Auxiliary Learning and its Statistical Understanding

[1] *Peking University;* [2] *Renmin University of China;*

[3] *Ministry of Commerce People's Republic of China.*

Supplementary Material

Appendix A: Verification Details

Appendix A.1: Verification for the form of $\mathbb{W}$

For the sake of simplicity, we write $\mathbb{W}_1 = \{w \in \mathbb{R}^{K+1} : Bw = \beta^{(0)}\}$ and $\mathbb{W}_2 = \{e_1 + \Theta u : u \in \mathbb{R}^{K-d+1}\}$. Then it suffices to show that $\mathbb{W}_1 = \mathbb{W}_2$. On one side, for any $w \in \mathbb{W}_2$, we can find u such that $w = e_1 + \Theta u$. We then have $Bw = B(e_1 + \Theta u) = \beta^{(0)} + B\Theta u$. In fact, we should have $B\Theta u = \mathbf{0}_p$; otherwise we have $u^\top \Theta^\top B^\top B\Theta u > 0$, which is a contradiction since $\Theta^\top B^\top B\Theta = \boldsymbol{O}_{K-d+1}$. Then $Bw = \beta^{(0)}$ and thus $w \in \mathbb{W}_1$, indicating $\mathbb{W}_2 \subset \mathbb{W}_1$. On the other side, for any $w \in \mathbb{W}_1$, we should have $B(w - e_1) = \mathbf{0}$. Recall that $\text{rank}(B) = d$, $\text{rank}(\Theta) = K + d - 1$, and $B\Theta = \boldsymbol{O}_{p\times(K+d-1)}$. Therefore, Θ is a basis of the null space for B. It follows that there exists $u \in \mathbb{R}^{K+d-1}$ such that $w - e_1 = \Theta u$. Therefore, $w \in \mathbb{W}_2$ and $\mathbb{W}_1 \subset \mathbb{W}_2$. Consequently, we have $\mathbb{W}_1 = \mathbb{W}_2$.

Appendix A.2: Derivation of w^*

We start with solving $w^* = \arg\min_{w\in\mathbb{W}} \mathcal{P}(w) = \arg\min_{w\in\mathbb{W}} w^\top \Sigma_\varepsilon w$. Recall that for any $w \in \mathbb{W}$, there exists some $u \in \mathbb{R}^{K+1-d}$ such that $w = e_1 + \Theta u$. Therefore, minimizing $w^\top \Sigma_\varepsilon w$ under the constraint $w \in \mathbb{W}$ is equivalent to minimizing $(e_1 + \Theta u)^\top \Sigma_\varepsilon (e_1 + \Theta u)$ with respect to $u \in \mathbb{R}^{K+1-d}$. Define $\mathcal{Q}(u) = (e_1 + \Theta u)^\top \Sigma_\varepsilon (e_1 + \Theta u)$, which is convex with respect to u. The first-order derivative of $\mathcal{Q}(u)$ is given by $\dot{\mathcal{Q}}(u) = \partial \mathcal{Q}(u)/\partial u = \Theta^\top \Sigma_\varepsilon e_1 + \Theta^\top \Sigma_\varepsilon \Theta u$. Setting this derivative to zero, we find $u^* = -(\Theta^\top \Sigma_\varepsilon \Theta)^{-1} \Theta^\top \Sigma_\varepsilon e_1$. Finally, substituting u^* into the expression for w, we obtain the optimal weight $w^* = e_1 - \Theta(\Theta^\top \Sigma_\varepsilon \Theta)^{-1} \Theta^\top \Sigma_\varepsilon e_1$.

Appendix B: The Proof of Theorem 1

Write $H = \Sigma_\varepsilon^{1/2}\Theta(\Theta^\top\Sigma_\varepsilon\Theta)^{-1}\Theta\Sigma_\varepsilon^{1/2}$ and $\widehat{H} = \widehat{\Sigma}_\varepsilon^{1/2}\widehat{\Theta}(\widehat{\Theta}\widehat{\Sigma}_\varepsilon\widehat{\Theta})^{-1}\widehat{\Theta}\widehat{\Sigma}_\varepsilon^{1/2}$. We then have $w^* = e_1 - \Sigma_\varepsilon^{-1/2}H\Sigma_\varepsilon^{1/2}e_1$ and $\widehat{w}^* = e_1 - \widehat{\Sigma}_\varepsilon^{-1/2}\widehat{H}\widehat{\Sigma}_\varepsilon^{1/2}e_1$. Then

$$\begin{aligned}
\|\widehat{\beta}_{\widehat{w}^*} - \widehat{\beta}_{w^*}\|_2 =& \|\widehat{B}(\widehat{w}^* - w^*)\|_{\text{op}} \le \|\widehat{B}\|_{\text{op}}\|\widehat{w}^* - w^*\|_{\text{op}} \\
\le& \|\widehat{B}\|_{\text{op}}\|\widehat{\Sigma}_\varepsilon^{-1/2}\widehat{H}\widehat{\Sigma}_\varepsilon^{1/2} - \Sigma_\varepsilon^{-1/2}H\Sigma_\varepsilon^{1/2}\|_{\text{op}} \\
\le& \|\widehat{B}\|_{\text{op}}\Big\{\|\widehat{\Sigma}_\varepsilon^{-1/2} - \Sigma_\varepsilon^{-1/2}\|_{\text{op}}\|\widehat{\Sigma}_\varepsilon^{1/2}\|_{\text{op}} + \tau_{\min}^{-1/2}\|\widehat{H} - H\|_{\text{op}}\|\widehat{\Sigma}_\varepsilon^{1/2}\| \\
&+ \tau_{\min}^{-1/2}\|\widehat{\Sigma}_\varepsilon^{1/2} - \Sigma_\varepsilon^{1/2}\|_{\text{op}}\Big\}.
\end{aligned}$$

