ONLINE SUPPLEMENTARY MATERIAL FOR

Multiple Descent in the Multiple Random Feature Model

By Xuran Meng, Jianfeng Yao and Yuan Cao

The supplementary material is organized as follows:

- In Section I, we give the proof of Proposition A.2 which introduces the decomposition of the asymptotic excess risk.
- In Section II, we present the proof of Proposition A.4 by showing how the key terms in the risk decomposition are related to the logarithmic potential of the linear pencil matrix.
- In Section III, we establish basic properties of the fixed point equation (A.2) to justify the definition of $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$ below Definition A.5.
- In Section IV, we provide the proof of Proposition A.6, which extends the definition of $m(\xi;\mathbf{q},\boldsymbol{\mu})$ to $\mathbb{C}_+$ and shows that it is the asymptotic limit of $M_d(\xi;\mathbf{q},\boldsymbol{\mu})$ as $d\to\infty$.
- In Section V, we give the proof of Proposition A.7 by relating the logarithmic potential $G_d(\xi;\mathbf{q},\boldsymbol{\mu})$ to $g(\xi;\mathbf{q},\boldsymbol{\mu})$ in (A.3).
- In Section VI, we prove Proposition A.8 to justify the definition of $\boldsymbol{\nu}(\xi)$ as the unique solution to the system (3.1).
- In Section VII, we display the proof of the lemmas and propositions given in Appendix C.

I. Proof of Proposition A.2

Proposition A.2 gives a decomposition of the risk $R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\boldsymbol{\beta}_d,\boldsymbol{\varepsilon})$. To prove this decomposition, we first introduce some additional notations and preliminary lemmas.

Definition I.1. *Define*

$$\mathbf{V}_0(F_0)=F_0\mathbb{E}_{\mathbf{x}}\big[\boldsymbol{\sigma}(\mathbf{x})\big]\in\mathbb{R}^{N\times 1},\ \mathbf{V}(\boldsymbol{\beta}_{1,d})=\mathbb{E}_{\mathbf{x}}\big[\boldsymbol{\sigma}(\mathbf{x})\mathbf{x}^\top\boldsymbol{\beta}_{1,d}\big]\in\mathbb{R}^{N\times 1},\ \mathbf{U}=\mathbb{E}_{\mathbf{x}}\big[\boldsymbol{\sigma}(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})^\top\big]\in\mathbb{R}^{N\times N},$$

where $\mathbf{x}$ *is a random vector uniformed distributed on the sphere* $\sqrt{d}\cdot\mathbb{S}^{d-1}$ *and* $\boldsymbol{\sigma}(\mathbf{x})$ *is defined in Definition A.1.* □

Note that by the definition of $\boldsymbol{\sigma}(\mathbf{x})$ in Definition A.1, $\boldsymbol{\sigma}(\mathbf{x})$ also depends on the random feature parameter matrix $\boldsymbol{\Theta}$. Therefore, $\mathbf{V}_0(F_0)$, $\mathbf{V}(\boldsymbol{\beta}_{1,d})$ and $\mathbf{U}$ also depends on $\boldsymbol{\Theta}$. Now with these notations, and by the definition of $\widehat{\mathbf{a}}$ in (A.1), we can rewrite the risk as follows:

$$\begin{aligned}R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\boldsymbol{\beta}_d,\boldsymbol{\varepsilon})=&\mathbb{E}_{\mathbf{x}}\big[\mathbf{x}^\top\boldsymbol{\beta}_{1,d}+F_0-\widehat{\mathbf{a}}^\top\boldsymbol{\sigma}(\mathbf{x})\big]^2\\=&F_0^2+F_{1,d}^2-2\mathbf{y}^\top\mathbf{Z}\boldsymbol{\Upsilon}[\mathbf{V}(\boldsymbol{\beta}_{1,d})+\mathbf{V}_0(F_0)]/\sqrt{d}+\mathbf{y}^\top\big[\mathbf{U}\big]_{\mathbf{Z}}\mathbf{y}/d.\end{aligned}\tag{I.1}$$

Therefore, to prove Proposition A.2, it suffices to further decompose the terms $\mathbf{U},\mathbf{V}(\boldsymbol{\beta}_{1,d})$ and $\mathbf{V}_0(F_0)$. To handle these terms, we consider the Gegenbauer decomposition (Hua, 1963) of the nonlinear activation functions. For $j=1,2$, let $\lambda_{d,k}(\sigma_j)$ be the coefficients of the Gegenbauer

decomposition of σ_j, i.e.,

$$\sigma_j(x) = \sum_{k=0}^{+\infty} \lambda_{d,k}(\sigma_j) B(d,k) \cdot Q_k^{(d)}(\sqrt{d} \cdot x),$$

where $B(d,0) = 1$, $B(d,k) = k^{-1}(2k+d-2)\binom{k+d-3}{k-1}$ with $k \geq 1$, and $Q_k^{(d)}$, $k \in \mathbb{N}$ are the Gegenbauer polynomials forms an orthogonal basis on $L^2([-d,d], \tau_d)$. τ_d is the distribution of $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle$ where $\mathbf{x}_1, \mathbf{x}_2 \sim \sqrt{d} \cdot \mathrm{Unif}(\mathbb{S}^{d-1})$. Then define

$$\mathbf{\Lambda}_{d,k} = \mathrm{diag}\big(\lambda_{d,k}(\sigma_1)\mathbf{I}_{N_1}, \lambda_{d,k}(\sigma_2)\mathbf{I}_{N_2}\big), \quad k \in \mathbb{N} = \{0, 1, ...\}. \tag{I.2}$$

The following lemma decomposes the three terms in Definition I.1.

Lemma I.2. *With $\mathbf{M}_1$ and $\mathbf{M}_*$ in Definition A.1, and $\mathbf{\Lambda}_{d,k}$ in equation (I.2), we have*

$$\begin{aligned}
&\mathbf{V}_0(F_0) = F_0 \mathbf{\Lambda}_{d,0} \mathbf{1}_N, \\
&\mathbf{V}(\boldsymbol{\beta}_{1,d}) = \mathbf{\Lambda}_{d,1} \mathbf{\Theta} \boldsymbol{\beta}_{1,d} = \left(\frac{\mathbf{M}_1 + \mathbf{\Delta}'}{\sqrt{d}} \right) \mathbf{\Theta} \boldsymbol{\beta}_{1,d}, \\
&\mathbf{U} = \mathbf{\Lambda}_{d,0} \mathbf{1}_N \mathbf{1}_N^\top \mathbf{\Lambda}_{d,0} + \mathbf{M}_1 \frac{\mathbf{\Theta}\mathbf{\Theta}^\top}{d} \mathbf{M}_1 + \mathbf{M}_* \mathbf{M}_* + \mathbf{\Delta}.
\end{aligned}$$

where the remainder matrices $\mathbf{\Delta}, \mathbf{\Delta}'$ satisfy $\mathbb{E}\|\mathbf{\Delta}\|_{\mathrm{op}}^2 \vee \mathbb{E}\|\mathbf{\Delta}'\|_{\mathrm{op}}^2 = o_d(1)$.

