\theta} - \nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1}\left(\nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) - \nabla\mathcal{L}_N(\widehat{\theta})\right)\right\|_2 \\
&= \left\|\widetilde{\theta}^{(t-1)} - \widehat{\theta} - \nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1}\int_0^1 \nabla^2\mathcal{L}_N(\widehat{\theta} + s(\widetilde{\theta}^{(t-1)} - \widehat{\theta}))ds\left(\widetilde{\theta}^{(t-1)} - \widehat{\theta}\right)\right\|_2 \\
&= \left\|\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1}\int_0^1 \nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) - \nabla^2\mathcal{L}_N(\widehat{\theta} + s(\widetilde{\theta}^{(t-1)} - \widehat{\theta}))ds\left(\widetilde{\theta}^{(t-1)} - \widehat{\theta}\right)\right\|_2 \\
&\le \left\|\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1}\right\|_2 \int_0^1 \left\|\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) - \nabla^2\mathcal{L}_N(\widehat{\theta} + s(\widetilde{\theta}^{(t-1)} - \widehat{\theta}))\right\|_2 ds \left\|\widetilde{\theta}^{(t-1)} - \widehat{\theta}\right\|_2.
\end{aligned}$$

By the proof of Lemma F.11, we have that

$$\begin{aligned}
\left\|\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1}\right\|_2 &\le \left\|\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1} - \nabla^2\mathcal{L}^*(\theta^*)^{-1}\right\|_2 + \left\|\nabla^2\mathcal{L}^*(\theta^*)^{-1}\right\|_2 \\
&\le \left\|\nabla^2\mathcal{L}^*(\theta^*)^{-1}\right\|_2^2 \left\|\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) - \nabla^2\mathcal{L}^*(\theta^*)\right\|_2 + \left\|\nabla^2\mathcal{L}^*(\theta^*)^{-1}\right\|_2 \\
&\lesssim \sqrt{\frac{d\log\frac{d}{\delta}}{N}} + \frac{d\log\frac{d}{\delta}}{N} + d^{3/2}\left\|\widetilde{\theta}^{(t-1)} - \theta^*\right\|_2 + 1,
\end{aligned}$$

with probability at least $1-\delta$, and

$$\left\|\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) - \nabla^2\mathcal{L}_N(\widehat{\theta} + s(\widetilde{\theta}^{(t-1)} - \widehat{\theta}))\right\|_2 \lesssim d^{3/2}\left\|\widetilde{\theta}^{(t-1)} - \widehat{\theta}\right\|_2,$$

and thus,

$$\left\|\widetilde{\theta}^{(t-1)} - \nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1}\nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) - \widehat{\theta}\right\|_2 \lesssim \left(\sqrt{\frac{d\log\frac{d}{\delta}}{N}} + \frac{d\log\frac{d}{\delta}}{N} + d^{3/2}\left\|\widetilde{\theta}^{(t-1)} - \theta^*\right\|_2 + 1\right) d^{3/2}\left\|\widetilde{\theta}^{(t-1)} - \widehat{\theta}\right\|_2^2.$$

To bound the second term, we have that

$$\begin{aligned}
&\left\|\left(\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1} - \nabla^2\mathcal{L}_1(\widetilde{\theta}^{(t-1)})^{-1}\right)\nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)})\right\|_2 \\
&\le \left\|\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1} - \nabla^2\mathcal{L}_1(\widetilde{\theta}^{(t-1)})^{-1}\right\|_2 \left\|\nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)})\right\|_2 \\
&= \left\|\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1} - \nabla^2\mathcal{L}_1(\widetilde{\theta}^{(t-1)})^{-1}\right\|_2 \left\|\nabla\mathcal{L}_N(\widetilde{\theta}^{(t-1)}) - \nabla\mathcal{L}_N(\widehat{\theta})\right\|_2 \\
&\le \left\|\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1} - \nabla^2\mathcal{L}_1(\widetilde{\theta}^{(t-1)})^{-1}\right\|_2 \int_0^1 \left\|\nabla^2\mathcal{L}_N(\widehat{\theta} + s(\widetilde{\theta}^{(t-1)} - \widehat{\theta}))\right\|_2 ds \left\|\widetilde{\theta}^{(t-1)} - \widehat{\theta}\right\|_2.
\end{aligned}$$