Consequently, it suffices to prove the following inequalities, i.e., (B.1)—(B.4). Their detailed proofs are presented in Appendix C.

$$P\left(\|\widehat{B} - B\|_{\text{op}} > C_1\sqrt{\frac{p+K}{N}}\right) \le C_2 \exp\big\{-(p+K)\big\}, \tag{B.1}$$

$$P\left\{\|\widehat{\Sigma}_\varepsilon^{1/2} - \Sigma_\varepsilon^{1/2}\|_{\text{op}} > C_3\left(\sqrt{\frac{K}{N}} + \frac{p}{N}\right)\right\} \le 2\exp(-K), \tag{B.2}$$

$$P\left\{\|\widehat{\Sigma}_\varepsilon^{-1/2} - \Sigma_\varepsilon^{-1/2}\|_{\text{op}} > C_4\left(\sqrt{\frac{K}{N}} + \frac{p}{N}\right)\right\} \le 2\exp(-K), \tag{B.3}$$

$$P\left\{\|\widehat{H} - H\|_{\text{op}} > C_5\left(\sqrt{\frac{K+d}{N}} + \frac{p}{N}\right)\right\} \le C_6\exp(-K). \tag{B.4}$$

The detailed proof of (B.1) is given in Lemma 1. The results of (B.2) and (B.3) are given in Lemma 2. The inequality (B.4) is proved in Lemma 3.

Based on (B.1)—(B.4), we have $\|\widehat{B}\|_{\text{op}} \le \|\widehat{B} - B\|_{\text{op}} + \|B\|_{\text{op}} \le 2\|B\|_{\text{op}}$ as long as $N > N_0$ for some sufficiently large constant N_0. Similarly, we have $\|\widehat{\Sigma}_\varepsilon^{1/2}\| \le 2\|\Sigma_\varepsilon^{1/2}\|_{\text{op}}$ as $N > N_0$ for the same N_0. It follows that $\|\widehat{\beta}_{\widehat{w}^*} - \widehat{\beta}_{w^*}\|_2 \le 4\tau_{\max}^{3/2}\|\widehat{\Sigma}_\varepsilon^{-1/2} - \Sigma_\varepsilon^{-1/2}\| + \tau_{\min}^{-1/2}\tau_{\max}^{3/2}\|\widehat{H} - H\|_{\text{op}} +$

$\tau_{\min}^{-1/2}\tau_{\max}\|\widehat{\Sigma}_\varepsilon^{1/2} - \Sigma_\varepsilon^{1/2}\|_{\text{op}}$. Then by (B.2)—(B.4), we have

$$P\left\{\left\|\widehat{\beta}_{\widehat{w}^*} - \widehat{\beta}_{w^*}\right\|_{\text{op}} > C_7\left(\sqrt{\frac{K+d}{N}} + \frac{p}{N}\right)\right\} \le C_8 \exp(-K),$$

for some constants C_7 and C_8 as long as $N > N_0$ for the same constant N_0. This concludes the entire proof.

Appendix C: Some Useful Lemmas for Theorem 1

Lemma 1. *(Convergence Rate of $\widehat{B}$) Assume the same conditions as Theorem 1. Then $\|\widehat{B}-B\|_{\rm op} \le C_1\sqrt{(p+K)/N}$ holds with probability at least $1-C_2\exp\{-(p+K)\}$ as long as $N > N_0$ for some sufficiently large constant N_0. Here C_1 and C_2 are constants independent of p, K, and N.*

Proof. Note that $\widehat{B}-B = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathcal{E}$, where $\mathbb{X} = (X_1, ..., X_N)^\top \in \mathbb{R}^{N\times p}$, $\mathcal{E} = (\varepsilon^{(0)}, ..., \varepsilon^{(K)}) \in \mathbb{R}^{N\times(K+1)}$, and $\varepsilon^{(k)} = (\varepsilon_1^{(k)}, ..., \varepsilon_N^{(k)})^\top \in \mathbb{R}^N$. Then we have

$$\begin{aligned}\|\widehat{B}-B\|_{\rm op} &\le \|\widehat{\Sigma}_{xx}^{-1}\|_{\rm op}\|N^{-1}\mathbb{X}^\top\mathcal{E}\|_{\rm op}\\ &\le \Big\{\lambda_{\min}\left(\Sigma_{xx}\right) - \lambda_{\max}\big(\Sigma_{xx} - \widehat{\Sigma}_{xx}\big)\Big\}^{-1}\|N^{-1}\mathbb{X}^\top\mathcal{E}\|_{\rm op},\end{aligned}$$

where the last inequality is due to the fact that $\lambda_{\min}(A) \ge \lambda_{\min}(B) - \lambda_{\max}(B-A)$ for two arbitrary but symmetric matrices A and B. Then it sufficies to prove the following two inequalities

$$P\Big(\|\widehat{\Sigma}_{xx} - \Sigma_{xx}\|_{\rm op} > \tau_{\min}/2\Big) \le \mathcal{C}_1\exp(-\mathcal{C}_2 N), \tag{C.1}$$

$$P\Big(\|N^{-1}\mathbb{X}^\top\mathcal{E}\|_{\rm op} > t\Big) \le \mathcal{C}_3\exp\left\{2(p+K) - \mathcal{C}_4 N\min\left(\frac{t^2}{4C_{\rm sub}^4}, \frac{t}{2C_{\rm sub}^2}\right)\right\}. \tag{C.2}$$