Lemma I.2 is proved in Section I.1. Plugging the decompositions in Lemma I.2 into (I.1) will then give a decomposition of the risk consisting of multiple terms. The next lemma establishes useful moment estimations for some of the terms in (I.1), which helps us get rid of the negligible terms in the decomposition.

Lemma I.3. *For any fixed $k \in \mathbb{N}\backslash\{0\}$, let $\mathbf{\Gamma}_1 \in \mathbb{R}^{N \times N}$ and $\mathbf{\Gamma}_2 \in \mathbb{R}^{n \times n}$ be symmetric random matrices with $\big[\mathbb{E}\|\mathbf{\Gamma}_j\|_{\mathrm{op}}^k\big]^{1/k} = O_d(1)$, $j = 1, 2$. Define*

$$\begin{aligned}
\mathcal{B} =& \frac{1}{d} \mathbf{1}_n^\top \big[\mathbf{\Gamma}_1\big]_{\mathbf{Z}} \mathbf{1}_n, \\
\mathcal{C} =& 1 - \frac{2}{\sqrt{d}} \mathrm{tr}(\mathbf{\Lambda}_{d,0} \mathbf{1}_N \mathbf{1}_n^\top \mathbf{Z} \mathbf{\Upsilon}) + \frac{1}{d} \mathbf{1}_n^\top \big[\mathbf{\Lambda}_{d,0} \mathbf{1}_N \mathbf{1}_N^\top \mathbf{\Lambda}_{d,0}\big]_{\mathbf{Z}} \mathbf{1}_n, \\
\mathcal{D} =& \frac{1}{d} \mathrm{tr}(\big[\mathbf{\Lambda}_{d,0} \mathbf{1}_N \mathbf{1}_N^\top \mathbf{\Lambda}_{d,0}\big]_{\mathbf{Z}} \mathbf{\Gamma}_2),
\end{aligned}$$

where $\mathbf{\Lambda}_{d,0}$ is defined in equation (I.2). Then if $\sum_j \mu_{j,0}^2 > 0$, for any fixed $\lambda > 0$, there exists a constant $C > 0$ such that

$$\big(\mathbb{E}|\mathcal{B}|^k\big)^{1/k} \vee \mathbb{E}|\mathcal{C}| \vee \big(\mathbb{E}|\mathcal{D}|^k\big)^{1/k} = O_d\Big(d^{-1} e^{C\sqrt{\log d}}\Big) = o_d(1).$$

If $\sum_j \mu_{j,0}^2 = 0$, it still holds that $\big(\mathbb{E}|\mathcal{D}|^k\big)^{1/k} = o_d(1)$.

The proof of the lemma is given in Section I.2. To further decompose and calculate the risk, we also need to study the impact of fixed vector $\boldsymbol{\beta}_{1,d}$ on the risk. To do so, we aim to show that the risk only depends on $F_{1,d}$ $(= \|\boldsymbol{\beta}_{1,d}\|_2)$ due to rotation invariance of the learning problem. The result is given in the following lemma.

Lemma I.4. *Suppose* $\widetilde{\boldsymbol{\beta}}_{1,d} \sim \mathrm{Unif}(F_{1,d} \cdot \mathbb{S}^{d-1})$ *is independent of* $(\mathbf{X}, \boldsymbol{\Theta}, \boldsymbol{\varepsilon})$, *and denote* $\widetilde{\boldsymbol{\beta}}_d = [F_0, \widetilde{\boldsymbol{\beta}}_{1,d}^\top]^\top$. *Then for any fixed* $\boldsymbol{\beta}_{1,d}$, *under the assumptions of Proposition A.2, we have*

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}}\big|R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\boldsymbol{\beta}_d,\boldsymbol{\varepsilon}) - \overline{R}_d(\mathbf{X},\boldsymbol{\Theta},\lambda,F_{1,d},\tau)\big| \\
&\qquad\qquad = \mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},\widetilde{\boldsymbol{\beta}}_d}\big|R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}) - \overline{R}_d(\mathbf{X},\boldsymbol{\Theta},\lambda,F_{1,d},\tau)\big|, \\
&\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\big[\,\mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}}(R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}))\big] = o_d(1).
\end{aligned}$$

The proof of the lemma is given in Section I.3. Based on the above lemmas, we are ready to present the proof of Proposition A.2 as follows.

Proof [Proof of Proposition A.2] Let $\widetilde{\boldsymbol{\beta}}_d = [F_0, \widetilde{\boldsymbol{\beta}}_{1,d}^\top]^\top$ with $\widetilde{\boldsymbol{\beta}}_{1,d} \sim \mathrm{Unif}(F_{1,d} \cdot \mathbb{S}^{d-1})$. Then we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}}\big|R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\boldsymbol{\beta}_d,\boldsymbol{\varepsilon}) - \overline{R}_d(\mathbf{X},\boldsymbol{\Theta},\lambda,F_{1,d},\tau)\big| \\
&\quad= \mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},\widetilde{\boldsymbol{\beta}}_d}\big|R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}) - \overline{R}_d(\mathbf{X},\boldsymbol{\Theta},\lambda,F_{1,d},\tau)\big| \\
&\quad\le \mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},\widetilde{\boldsymbol{\beta}}_d}\big|R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}) - \mathbb{E}_{\boldsymbol{\varepsilon},\widetilde{\boldsymbol{\beta}}_d}R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon})\big| \\
&\qquad+ \mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\big|\mathbb{E}_{\boldsymbol{\varepsilon},\widetilde{\boldsymbol{\beta}}_d}R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}) - \overline{R}_d(\mathbf{X},\boldsymbol{\Theta},\lambda,F_{1,d},\tau)\big| \\
&\quad\le \mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\Big[\sqrt{\mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}}(R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}))}\Big] \\
&\qquad+ \mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\big|\mathbb{E}_{\boldsymbol{\varepsilon},\widetilde{\boldsymbol{\beta}}_d}R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}) - \overline{R}_d(\mathbf{X},\boldsymbol{\Theta},\lambda,F_{1,d},\tau)\big| \\
&\quad\le \sqrt{\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\big[\,\mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}}(R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}))\big]} \\
&\qquad+ \mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\big|\mathbb{E}_{\boldsymbol{\varepsilon},\widetilde{\boldsymbol{\beta}}_d}R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}) - \overline{R}_d(\mathbf{X},\boldsymbol{\Theta},\lambda,F_{1,d},\tau)\big| \\
&\quad= o_d(1) + \mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\big|\mathbb{E}_{\boldsymbol{\varepsilon},\widetilde{\boldsymbol{\beta}}_d}R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}) - \overline{R}_d(\mathbf{X},\boldsymbol{\Theta},\lambda,F_{1,d},\tau)\big|,
\end{aligned}$$