By the proof of Lemma F.11, we have that

$$\begin{aligned}
\left\|\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1} - \nabla^2\mathcal{L}_1(\widetilde{\theta}^{(t-1)})^{-1}\right\|_2 &\le \left\|\nabla^2\mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1} - \nabla^2\mathcal{L}^*(\theta^*)^{-1}\right\|_2 + \left\|\nabla^2\mathcal{L}_1(\widetilde{\theta}^{(t-1)})^{-1} - \nabla^2\mathcal{L}^*(\theta^*)^{-1}\right\|_2 \\
&\lesssim \sqrt{\frac{d\log\frac{d}{\delta}}{n}} + \frac{d\log\frac{d}{\delta}}{n} + d^{3/2}\left\|\widetilde{\theta}^{(t-1)} - \theta^*\right\|_2,
\end{aligned}$$

with probability at least $1-\delta$, and

$$\begin{aligned}
\left\|\nabla^2\mathcal{L}_N(\widehat{\theta} + s(\widetilde{\theta}^{(t-1)} - \widehat{\theta}))\right\|_2 &\le \left\|\nabla^2\mathcal{L}_N(\widehat{\theta} + s(\widetilde{\theta}^{(t-1)} - \widehat{\theta})) - \nabla^2\mathcal{L}^*(\theta^*)\right\|_2 + \left\|\nabla^2\mathcal{L}^*(\theta^*)\right\|_2 \\
&\lesssim \sqrt{\frac{d\log\frac{d}{\delta}}{N}} + \frac{d\log\frac{d}{\delta}}{N} + d^{3/2}\left(\left\|\widetilde{\theta}^{(t-1)} - \theta^*\right\|_2 + \left\|\widehat{\theta} - \theta^*\right\|_2\right) + 1 \\
&\lesssim d^{3/2}\left\|\widetilde{\theta}^{(t-1)} - \theta^*\right\|_2 + 1,
\end{aligned}$$

for $\delta \gtrsim e^{-N/d^4}$, provided that $N \gtrsim d^4 \log d$, and thus,

$$\begin{aligned}
&\left\| \left(\nabla^2 \mathcal{L}_N(\widetilde{\theta}^{(t-1)})^{-1} - \nabla^2 \mathcal{L}_1(\widetilde{\theta}^{(t-1)})^{-1} \right) \nabla \mathcal{L}_N(\widetilde{\theta}^{(t-1)}) \right\|_2 \\
&\lesssim \left(\sqrt{\frac{d \log \frac{d}{\delta}}{n}} + \frac{d \log \frac{d}{\delta}}{n} + d^{3/2} \left\| \widetilde{\theta}^{(t-1)} - \theta^* \right\|_2 \right) \left(d^{3/2} \left\| \widetilde{\theta}^{(t-1)} - \theta^* \right\|_2 + 1 \right) \left\| \widetilde{\theta}^{(t-1)} - \widehat{\theta} \right\|_2 .
\end{aligned}$$

Provided that $n \gtrsim d^4 \log d$ and $\delta \gtrsim e^{-n/d^4}$, we have $d^{3/2} \left\| \widetilde{\theta}^{(t-1)} - \theta^* \right\|_2 \lesssim 1$ for any $t \geq 1$, and then,

$$\left\| \widetilde{\theta}^{(t)} - \widehat{\theta} \right\|_2 \lesssim d^{3/2} \left\| \widetilde{\theta}^{(t-1)} - \widehat{\theta} \right\|_2^2 + \left(\sqrt{\frac{d \log \frac{d}{\delta}}{n}} + d^{3/2} \left\| \widetilde{\theta}^{(t-1)} - \theta^* \right\|_2 \right) \left\| \widetilde{\theta}^{(t-1)} - \widehat{\theta} \right\|_2 .$$

Since

$$\left\| \widetilde{\theta}^{(t-1)} - \theta^* \right\|_2 \leq \left\| \widetilde{\theta}^{(t-1)} - \widehat{\theta} \right\|_2 + \left\| \widehat{\theta} - \theta^* \right\|_2 \leq \left\| \widetilde{\theta}^{(t-1)} - \widehat{\theta} \right\|_2 + \sqrt{\frac{d \log \frac{d}{\delta}}{N}},$$

we obtain the bound in the lemma. □