The detailed proofs are given in the follwing Step 1 and Step 2. With the help of (C.1) and (C.2), we then have

$$\begin{aligned}P\Big(\|\widehat{B}-B\|_{\rm op} > \delta\Big) &\le P\Big(\|\widehat{\Sigma}_{xx} - \Sigma_{xx}\|_{\rm op} > \tau_{\min}/2\Big) + P\Big(2\tau_{\min}^{-1}\|N^{-1}\mathbb{X}\mathcal{E}\| > \delta\Big)\\ &\le 2\mathcal{C}_3\exp\Big\{2(p+K) - \mathcal{C}_4 N\min\big(\mathcal{C}_5^2\delta^2, \mathcal{C}_5\delta\big)\Big\}, \end{aligned}\tag{C.3}$$

where $\mathcal{C}_5 = \tau_{\min}/(4C_{\rm sub}^2)$ is a constant. Recall that by Condition (C4), $p/N \to 0$ holds as

$N \to \infty$. Therefore, we should have $p/N \leq \sqrt{p/N}$ as long as $N > N_0$ for the same constant N_0. Subsequenlt, we take

$$\delta = \frac{1}{\mathcal{C}_5} \max\left(\sqrt{\frac{3(p+K)}{\mathcal{C}_4 N}}, \frac{3(p+K)}{\mathcal{C}_4 N}\right) = \frac{1}{\mathcal{C}_5}\sqrt{\frac{3(p+K)}{\mathcal{C}_4 N}}.$$

Then (C.3) suggests that $\|\widehat{B} - B\|_{\text{op}} \leq C_1\sqrt{(p+K)/N}$ holds with probability at least $1 - C_2 \exp\big\{-(p+K)\big\}$, where $C_1 = 3/(\sqrt{\mathcal{C}_4}\mathcal{C}_5)$ and $C_2 = 2\mathcal{C}_3$ are constants. This leads to the conclusion of Lemma 1. We next verify the inequalities (C.1) and (C.2) in the following two steps.

Step 1: Proof of (C.1). By condition (C1), we know X_i is an independently and identically distributed sub-Gaussian random variable. Therefore, we can apply Theorem 6.5 of Wainwright (2019) and obtain

$$P\left\{C_{\text{sub}}^{-2}\left\|\widehat{\Sigma}_{xx} - \Sigma_{xx}\right\|_{\text{op}} \geq C_1\left(\sqrt{\frac{p}{N}} + \frac{p}{N}\right) + \epsilon\right\} \leq C_2 \exp\Big\{-C_3 N \min(\epsilon, \epsilon^2)\Big\}, \tag{C.4}$$

where C_1, C_2 and C_3 are some fixed constants. Next define $\epsilon = \tau_{\min}/(4C_{\text{sub}})$. Then by Condition (C2) that $p/N \to 0$ as $N \to \infty$, we should have $C_1 C_{\text{sub}}^2(\sqrt{p/N} + p/N) + C_{\text{sub}}^2\epsilon \leq \tau_{\min}/2$ as long as $N > N_0$ for some sufficiently large constant N_0. Therefore, by the inequality (C.4) we know that as long as $N > N_\delta$, we have

$$P\Big(\|\Sigma_{xx} - \widehat{\Sigma}_{xx}\|_{\text{op}} \geq \tau_{\min}/2\Big) \leq C_2 \exp(-C_4 N), \tag{C.5}$$

where $C_4 = C_3 \min\Big\{\tau_{\min}/(4C_{\text{sub}}), \tau_{\min}^2/(16C_{\text{sub}^4})\Big\}$ is a constant independent of N.

Step 2: Proof of (C.2). We consider an ε-net to bound the term $\|N^{-1}\mathbb{X}^\top\mathcal{E}\|_{\text{op}}^2$. Let $\varepsilon = 1/3$ and we can find two ε-nets $\mathcal{U}$ and $\mathcal{V}$ of the unit spheres $\mathcal{S}^{p-1}$ and $\mathcal{S}^K$ with cardinalities $|\mathcal{U}| \leq 7^p \leq e^{2p}$ and $|\mathcal{V}| \leq 7^{K+1} \leq e^{2(K+1)}$, respectively (Vershynin, 2018, Corollary 4.2.13). Then we have $\|N^{-1}\mathbb{X}^\top\mathcal{E}\|_{\text{op}} \leq 2\max_{u\in\mathcal{U},v\in\mathcal{V}} |N^{-1}(\mathcal{E}v)^\top(\mathbb{X}u)|$ (Vershynin, 2018, Lemma 4.4.1). Note that $N^{-1}(\mathcal{E}v)^\top(\mathbb{X}u) = N^{-1}\sum_{i=1}^N \widetilde{X}_i\widetilde{\varepsilon}_i$, where $\widetilde{X}_i = X_i^\top u \in \mathbb{R}$ and $\widetilde{\varepsilon}_i = \varepsilon_i^\top v \in \mathbb{R}$. Here $\widetilde{X}_i$ and $\widetilde{\varepsilon}_i$ are independent sub-Gaussian variables with $\|\widetilde{X}_i\|_{\psi_2} \leq C_{\text{sub}}$ and $\|\widetilde{\varepsilon}_i\|_{\psi_2} \leq C_{\text{sub}}$ by Condition (C1). We further note that $\widetilde{X}_i\widetilde{\varepsilon}_i$ are sub-exponential variables with $E(\widetilde{X}_i\widetilde{\varepsilon}_i) = \mathbf{0}$ and $\|\widetilde{X}_i\widetilde{\varepsilon}_i\|_{\psi_1} \leq \|\widetilde{X}_i\|_{\psi_2}\|\widetilde{\varepsilon}_i\|_{\psi_2} \leq C_{\text{sub}}^2$ (Vershynin, 2018, Lemma 2.7.7). Then for some fixed positive constant C_5 and C_6, we have