where the first equality follows by Lemma I.4, the first inequality follows by triangle inequality, the second and third inequalities are by Jensen's inequality, and the last equality follows by Lemma I.4 again. Therefore, to prove the proposition, it suffices to show that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\big|\mathbb{E}_{\boldsymbol{\varepsilon},\widetilde{\boldsymbol{\beta}}_d}R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}) - \overline{R}_d(\mathbf{X},\boldsymbol{\Theta},\lambda,F_{1,d},\tau)\big| = o_d(1).$$

Similar to (I.1), we have

$$R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}) = F_0^2 + F_{1,d}^2 - \frac{2\widetilde{\mathbf{y}}^\top\mathbf{Z}\boldsymbol{\Upsilon}(\mathbf{V}(\widetilde{\boldsymbol{\beta}}_{1,d}) + \mathbf{V}_0(F_0))}{\sqrt{d}} + \frac{\widetilde{\mathbf{y}}^\top[\mathbf{U}]_{\mathbf{Z}}\widetilde{\mathbf{y}}}{d}, \tag{I.3}$$

where $\widetilde{\mathbf{y}} = \mathbf{1}_n F_0 + \mathbf{X}\widetilde{\boldsymbol{\beta}}_{1,d} + \boldsymbol{\varepsilon}$. From Lemma I.2, we further have

$$\mathbf{V}_0(F_0)F_0 = F_0^2\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N, \quad \mathbb{E}_{\widetilde{\boldsymbol{\beta}}_{1,d}}\big(\mathbf{V}(\widetilde{\boldsymbol{\beta}}_{1,d})\widetilde{\boldsymbol{\beta}}_{1,d}^\top\big) = F_{1,d}^2\bigg(\frac{\mathbf{M}_1 + \boldsymbol{\Delta}'}{\sqrt{d}}\bigg)\frac{\boldsymbol{\Theta}}{d}, \tag{I.4}$$

and

$$\mathbf{U} = \boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0} + \mathbf{M}_1\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\top}{d}\mathbf{M}_1 + \mathbf{M}_*\mathbf{M}_* + \boldsymbol{\Delta}. \tag{I.5}$$

By (I.3), (I.4), (I.5) and the definition of $\overline{R}_d(\mathbf{X}, \mathbf{\Theta}, \lambda, F_{1,d}, \tau)$, we obtain the following equation with direct calculation:

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\boldsymbol{\beta}}_d, \boldsymbol{\varepsilon}} R_d(\mathbf{X}, \mathbf{\Theta}, \lambda, \widetilde{\boldsymbol{\beta}}_d, \boldsymbol{\varepsilon}) - \overline{R}_d(\mathbf{X}, \mathbf{\Theta}, \lambda, F_{1,d}, \tau) = \\
&\quad \underbrace{F_0^2 - \frac{2F_0^2}{\sqrt{d}} \mathrm{tr}\big(\mathbf{\Lambda}_{d,0} \mathbf{1}_N \mathbf{1}_n^\top \mathbf{Z} \mathbf{\Upsilon}\big) + \frac{F_0^2}{d} \mathrm{tr}\big([\mathbf{\Lambda}_{d,0} \mathbf{1}_N \mathbf{1}_N^\top \mathbf{\Lambda}_{d,0}]_{\mathbf{Z}} \mathbf{1}_n \mathbf{1}_n^\top\big)}_{I_1} \\
&\quad \underbrace{- \frac{2F_{1,d}^2}{d} \mathrm{tr}\left(\mathbf{\Delta}' \frac{\mathbf{\Theta} \mathbf{X}^\top}{d} \mathbf{Z} \mathbf{\Upsilon}\right)}_{I_2} + \underbrace{\frac{F_0^2}{d} \mathrm{tr}\left(\left[\mathbf{M}_1 \frac{\mathbf{\Theta}\mathbf{\Theta}^\top}{d} \mathbf{M}_1 + \mathbf{M}_* \mathbf{M}_*\right]_{\mathbf{Z}} \mathbf{1}_n \mathbf{1}_n^\top\right)}_{I_3} \\
&\quad + \underbrace{\frac{F_{1,d}^2}{d} \mathrm{tr}\left([\mathbf{\Lambda}_{d,0} \mathbf{1}_N \mathbf{1}_N^\top \mathbf{\Lambda}_{d,0}]_{\mathbf{Z}} \frac{\mathbf{X}\mathbf{X}^\top}{d}\right)}_{I_4} + \underbrace{\frac{\tau^2}{d} \mathrm{tr}\big([\mathbf{\Lambda}_{d,0} \mathbf{1}_N \mathbf{1}_N^\top \mathbf{\Lambda}_{d,0}]_{\mathbf{Z}}\big)}_{I_5} \\
&\quad + \underbrace{\frac{F_{1,d}^2}{d} \mathrm{tr}\left([\mathbf{\Delta}]_{\mathbf{Z}} \frac{\mathbf{X}\mathbf{X}^\top}{d}\right)}_{I_6} + \underbrace{\frac{F_{1,d}^2}{d} \mathrm{tr}[\mathbf{\Delta}]_{\mathbf{Z}}}_{I_7} + \underbrace{\frac{F_0^2}{d} \mathrm{tr}\big([\mathbf{\Delta}]_{\mathbf{Z}} \mathbf{1}_n \mathbf{1}_n^\top\big)}_{I_8}.
\end{aligned}$$

We now show that all the terms $I_1, \ldots, I_8$ on the right hand side above are negligible terms. We note that by definition, $\|\mathbf{Z}\mathbf{\Upsilon}\|_{\mathrm{op}} = \|\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z} + \lambda \mathbf{I})^{-1}\|_{\mathrm{op}} \leq 1/(2\sqrt{\lambda})$ is deterministically bounded. Therefore we have

$$\mathbb{E}|I_2| \leq 2F_{1,d}^2 \cdot \mathbb{E}\left\|\left(\mathbf{\Delta}' \frac{\mathbf{\Theta}\mathbf{X}^\top}{d} \mathbf{Z}\mathbf{\Upsilon}\right)\right\|_{\mathrm{op}} \leq O_d\Big(\frac{1}{2\sqrt{\lambda}}\Big) \cdot \big(\mathbb{E}\|\mathbf{\Delta}'\|_{\mathrm{op}}^2\big)^{\frac{1}{2}} \cdot \left(\mathbb{E}\left\|\frac{\mathbf{\Theta}\mathbf{X}^\top}{d}\right\|_{\mathrm{op}}^2\right)^{\frac{1}{2}} = o_d(1),$$

where the last equality follows by $\mathbb{E}\|\mathbf{\Delta}'\|_{\mathrm{op}}^2 = o_d(1)$ in Lemma I.2. Moreover, by definition, we have

$$\|[\mathbf{\Delta}]_{\mathbf{Z}}\|_{\mathrm{op}} = \|\mathbf{Z}\mathbf{\Upsilon}\mathbf{\Delta}(\mathbf{Z}\mathbf{\Upsilon})^\top\|_{\mathrm{op}} \leq \frac{1}{4\lambda}\|\mathbf{\Delta}\|_{\mathrm{op}}.$$

Therefore, by Lemma I.2 that $\mathbb{E}\|\mathbf{\Delta}\|_{\mathrm{op}}^2 = o_d(1)$, we have

$$\begin{aligned}
&\mathbb{E}|I_6| \leq F_{1,d}^2 \cdot \mathbb{E}\left\|[\mathbf{\Delta}]_{\mathbf{Z}} \frac{\mathbf{X}\mathbf{X}^\top}{d}\right\|_{\mathrm{op}} \leq F_{1,d}^2 \cdot \mathbb{E}\left[\|[\mathbf{\Delta}]_{\mathbf{Z}}\|_{\mathrm{op}} \cdot \left\|\frac{\mathbf{X}\mathbf{X}^\top}{d}\right\|_{\mathrm{op}}\right] = O_d(\mathbb{E}\|\mathbf{\Delta}\|_{\mathrm{op}}^2) = o_d(1), \\
&\mathbb{E}|I_7| \leq F_{1,d}^2 \cdot \mathbb{E}\|[\mathbf{\Delta}]_{\mathbf{Z}}\|_{\mathrm{op}} \leq \frac{F_{1,d}^2}{4\lambda} \cdot \mathbb{E}\|\mathbf{\Delta}\|_{\mathrm{op}} = o_d(1).
\end{aligned}$$

For the remaining terms, we discuss them according to the value of F_0. When $F_0 = 0$, it is clear that $I_1 = I_3 = 0$. Note that under this situation, the condition $\sum_j \mu_{j,0}^2 > 0$ in Lemma I.3 may not hold. If $\sum_j \mu_{j,0}^2 = 0$, from Lemma I.3, it still holds that

$$\mathbb{E}|I_4| = o_d(1), \qquad \mathbb{E}|I_5| = o_d(1).$$

Therefore, when $F_0 = 0$, Proposition A.2 holds.

When $F_0 \neq 0$, $\sum_j \mu_{j,0}^2 > 0$ holds from Assumption 3.4, the result for $\mathcal{C}$ in Lemma I.3 gives the bound for I_1, the result for $\mathcal{B}$ in Lemma I.3 gives the bounds on I_3 and I_8, and the result for $\mathcal{D}$ in

Lemma I.3 gives the bounds on I_4 and I_5. Therefore we have

$$\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}\left|\mathbb{E}_{\widetilde{\boldsymbol{\beta}}_d,\boldsymbol{\varepsilon}} R_d(\mathbf{X},\mathbf{\Theta},\lambda,\widetilde{\boldsymbol{\beta}}_d,\varepsilon) - \overline{R}_d(\mathbf{X},\mathbf{\Theta},\lambda,F_{1,d},\tau)\right| = o_d(1),$$

which proves Proposition A.2. ■

I.1 Proof of Lemma I.2

The proof of Lemma I.2 is mainly based on the decomposition of the nonlinear activation function. We first present several classical lemmas about Gegenbauer polynomials and their relation to Hermite polynomials. The following lemma can be found in Mei and Montanari (2022) (see Lemma 9.4 and its proof in the reference).

Lemma I.5. *Let $Q_k^{(d)}(\cdot)$, $k \in \mathbb{N}$ be the Gegenbauer polynomials. The following properties hold:*

1. *For $\mathbf{v}_1$,$\mathbf{v}_2 \in \sqrt{d}\cdot\mathbb{S}^{d-1}$, suppose $\mathbf{x} \sim \mathrm{Unif}(\sqrt{d}\cdot\mathbb{S}^{d-1})$, then for $k,l \in \mathbb{N}$,*

$$\mathbb{E}_{\mathbf{x}}\left[Q_k^{(d)}(\mathbf{v}_1^\top\mathbf{x})Q_l^{(d)}(\mathbf{x}^\top\mathbf{v}_2)\right] = \frac{\delta_{kl}}{B(d,k)}\cdot Q_k^{(d)}(\mathbf{v}_1^\top\mathbf{v}_2),$$

   where $\delta_{kl} = 1$ if $k = l$ and $\delta_{kl} = 0$ if $k \neq l$.

2. *For $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ defined in Section 2, $Q_k^{(d)}(\cdot)$ the point wise function on matrices, the following equality holds:*

$$\mathbb{E}\Big[\sup_{k\geq 2}\big\|Q_k^{(d)}(\mathbf{\Theta}_j\mathbf{\Theta}_j^\top) - \mathbf{I}_{N_j}\big\|_{\mathrm{op}}^2\Big] = o_d(1), \quad j = 1,2,$$

$$\mathbb{E}\Big[\sup_{k\geq 2}\big\|Q_k^{(d)}(\mathbf{\Theta}_1\mathbf{\Theta}_2^\top)\big\|_{\mathrm{op}}^2\Big] = o_d(1).$$

The next lemma gives the connection between the coefficients in Hermite polynomials H_k and the coefficients in Gegenbauer polynomials $Q_k^{(d)}$.