$$P\left(\left\|N^{-1}\mathbb{X}^\top\mathcal{E}\right\|_{\text{op}} > t\right) \leq |\mathcal{U}|\,|\mathcal{V}|\,P\left(\left|\frac{1}{N}\sum_{i=1}^N \widetilde{X}_i\widetilde{\varepsilon}_i\right| > \frac{t}{2}\right)$$
$$\leq C_5\exp\left\{2(p+K) - C_6 N\min\left(\frac{t^2}{4C_{\text{sub}}^4}, \frac{t}{2C_{\text{sub}}^2}\right)\right\}. \tag{C.6}$$

□

Lemma 2. *(Convergence Rate of $\widehat{\Sigma}_\varepsilon^{1/2}$ and $\widehat{\Sigma}_\varepsilon^{-1/2}$) Assume the same conditions as Theorem 1. We then have $\|\widehat{\Sigma}_\varepsilon^{1/2} - \Sigma_\varepsilon^{1/2}\|_{\text{op}} \leq C_1(\sqrt{K/N} + p/N)$ and $\|\widehat{\Sigma}_\varepsilon^{-1/2} - \Sigma_\varepsilon^{-1/2}\|_{\text{op}} \leq C_2(\sqrt{K/N} + p/N)$ hold with a probability at least $1 - 2\exp(-K)$ as long as $N > N_0$ for some sufficiently large constant N_0. Here C_1 and C_2 are constants independent of K, N, and N_0.*

Proof. We first argue the conclusion that $\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\|_{\text{op}} \leq C(\sqrt{K/N} + p/N)$ holds with a probability at least $1 - 2\exp(-K)$ as long as $N > N_0$. We prove this conclusion in Lemma 4. Then by Lemma 4, we should have $\widehat{\Sigma}_\varepsilon$ to be consistent as $N \to \infty$. Define $\Delta = \widehat{\Sigma}_\varepsilon^{1/2} - \Sigma_\varepsilon^{1/2}$, we should have $\|\Delta\|_{\text{op}} \to_p 0$ as $N \to \infty$. We next study the probabilistic upper bound for $\|\Delta\|_{\text{op}}$. Note that $\widehat{\Sigma}_\varepsilon = (\Sigma_\varepsilon^{1/2} + \Delta)(\Sigma_\varepsilon^{1/2} + \Delta) = \Sigma_\varepsilon + \Sigma_\varepsilon^{1/2}\Delta + \Delta\Sigma_\varepsilon^{1/2} + \Delta^2$. It follows that $\|\widehat{\Sigma}_\varepsilon^{-1/2} - \Sigma_\varepsilon^{-1/2}\|_{\text{op}} \leq C_0\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\|_{\text{op}}$ for some constant C_0. Then by Lemma 4, we should have $\|\Delta\|_{\text{op}} = \|\widehat{\Sigma}_\varepsilon^{1/2} - \Sigma_\varepsilon^{1/2}\|_{\text{op}} \leq C_1(\sqrt{K/N} + p/N)$ holds for some constant C_0 with a probability at least $1-2\exp(-K)$ as long as $N > N_0$. Further note that $\widehat{\Sigma}_{\varepsilon}^{-1/2} - \Sigma_{\varepsilon}^{-1/2} = (\widehat{\Sigma}_{\varepsilon}^{1/2} - \Sigma_{\varepsilon}^{1/2})\widehat{\Sigma}_{\varepsilon}^{-1} - \Sigma_{\varepsilon}^{-1/2}(\widehat{\Sigma}_{\varepsilon} - \Sigma_{\varepsilon})\widehat{\Sigma}_{\varepsilon}^{-1}$. This suggests that $\|\widehat{\Sigma}_{\varepsilon}^{-1/2} - \Sigma_{\varepsilon}^{-1/2}\|_{\text{op}} \le C_2(\sqrt{K/N} + p/N)$ holds with a probability at least $1-2\exp(-K)$ as long as $N > N_0$ for some sufficiently large constant N_0 and a constant C_2. □

Lemma 3. *(Convergence rate of $\widehat{H}$) Assume the same conditions as Theorem 1. Then $\|\widehat{H} - H\|_{\text{op}} \le C_1\Big\{\sqrt{(K+d)/N} + (p/N)\Big\}$ holds with a probability at least $1 - C_2\exp(-K)$ as long as $N > N_0$ for some sufficiently large constant N_0.*

Proof. We consider the symmetric matrices $M = \Sigma_{\varepsilon}^{-1/2} B^\top B \Sigma_{\varepsilon}^{-1/2}$ and $\widehat{M} = \widehat{\Sigma}_{\varepsilon}^{-1/2}(\widehat{B}^\top\widehat{B} - \sum_{k=d+1}^{K+1}\widehat{\lambda}_k\widehat{\nu}_k\widehat{\nu}_k^\top)\widehat{\Sigma}_{\varepsilon}^{-1/2}$, where $\widehat{\lambda}_k$ is the k-th largest eigenvalue of $\widehat{B}^\top\widehat{B}$ and $\widehat{\nu}_k$ is the associated eigenvector. Then we have $M\Sigma_{\varepsilon}^{1/2}\Theta = \boldsymbol{O}$ and $\widehat{M}\widehat{\Sigma}_{\varepsilon}^{1/2}\widehat{\Theta} = \boldsymbol{O}$, where $\boldsymbol{O}$ is a zero matrix. We conduct the eigenvalue decomposition on M and $\widehat{M}$ as $M = \sum_{k=1}^{K+1}\lambda_k^*\nu_k^*\nu_k^{*\top}$ and $\widehat{M} = \sum_{k=1}^{K+1}\widehat{\lambda}_k^*\widehat{\nu}_k^*\widehat{\nu}_k^{*\top}$, where λ_k^* and $\widehat{\lambda}_k^*$ are the k-th largest eigenvalues and ν_k^* and $\widehat{\nu}_k^*$ are the associated eigenvectors of M and $\widehat{M}$, respectively. Let $\Theta^* = (\nu_{d+1}^*, \ldots, \nu_{K+1}^*) \in \mathbb{R}^{(K+1)\times(K-d+1)}$ and $\widehat{\Theta}^* = (\widehat{\nu}_{d+1}^*, \ldots, \widehat{\nu}_{K+1}^*) \in \mathbb{R}^{(K+1)\times(K-d+1)}$. Then we have Θ^* and $\widehat{\Theta}^*$ as matrices with orthogonal columns, satisfying $M\Theta^* = \boldsymbol{O}$ and $\widehat{M}\widehat{\Theta}^* = \boldsymbol{O}$. Under the condition that $\widehat{\lambda}_d^* > 0$, we should have $\mathcal{S}(\Sigma_{\varepsilon}^{1/2}\Theta) = \mathcal{S}(\Theta^*)$ and $\mathcal{S}(\widehat{\Sigma}_{\varepsilon}^{1/2}\widehat{\Theta}) = \mathcal{S}(\widehat{\Theta}^*)$ with projection matrices H and $\widehat{H}$, respectively. Therefore, we have $\|H - \widehat{H}\|_{\text{op}} = \|\widehat{\Theta}^*\widehat{\Theta}^{*\top} - \Theta^*\Theta^{*\top}\|_{\text{op}} = \|\sin\theta(\widehat{\Theta}^*, \Theta^*)\|_{\text{op}}$; see Lemma 2.5 of Chen et al. (2021). Here $\sin\theta(\widehat{\Theta}^*, \Theta^*) = \text{diag}(\sin\theta_k^* : 0 \le k \le K) \in \mathbb{R}^{(K+1)\times(K+1)}$, where $\theta_k^* = \arccos(\sigma_k^*)$ and σ_k^* is the k-th largest singular value of $\widehat{\Theta}^{*\top}\Theta^*$. Note that $\lambda_d^* \ge \tau_{\min}\tau_{\max}^{-1}$ and $\widehat{\lambda}_k^* = 0$ for any $d < k \le K+1$. Under the condition that $\widehat{\lambda}_k^* > 0$ for $1 \le k \le d$, we can apply the Davis-Kahan Theorem (Chen et al., 2021, Theorem 2.7) as

$$\left\|\widehat{H} - H\right\|_{\text{op}} = \left\|\sin\theta\left(\widehat{\Theta}^*, \Theta^*\right)\right\|_{\text{op}} \le \left(\frac{\tau_{\max}}{\tau_{\min}}\right)\left\|\widehat{M} - M\right\|_{\text{op}}. \tag{C.7}$$

We write $\widehat{M} = \widehat{\Sigma}_\varepsilon^{-1/2}(\widehat{B}^\top \widehat{B} - \sum_{k=d+1}^{K+1} \widehat{\lambda}_k \widehat{\nu}_k \widehat{\nu}_k^\top)\widehat{\Sigma}_\varepsilon^{-1/2} = \widehat{\Sigma}_\varepsilon^{-1/2}\widehat{B}^\top \widehat{B}\widehat{\Sigma}_\varepsilon^{-1/2} - \widehat{\Sigma}_\varepsilon^{-1/2}\widehat{\Theta}\widehat{\Lambda}\widehat{\Theta}^\top \widehat{\Sigma}_\varepsilon^{-1/2}$, where $\widehat{\Lambda} = \mathrm{diag}(\widehat{\lambda}_k : d < k \leq K+1) \in \mathbb{R}^{(K+1)\times(K+1)}$. Then $\|\widehat{M} - M\|_{\mathrm{op}}$ can be further bounded by $M_1 + M_2$, where $M_1 = \|\widehat{\Sigma}_\varepsilon^{-1/2}\widehat{B}^\top \widehat{B}\widehat{\Sigma}_\varepsilon^{-1/2} - \Sigma_\varepsilon^{-1/2} B^\top B \Sigma_\varepsilon^{-1/2}\|_{\mathrm{op}}$ and $M_2 = \|\widehat{\Sigma}_\varepsilon^{-1/2}\widehat{\Theta}\widehat{\Lambda}\widehat{\Theta}^\top \widehat{\Sigma}_\varepsilon^{-1/2}\|_{\mathrm{op}}$.