Lemma I.6. *Let $Q_k^{(d)}(\cdot)$, $H_k(\cdot)$, $k \in \mathbb{N}$ be the Gegenbauer and Hermite polynomials respectively. For $j = 1,2$, suppose that $\sigma_j(x)$ has Gegenbauer decomposition*

$$\sigma_j(x) = \sum_{k=0}^{+\infty}\lambda_{d,k}(\sigma_j)B(d,k)\cdot Q_k^{(d)}(\sqrt{d}\cdot x)$$

and Hermite polynomial decomposition

$$\sigma_j(x) = \sum_{k=0}^{+\infty}\alpha_k(\sigma_j)/k!\cdot H_k(x).$$

Then for each $k \in \mathbb{N}$, $\lambda_{d,k}^2(\sigma_j)B(d,k)k! \to \alpha_k^2(\sigma_j)$ as $d \to +\infty$.

The proof of Lemma I.6 can be found in Appendix A.3 in Mei and Montanari (2022). Note that the orthogonality of the standard Hermite polynomials ($H_1(x) = x$) implies that for $G \sim N(0,1)$,

$$\mathbb{E}[H_k(G)H_l(G)] = \delta_{kl}\cdot k!.$$

Based on this property, let $\alpha_k(\sigma_j)$ be defined in Lemma I.6. Then for $j = 1, 2$, we have

$$\alpha_k(\sigma_j) = \mu_{j,k},\ k = 0, 1, \qquad \mu_{j,*}^2 = \sum_{k\geq 2} \frac{\alpha_k^2(\sigma_j)}{k!},$$

where the constants $\mu_{j,k}$ and $\mu_{j,*}$ are defined in Definition 3.1. Therefore, by Lemma I.6, we further have

$$\sum_{k\geq 2} \lambda_{d,k}^2(\sigma_j) B(d,k) \to \mu_{j,*}^2. \tag{I.6}$$

Recall that $\boldsymbol{\sigma}(\mathbf{x}) = \big(\sigma_1(\mathbf{x}^\top \boldsymbol{\Theta}_1^\top/\sqrt{d}), \sigma_2(\mathbf{x}^\top \boldsymbol{\Theta}_2^\top/\sqrt{d})\big)^\top$. Moreover, note that the zeroth order Gegenbauer polynomial $Q_0^d(x) = 1$. Therefore by Lemma I.5 and the Gegenbauer decomposition of $\boldsymbol{\sigma}_j$ in Lemma I.6, we have

$$\mathbf{V}_0(F_0) = F_0 \mathbb{E}_{\mathbf{x}}[\boldsymbol{\sigma}(\mathbf{x}) \cdot Q_0^d(\mathbf{x}^\top \mathbf{1}_d)] = \frac{F_0}{B(d,0)} \cdot \boldsymbol{\Lambda}_{d,0} \cdot Q_0^d(\boldsymbol{\Theta}\mathbf{1}_d) \cdot B(d,0) = F_0 \boldsymbol{\Lambda}_{d,0} \mathbf{1}_N.$$

Here, the equality holds from the fact that $Q_0^d(\mathbf{x}^\top \mathbf{1}_d) = 1$ and $Q_0^d(\boldsymbol{\Theta}\mathbf{1}_d) = \mathbf{1}_N$. Similarly, $Q_1^d(x) = x/d$ holds. Again from Lemma I.5 and Lemma I.6, we have

$$\begin{aligned}\mathbf{V}(\boldsymbol{\beta}_{1,d}) &= \mathbb{E}_{\mathbf{x}} \boldsymbol{\sigma}(\mathbf{x}) \mathbf{x}^\top \boldsymbol{\beta}_{1,d} = d \cdot \mathbb{E}_{\mathbf{x}} \boldsymbol{\sigma}(\mathbf{x}) Q_1^d(\mathbf{x}^\top \boldsymbol{\beta}_{1,d}) = \frac{d}{B(d,1)} \cdot \boldsymbol{\Lambda}_{d,1} \cdot Q_1^d(\boldsymbol{\Theta}\boldsymbol{\beta}_{1,d}) \cdot B(d,1) \\ &= \boldsymbol{\Lambda}_{d,1} \boldsymbol{\Theta} \boldsymbol{\beta}_{1,d} = \left(\frac{\mathbf{M}_1 + \boldsymbol{\Delta}'}{\sqrt{d}}\right) \boldsymbol{\Theta} \boldsymbol{\beta}_{1,d}.\end{aligned}$$

Here, $\boldsymbol{\Delta}' = \sqrt{d} \cdot \boldsymbol{\Lambda}_{d,1} - \mathbf{M}_1$. From Lemma I.6, set $k = 1$ and we have $\sqrt{d}\lambda_{d,1}(\sigma_j) \to \mu_{j,1}$. Thus Δ' satisfies $\mathbb{E}\|\boldsymbol{\Delta}'\|_{\text{op}}^2 = o_d(1)$. As for $\mathbf{U} = \mathbb{E}_{\mathbf{x}}\big[\boldsymbol{\sigma}(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})^\top\big]$, $\mathbf{U}$ could be divided into the following block matrix:

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_{1,1} & \mathbf{U}_{1,2} \\ \mathbf{U}_{2,1} & \mathbf{U}_{2,2} \end{bmatrix},$$

where

$$\mathbf{U}_{i,j} = \mathbb{E}_{\mathbf{x}}[\sigma_i(\boldsymbol{\Theta}_i \mathbf{x}/\sqrt{d}) \sigma_j(\mathbf{x}^\top \boldsymbol{\Theta}_j^\top/\sqrt{d})], \quad i, j = 1, 2.$$

Now by Lemma I.5, we have

$$\mathbf{U}_{i,j} = \sum_{k=0}^{+\infty} \lambda_{d,k}(\sigma_i)\lambda_{d,k}(\sigma_j) B(d,k) Q_k^{(d)}(\boldsymbol{\Theta}_i \boldsymbol{\Theta}_j^\top), \quad i, j = 1, 2. \tag{I.7}$$