For the term M_1, simple algebra suggests that $M_1 \leq M_{11} + M_{12}$, where

$$M_{11} = \left\|\widehat{\Sigma}_\varepsilon^{-1/2}\right\|_{\mathrm{op}}^2 \left\|\widehat{B}^\top \widehat{B} - B^\top B\right\|_{\mathrm{op}},$$
$$M_{12} = \tau_{\max}\left(\left\|\widehat{\Sigma}_\varepsilon^{-1/2}\right\|_{\mathrm{op}} + \left\|\Sigma_\varepsilon^{-1/2}\right\|_{\mathrm{op}}\right)\left\|\widehat{\Sigma}_\varepsilon^{-1/2} - \Sigma_\varepsilon^{-1/2}\right\|_{\mathrm{op}}.$$

Furthermore, note that $M_2 \leq \|\widehat{\Sigma}_\varepsilon^{-1/2}\|_{\mathrm{op}}^2 |\widehat{\lambda}_{d+1}|$, which is due to the fact that $\|\widehat{V}\|_{\mathrm{op}} = \widehat{\lambda}_{d+1}$. Recall that $\widehat{\lambda}_d \geq 0$ and $\lambda_{d+1} = 0$. Then by Weyl's inequality, we should have $\widehat{\lambda}_{d+1} = |\widehat{\lambda}_{d+1} - \lambda_{d+1}| \leq \|\widehat{B}^\top \widehat{B} - B^\top B\|_{\mathrm{op}}$ (Chen et al., 2021, Lemma 2.2). Consequently, we have $M_2 \leq \|\widehat{\Sigma}_\varepsilon^{-1/2}\|_{\mathrm{op}}^2 \|\widehat{B}^\top \widehat{B} - B^\top B\|_{\mathrm{op}}$. Thus we have $\|\widehat{M} - M\|_{\mathrm{op}} \leq 2M_{11} + M_{12}$. Next by Lemma 2, we should have

$$M_{11} \leq M_{11}' = 4\|\widehat{\Sigma}_\varepsilon^{-1/2}\|_{\mathrm{op}}^2 \|\widehat{B}^\top \widehat{B} - B^\top B\|_{\mathrm{op}},$$
$$M_{12} \leq M_{12}' = 3\|\Sigma_\varepsilon^{-1/2}\|_{\mathrm{op}} \|\widehat{\Sigma}_\varepsilon^{-1/2} - \Sigma_\varepsilon^{-1/2}\|_{\mathrm{op}},$$

as long as $N > N_0$ for some sufficiently large constant N_0. Then by Lemma 2 and Lemma

5, we obtain the probabilistic upper bound for $\|\widehat{H} - H\|_{\rm op}$ as

$$P\Bigg\{\|\widehat{H} - H\|_{\rm op} > \frac{4C_1\tau_{\max}}{\tau_{\min}^2}\left(\sqrt{\frac{K+d}{N}} + \frac{p}{N}\right) + \frac{3C_2\tau_{\max}}{\tau_{\min}^{3/2}}\left(\sqrt{\frac{K}{N}} + \frac{p}{N}\right)\Bigg\}$$
$$\leq P\left\{\|\widehat{B}^\top\widehat{B} - B^\top B\|_{\rm op} > C_1\left(\sqrt{\frac{K+d}{N}} + \frac{p}{N}\right)\right\}$$
$$+P\left\{\|\widehat{\Sigma}_\varepsilon^{-1/2} - \Sigma_\varepsilon^{-1/2}\|_{\rm op} > C_2\left(\sqrt{\frac{K}{N}} + \frac{p}{N}\right)\right\} \leq C_3\exp(-K).$$

Consequently, we should have $P\Big[\|\widehat{H} - H\|_{\rm op} > C_4\{\sqrt{(K+d)/N} + (p/N)\}\Big] \leq C_3\exp(-K)$ holds as long as $N > N_0$ for the same constant N_0 as mentioned before. □

Lemma 4. *(Convergence Rate of $\widehat{\Sigma}_\varepsilon$.) Assume the same conditions as Theorem 1. We then have $\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\|_{\rm op} \leq C(\sqrt{K/N} + p/N)$ holds with a probability at least $1 - 2\exp(-K)$ as long as $N > N_0$ for some sufficiently large constant N_0. Here C is a constant independent of K, N, and N_0.*

Proof. Recall that $\widehat{\Sigma}_\varepsilon = N^{-1}(\mathbb{Y} - \mathbb{X}\widehat{B})^\top(\mathbb{Y} - \mathbb{X}\widehat{B}) = N^{-1}\mathcal{E}^\top(I_N - H_{\mathbb{X}})\mathcal{E} \in \mathbb{R}^{(K+1)\times(K+1)}$, where $\mathbb{Y} = (Y_i^{(0)}, ..., Y_i^{(K)}) \in \mathbb{R}^{N\times(K+1)}$, and $H_{\mathbb{X}} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}$. It follows that $E(\widehat{\Sigma}_\varepsilon) = \{(N-p)/N\}\Sigma_\varepsilon$. Therefore, we should have $\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\|_{\rm op} \leq \|\widehat{\Sigma}_\varepsilon - E(\widehat{\Sigma}_\varepsilon)\|_{\rm op} + \tau_{\max}p/N$. Then it suffices to prove the following inequality as

$$P\Big\{\|\widehat{\Sigma}_\varepsilon - E(\widehat{\Sigma}_\varepsilon)\|_{\rm op} > \delta\Big\} \leq 2\exp\left\{4K - C_1 N\min\left(\frac{\delta^2}{4C_{\rm sub}^4}, \frac{\delta}{2C_{\rm sub}}\right)\right\}. \tag{C.8}$$