Note that $Q_0^{(d)}(x) = 1$, $Q_1^{(d)}(x) = x/d$, so that the first two terms in the decomposition (I.7) have a simple form. For $k \geq 2$, we approximate the terms using the approximation given in the second item of Lemma I.5. Consider first $\mathbf{U}_{1,1}$. We have

$$\mathbf{U}_{1,1} = \lambda_{d,0}^2(\sigma_1) \cdot \mathbf{1}_{N_1} \mathbf{1}_{N_1}^\top + \lambda_{d,1}^2(\sigma_1) \cdot B(d,1) \cdot \frac{\boldsymbol{\Theta}_1 \boldsymbol{\Theta}_1^\top}{d} + \sum_{k=2}^{+\infty} \lambda_{d,k}^2(\sigma_1) \cdot B(d,k) \cdot Q_k^{(d)}(\boldsymbol{\Theta}_1 \boldsymbol{\Theta}_1^\top)$$

$$= \lambda_{d,0}^2(\sigma_1) \cdot \mathbf{1}_{N_1}\mathbf{1}_{N_1}^\top + \lambda_{d,1}^2(\sigma_1) \cdot B(d,1) \cdot \frac{\mathbf{\Theta}_1\mathbf{\Theta}_1^\top}{d} + \sum_{k=2}^{+\infty} \lambda_{d,k}^2(\sigma_1) \cdot B(d,k) \cdot \mathbf{I}_{N_1}$$
$$+ \sum_{k=2}^{+\infty} \lambda_{d,k}^2(\sigma_1) \cdot B(d,k) \cdot \big[Q_k^{(d)}(\mathbf{\Theta}_1\mathbf{\Theta}_1^\top) - \mathbf{I}_{N_1}\big], \tag{I.8}$$

where we have used the fact that $\sum_{k=2}^{+\infty} \lambda_{d,k}^2(\sigma_1) B(d,k) < +\infty$ for sufficiently large d, which is implied by (I.6). Moreover, by Lemma I.5, the convergence of this series also implies that

$$\mathbb{E}\left\| \sum_{k=2}^{+\infty} \lambda_{d,k}^2(\sigma_1) \cdot B(d,k) \cdot \big[Q_k^{(d)}(\mathbf{\Theta}_1\mathbf{\Theta}_1^\top) - \mathbf{I}_{N_1}\big] \right\|_{\text{op}}^2 = o_d(1). \tag{I.9}$$

Therefore by (I.8) and (I.9), we have

$$\mathbb{E}\left\| \mathbf{U}_{1,1} - \lambda_{d,0}^2(\sigma_1) \cdot \mathbf{1}_{N_1}\mathbf{1}_{N_1}^\top - \lambda_{d,1}^2(\sigma_1) \cdot B(d,1) \cdot \frac{\mathbf{\Theta}_1\mathbf{\Theta}_1^\top}{d} - \sum_{k=2}^{+\infty} \lambda_{d,k}^2(\sigma_1) \cdot B(d,k) \cdot \mathbf{I}_{N_1} \right\|_{\text{op}}^2 = o_d(1). \tag{I.10}$$

Now by Lemma I.6 and equations (I.6), (I.10), we have

$$\mathbb{E}\left\| \mathbf{U}_{1,1} - \lambda_{d,0}^2(\sigma_1) \cdot \mathbf{1}_{N_1}\mathbf{1}_{N_1}^\top - \mu_{1,1}^2 \cdot \frac{\mathbf{\Theta}_1\mathbf{\Theta}_1^\top}{d} - \mu_{1,*}^2 \cdot \mathbf{I}_{N_1} \right\|_{\text{op}}^2 = o_d(1).$$

This establishes the approximation for $\mathbf{U}_{1,1}$.

For the other sub-matrices $\mathbf{U}_{1,2}$, $\mathbf{U}_{2,1}$ and $\mathbf{U}_{2,2}$, the derivations are exactly the same, and we obtain the following results:

$$\mathbb{E}\left\| \mathbf{U}_{1,2} - \lambda_{d,0}(\sigma_1)\lambda_{d,0}(\sigma_2)\mathbf{1}_{N_1}\mathbf{1}_{N_2}^\top - \mu_{1,1}\mu_{2,1} \cdot \frac{\mathbf{\Theta}_1\mathbf{\Theta}_2^\top}{d} \right\|_{\text{op}}^2 = o_d(1),$$
$$\mathbb{E}\left\| \mathbf{U}_{2,1} - \lambda_{d,0}(\sigma_1)\lambda_{d,0}(\sigma_2)\mathbf{1}_{N_1}\mathbf{1}_{N_2}^\top - \mu_{1,1}\mu_{2,1} \cdot \frac{\mathbf{\Theta}_2\mathbf{\Theta}_1^\top}{d} \right\|_{\text{op}}^2 = o_d(1),$$
$$\mathbb{E}\left\| \mathbf{U}_{2,2} - \lambda_{d,0}^2(\sigma_2)\mathbf{1}_{N_2}\mathbf{1}_{N_2}^\top - \mu_{2,1}^2 \frac{\mathbf{\Theta}_2\mathbf{\Theta}_2^\top}{d} - \mu_{2,*}^2 \cdot \mathbf{I}_{N_2} \right\|_{\text{op}}^2 = o_d(1).$$

Note that the collection of the approximations for the four blocks $\mathbf{U}_{1,1}, \mathbf{U}_{1,2}, \mathbf{U}_{2,1}$ and $\mathbf{U}_{1,1}$ gives the matrix

$$\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0} + \mathbf{M}_1\frac{\mathbf{\Theta}\mathbf{\Theta}^\top}{d}\mathbf{M}_1 + \mathbf{M}_*\mathbf{M}_*,$$

so finally we have

$$\mathbb{E}\left\| \mathbf{U} - \mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0} - \mathbf{M}_1\frac{\mathbf{\Theta}\mathbf{\Theta}^\top}{d}\mathbf{M}_1 - \mathbf{M}_*\mathbf{M}_* \right\|_{\text{op}}^2 = o_d(1).$$

The proof of Lemma I.2 is complete.