Recall that by Condition (C1), $p/N \to 0$ as $N \to \infty$. Assume $p/N \leq \sqrt{p/N}$ holds for any $N > N_0$ with a sufficiently large constant N_0. Next, in order to apply (C.8), we set

$$\delta = 2C_{\rm sub}\max\left(\sqrt{\frac{5K}{C_1 N}}, \frac{5K}{C_1 N}\right) = 2C_{\rm sub}\sqrt{\frac{5K}{C_1 N}}.$$

It follows that $P(\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\|_{\text{op}} \geq C_2\sqrt{K/N}) \leq \exp(-K)$, where $C_2 = 2\sqrt{5}C_{\text{sub}}/\sqrt{C_1}$ is a constant. Consequently, taking $C_3 = \max(C_2, \tau_{\min})$, we can prove that $\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\|_{\text{op}} \leq C_3(\sqrt{K/N} + p/N)$ holds with probability at least $1 - 2\exp(-K)$. We then verify (C.8) as follows.

Proof of (C.8). We consider an ε-net $\mathcal{U}$ on the unit sphere with $\varepsilon = 1/3$. It follows that $|\mathcal{U}| \leq e^{2(K+1)}$. We then have $\|\widehat{\Sigma}_\varepsilon - E(\widehat{\Sigma}_\varepsilon)\|_{\text{op}} \leq 2\max_{u,v\in\mathcal{U}} |u^\top\{\widehat{\Sigma}_\varepsilon - E(\widehat{\Sigma}_\varepsilon)\}v|$. Note that $\|u^\top\widehat{\Sigma}_\varepsilon v\|_{\psi_1} \leq \|N^{-1}u^\top\mathcal{E}^\top\mathcal{E}v\|_{\psi_1} \leq C_{\text{sub}}^2$, which suggests that $u^\top\widehat{\Sigma}_\varepsilon v$ is a sub-exponential variable. Then the Hanson-Wright inequality can be applied to obtain the upper bound as

$$P\Big\{\|\widehat{\Sigma}_\varepsilon - E(\widehat{\Sigma}_\varepsilon)\|_{\text{op}} > \delta\Big\} \leq \sum_{u,v\in\mathcal{U}} P\left[\left|u^\top\{\widehat{\Sigma}_\varepsilon - E(\widehat{\Sigma}_\varepsilon)\}v\right| > \frac{\delta}{2}\right]$$
$$\leq 2\exp\left\{4K - C_1 N \min\left(\frac{\delta^2}{4C_{\text{sub}}^4}, \frac{\delta}{2C_{\text{sub}}}\right)\right\}.$$

This concludes the entire proof. ☐

Lemma 5. *(Convergence Rate of $\widehat{B}^\top\widehat{B}$) Assume the same conditions as Theorem 1. We then have $\|\widehat{B}^\top\widehat{B} - B^\top B\|_{\text{op}} \leq C_1\Big\{\sqrt{(K+d)/N} + p/N\Big\}$ holds with a probability at least $1 - C_2\exp(-K)$ as long as $N > N_0$ for some sufficiently large constant N_0.*

Proof. Note that $\|\widehat{B}^\top\widehat{B} - B^\top B\|_{\text{op}} \leq 2\|B^\top(\widehat{B} - B)\|_{\text{op}} + \|\widehat{B} - B\|_{\text{op}}^2$, where the probabilistic upper bound for $\|\widehat{B} - B\|_{\text{op}}^2$ can be obtained from Lemma 1. Therefore, it sufficies to prove the following inequality

$$P\Big\{\|B^\top(\widehat{B} - B)\|_{\text{op}} > \delta\Big\} \leq C_1\exp\Big\{(K+d) - C_2 N\min(C_3^2\delta^2, C_3\delta)\Big\} \tag{C.9}$$

for some constants C_1, C_2, and C_3 as long as $N > N_0$ for a sufficiently large constant N_0.

To apply (C.9), we take

$$\delta = \frac{1}{C_3} \max\left(\sqrt{\frac{2(K+d)}{C_2 N}}, \frac{2(K+d)}{C_2 N}\right) = \frac{1}{C_3}\sqrt{\frac{2(K+d)}{C_2 N}},$$

where the last equation is due to Condition (C4) that $(p/N) \leq \sqrt{p/N}$ as long as $N > N_0$. It follows that

$$P\left(\|B^\top(\widehat{B} - B)\|_{\text{op}} > C_4\sqrt{\frac{K+d}{N}}\right) \leq C_5 \exp(-K)$$

as long as $N > N_0$. Then by the conclusion of Lemma 1, we should have

$$P\left(\|\widehat{B}^\top \widehat{B} - B^\top B\|_{\text{op}} > 2C_4\sqrt{\frac{K+d}{N}} + \frac{C_6^2(p+K)}{N}\right)$$
$$\leq P\left(\|B^\top(\widehat{B} - B)\|_{\text{op}} > C_4\sqrt{\frac{K+d}{N}}\right) + P\left(\|\widehat{B} - B\|_{\text{op}} > C_6\sqrt{\frac{p+K}{N}}\right)$$
$$\leq C_5 \exp(-K) + C_7 \exp(-p).$$

Consequently, we should have $\|\widehat{B}^\top \widehat{B} - B^\top B\|_{\text{op}} \leq C_8\Big\{\sqrt{(K+d)/N} + p/N\Big\}$ holds with probability at least $1 - C_9 \exp(-K)$. This leads to the desired conclusion. Then it only sufficies to prove (C.9).