I.2 Proof of Lemma I.3

We first prove that $\left(\mathbb{E}|\mathcal{D}|^k\right)^{1/k} = o_d(1)$ if $\sum_j \mu_{j,0}^2 = 0$. Note that the rank-1 matrix $\mathbf{A}$ satisfies $|\mathrm{tr}(\mathbf{A})| = \|\mathbf{A}\|_{\mathrm{op}}$. Moreover, $\sum_j \mu_{j,0}^2 = 0$ implies $\|\Lambda_{d,0}\|_{\mathrm{op}} = o_d(1)$. We have $\left(\mathbb{E}|\mathcal{D}|^k\right)^{1/k} = O_d(\|\mathbf{\Lambda}_{d,0}\frac{\mathbf{1}_N\mathbf{1}_N^\top}{d}\mathbf{\Lambda}_{d,0}\|_{\mathrm{op}}) \cdot (\mathbb{E}\|\mathbf{\Gamma}_2\|_{\mathrm{op}}^k)^{1/k} = o_d(1)\cdot O_d(1) = o_d(1)$.

In the following proof of Lemma I.3, we have the condition $\sum_j \mu_{j,0}^2 > 0$. We separate the proof into two parts, estimating $\mathcal{B}$ and $\mathcal{C}$, and $\mathcal{D}$, respectively.

I.2.1 Estimation for $\mathcal{B}$ and $\mathcal{C}$

Let

$$L_1 = \frac{1}{\sqrt{d}}\mathrm{tr}(\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_n^\top\mathbf{Z}\mathbf{\Upsilon}), \qquad L_2(\mathbf{\Gamma}) = \frac{1}{d}\mathrm{tr}(\left[\mathbf{\Gamma}\right]_{\mathbf{Z}}\mathbf{1}_n\mathbf{1}_n^\top) = \frac{1}{d}\mathrm{tr}(\mathbf{Z}\mathbf{\Upsilon}\mathbf{\Gamma}\mathbf{\Upsilon}\mathbf{Z}^\top\mathbf{1}_n\mathbf{1}_n^\top),$$

where $\mathbf{\Gamma}\in\mathbb{R}^{N\times N}$ is a symmetric matrix. Then we have

$$\mathcal{B} = L_2(\mathbf{\Gamma}), \quad \mathcal{C} = 1 - 2L_1 + L_2(\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}).$$

Define further the following terms:

$$\begin{array}{lll} K_{11} = \mathbf{T}_1^\top\mathbf{E}_0^{-1}\mathbf{T}_1, & K_{12} = \mathbf{T}_1^\top\mathbf{E}_0^{-1}\mathbf{T}_2, & K_{22} = \mathbf{T}_2^\top\mathbf{E}_0^{-1}\mathbf{T}_2, \\ G_{11} = \mathbf{T}_1^\top\mathbf{E}_0^{-1}\mathbf{\Gamma}\mathbf{E}_0^{-1}\mathbf{T}_1, & G_{12} = \mathbf{T}_1^\top\mathbf{E}_0^{-1}\mathbf{\Gamma}\mathbf{E}_0^{-1}\mathbf{T}_2, & G_{22} = \mathbf{T}_2^\top\mathbf{E}_0^{-1}\mathbf{\Gamma}\mathbf{E}_0^{-1}\mathbf{T}_2, \end{array}$$

where

$$\begin{array}{ll} \mathbf{J} = \mathbf{Z} - \mathbf{1}_n\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}/\sqrt{d}, & \mathbf{E}_0 = \mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N, \\ \mathbf{T}_1 = \psi_3^{1/2}\mathbf{\Lambda}_{d,0}\mathbf{1}_N, & \mathbf{T}_2 = \frac{1}{\sqrt{n}}\mathbf{J}^\top\mathbf{1}_n. \end{array}$$

We denote $\psi_3 = n/d$ for notation simplification. The proof is organized in two steps:

1. Express $\mathcal{B}$ and $\mathcal{C}$ in function of K_{ij} and G_{ij}, $i,j\in\{1,2\}$.
2. Estimate the order of K_{ij} and G_{ij}, and show that $\mathbb{E}|\mathcal{B}|$ and $\mathbb{E}|\mathcal{C}|$ are both $o_d(1)$.

Denote $\mathbf{F}_1 = [\mathbf{T}_1, \mathbf{T}_1, \mathbf{T}_2]\in\mathbb{R}^{N\times 3}$, $\mathbf{F}_2 = [\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_1]\in\mathbb{R}^{N\times 3}$, it is easy to see

$$\begin{aligned} \mathbf{\Upsilon} =& \left(\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{I}_N\right)^{-1} = \left((\mathbf{J} + \mathbf{1}_n\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0})^\top(\mathbf{J} + \mathbf{1}_n\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}) + \lambda\mathbf{I}_N\right)^{-1} \\ =& \left(\psi_3\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top + \psi_3^{1/2}\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{T}_2^\top + \psi_3^{1/2}\mathbf{T}_2\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0} + \mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N\right)^{-1} \\ =& \left(\mathbf{E}_0 + \mathbf{F}_1\mathbf{F}_2^\top\right)^{-1}. \end{aligned}$$

For L_1, replacing $\mathbf{Z}$ by $\mathbf{J} + \mathbf{1}_n\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}/\sqrt{d}$, we have

$$\begin{aligned} L_1 =& \mathrm{tr}\left[(\psi_3\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0} + \psi_3^{1/2}\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{T}_2^\top)\cdot\left(\mathbf{E}_0 + \mathbf{F}_1\mathbf{F}_2^\top\right)^{-1}\right] \\ =& \mathrm{tr}\left[(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{T}_1\mathbf{T}_2^\top)\cdot\left(\mathbf{E}_0 + \mathbf{F}_1\mathbf{F}_2^\top\right)^{-1}\right]. \end{aligned} \tag{I.11}$$

By the Sherman-Morrison-Woodbury formula,

$$\mathbf{\Upsilon} = \left(\mathbf{E}_0 + \mathbf{F}_1\mathbf{F}_2^\top\right)^{-1} = \mathbf{E}_0^{-1} - \mathbf{E}_0^{-1}\mathbf{F}_1(\mathbf{I}_3 + \mathbf{F}_2^\top\mathbf{E}_0^{-1}\mathbf{F}_1)^{-1}\mathbf{F}_2^T\mathbf{E}_0^{-1}. \tag{I.12}$$