Proof of (C.9). Note that $\|B^\top(\widehat{B} - B)\|_{\text{op}} = \|N^{-1}B^\top\widehat{\Sigma}_\varepsilon^{-1}\mathbb{X}^\top\mathcal{E}\|_{\text{op}} \leq \|N^{-1}B^\top(\widehat{\Sigma}_\varepsilon^{-1} - \Sigma_\varepsilon^{-1})\mathbb{X}^\top\mathcal{E}\|_{\text{op}} + \|N^{-1}B^\top\Sigma_\varepsilon^{-1}\mathbb{X}^\top\mathcal{E}\|_{\text{op}}$. We first study the term $\|N^{-1}B^\top\Sigma_\varepsilon^{-1}\mathbb{X}^\top\mathcal{E}\|_{\text{op}}$. Recall that $\nu_k \in \mathbb{R}^{K+1}$ is the eigenvector corresponding to the k-th largest eigenvalue of $B^\top B$. We then consider two ε-nets $\mathcal{U}$ and $\mathcal{V}$ of the set $\{\|x\| = 1 : x \in \mathcal{S}(\nu_k : 1 \leq k \leq d)\}$ and the unit sphere $\mathcal{S}^K$ with $\varepsilon = 1/3$. It follows that $|\mathcal{U}| \leq e^{2d}$ and $|\mathcal{V}| \leq e^{2(K+1)}$. Note that $\|N^{-1}B^\top\Sigma_\varepsilon^{-1}\mathbb{X}^\top\mathcal{E}\|_{\text{op}} \leq 2\sup_{u\in\mathcal{U},v\in\mathcal{V}} |N^{-1}u^\top B^\top\Sigma_\varepsilon^{-1}\mathbb{X}^\top\mathcal{E}v|$. Further write $t = \Sigma_\varepsilon^{-1}Bu/\|\Sigma_\varepsilon^{-1}Bu\|$ with $\|t\| = 1$. Then $\|N^{-1}B^\top\Sigma_\varepsilon^{-1}\mathbb{X}^\top\mathcal{E}\|_{\text{op}}$ can be further bounded by

$2\tau_{\max}\tau_{\min}^{-1}\sup_{u\in\mathcal{U},v\in\mathcal{V}} N^{-1}t^\top\mathbb{X}^\top\mathcal{E}v$. Subsequenltly, by Hanson-Wright's inequality, we have

$$
\begin{aligned}
P\Big(\big\|N^{-1}B^\top\Sigma_\varepsilon^{-1}\mathbb{X}^\top\mathcal{E}\big\|_{\mathrm{op}} > \delta\Big) &\le |\mathcal{U}||\mathcal{V}|P\left(N^{-1}t^\top\mathbb{X}^\top\mathcal{E}v > \frac{\tau_{\min}\delta}{2\tau_{\max}}\right)\\
&\le \mathcal{C}_1\exp\left\{2(K+d) - \mathcal{C}_2 N\min\left(\frac{\tau_{\min}^2\delta^2}{4\tau_{\max}^2C_{\mathrm{sub}}^2}, \frac{\tau_{\min}\delta}{2\tau_{\max}C_{\mathrm{sub}}}\right)\right\},
\end{aligned}
$$

where $\mathcal{C}_1 = 2e^2$. By Lemma 4 that $\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\|_{\mathrm{op}} \to_p 0$ as $N \to \infty$, it follows that $\|N^{-1}B^\top(\widehat{\Sigma}_\varepsilon^{-1} - \Sigma_\varepsilon^{-1})\mathbb{X}^\top\mathcal{E}\|_{\mathrm{op}} \le \|N^{-1}B^\top\Sigma_\varepsilon^{-1}\mathbb{X}^\top\mathcal{E}\|_{\mathrm{op}}$ holds almost surely as $N \to \infty$. Therefore, as long as $N > N_0$ for some constant N_0, we should have

$$
\begin{aligned}
&P\Big\{\|B^\top(\widehat{B} - B)\|_{\mathrm{op}} > \delta\Big\}\\
&\le \mathcal{C}_3\exp\left\{2(K+d) - \mathcal{C}_2 N\min\left(\frac{\tau_{\min}^2\delta^2}{4\tau_{\max}^2C_{\mathrm{sub}}^2}, \frac{\tau_{\min}\delta}{2\tau_{\max}C_{\mathrm{sub}}}\right)\right\}
\end{aligned}
$$

for some constants $\mathcal{C}_2$ and $\mathcal{C}_3$. □

References

Chen, Y., Y. Chi, J. Fan, C. Ma, et al. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning 14*(5), 566–806.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge university press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge university press.