Plugging (I.12) into (I.11), we have

$$\begin{aligned}
L_1 = \; & (\mathbf{T}_1^\top \mathbf{E}_0^{-1}\mathbf{T}_1 - \mathbf{T}_1^\top \mathbf{E}_0^{-1}\mathbf{F}_1(\mathbf{I}_3 + \mathbf{F}_2^\top \mathbf{E}_0^{-1}\mathbf{F}_1)^{-1}\mathbf{F}_2^\top \mathbf{E}_0^{-1}\mathbf{T}_1) \\
& + (\mathbf{T}_2^\top \mathbf{E}_0^{-1}\mathbf{T}_1 - \mathbf{T}_2^\top \mathbf{E}_0^{-1}\mathbf{F}_1(\mathbf{I}_3 + \mathbf{F}_2^\top \mathbf{E}_0^{-1}\mathbf{F}_1)^{-1}\mathbf{F}_2^\top \mathbf{E}_0^{-1}\mathbf{T}_1) \\
= \; & (K_{11} - [K_{11}, K_{11}, K_{12}](\mathbf{I}_3 + \boldsymbol{K})^{-1}[K_{11}, K_{12}, K_{11}]^\top) \\
& + (K_{12} - [K_{12}, K_{12}, K_{22}](\mathbf{I}_3 + \boldsymbol{K})^{-1}[K_{11}, K_{12}, K_{11}]^\top) \\
= \; & [K_{11}, K_{11}, K_{12}](\mathbf{I}_3 + \boldsymbol{K})^{-1}[1,0,0]^\top + [K_{12}, K_{12}, K_{22}](\mathbf{I}_3 + \boldsymbol{K})^{-1}[1,0,0]^\top,
\end{aligned}$$

where

$$\boldsymbol{K} = \mathbf{F}_2^\top \mathbf{E}_0^{-1}\mathbf{F}_1 = \begin{bmatrix} K_{11} & K_{11} & K_{12} \\ K_{12} & K_{12} & K_{22} \\ K_{11} & K_{11} & K_{12} \end{bmatrix}.$$

Thus by simple calculation,

$$L_1 = 1 - \frac{K_{12}+1}{K_{11}(1-K_{22}) + (K_{12}+1)^2}. \tag{I.13}$$

As for $L_2(\boldsymbol{\Gamma})$, we have

$$\begin{aligned}
\mathbf{Z}^\top \mathbf{1}_n \mathbf{1}_n^\top \mathbf{Z}/d =& (\mathbf{J} + \mathbf{1}_n \mathbf{1}_N^\top \boldsymbol{\Lambda}_{d,0}/\sqrt{d})^\top \mathbf{1}_n \mathbf{1}_n^\top (\mathbf{J} + \mathbf{1}_n \mathbf{1}_N^\top \boldsymbol{\Lambda}_{d,0}/\sqrt{d})/d \\
=& \psi_3(\psi_3^{1/2}\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N + \frac{1}{\sqrt{n}}\mathbf{J}^\top \mathbf{1}_n)(\psi_3^{1/2}\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N + \frac{1}{\sqrt{n}}\mathbf{J}^\top \mathbf{1}_n)^\top \\
=& \psi_3(\mathbf{T}_1 + \mathbf{T}_2)(\mathbf{T}_1 + \mathbf{T}_2)^\top.
\end{aligned}$$

Then after similar calculation by (I.12).

$$\begin{aligned}
\mathcal{B} = L_2(\boldsymbol{\Gamma}) =& \frac{1}{d}\mathrm{tr}\big(\mathbf{Z}^\top \mathbf{1}_n \mathbf{1}_n^\top \mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Gamma}\boldsymbol{\Upsilon}\big) = \mathrm{tr}\big(\psi_3(\mathbf{T}_1 + \mathbf{T}_2)(\mathbf{T}_1 + \mathbf{T}_2)^\top \boldsymbol{\Upsilon}\boldsymbol{\Gamma}\boldsymbol{\Upsilon}\big) \\
=& \psi_3(\mathbf{T}_1 + \mathbf{T}_2)^\top \big(\mathbf{E}_0 + \mathbf{F}_1\mathbf{F}_2^\top\big)^{-1}\boldsymbol{\Gamma}\big(\mathbf{E}_0 + \mathbf{F}_1\mathbf{F}_2^\top\big)^{-1}(\mathbf{T}_1 + \mathbf{T}_2) \\
=& \psi_3(\mathbf{T}_1 + \mathbf{T}_2)^\top \big(\mathbf{E}_0^{-1} - \mathbf{E}_0^{-1}\mathbf{F}_1(\mathbf{I}_3 + \mathbf{F}_2^\top \mathbf{E}_0^{-1}\mathbf{F}_1)^{-1}\mathbf{F}_2^T \mathbf{E}_0^{-1}\big) \\
& \cdot \boldsymbol{\Gamma}\big(\mathbf{E}_0^{-1} - \mathbf{E}_0^{-1}\mathbf{F}_1(\mathbf{I}_3 + \mathbf{F}_2^\top \mathbf{E}_0^{-1}\mathbf{F}_1)^{-1}\mathbf{F}_2^T \mathbf{E}_0^{-1}\big)(\mathbf{T}_1 + \mathbf{T}_2) \\
=& \psi_3 \frac{G_{11}(1-K_{22})^2 + G_{22}(K_{12}+1)^2 + 2G_{12}(K_{12}+1)(1-K_{22})}{(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2}.
\end{aligned} \tag{I.14}$$

When $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}$, the G_{11}, G_{12} and G_{22} above can be given as

$$G_{11} = K_{11}^2/\psi_3, \qquad G_{12} = K_{11}K_{12}/\psi_3, \qquad G_{22} = K_{12}^2/\psi_3.$$

Then by (I.13), (I.14), we have

$$\mathcal{C} = 1 - 2L_1 + L_2(\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}) = \frac{(K_{12}+1)^2}{(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2}. \tag{I.15}$$

We next estimate the order for K_{11}, K_{12}, K_{22}, G_{11}, G_{12} and G_{22} respectively. By the inequality $\left\|\begin{bmatrix}\mathbf{A} & \mathbf{B}\end{bmatrix}\right\|_{\mathrm{op}} \le \|\mathbf{A}\|_{\mathrm{op}} + \|\mathbf{B}\|_{\mathrm{op}}$ for any matrices $\mathbf{A}$ and $\mathbf{B}$, we have

$$\begin{aligned}\|\mathbf{J}\|_{\mathrm{op}} &\le \|\mathbf{Z}_1 - \lambda_{d,0}(\sigma_1)\mathbf{1}_n\mathbf{1}_{N_1}^\top/\sqrt{d}\|_{\mathrm{op}} + \|\mathbf{Z}_2 - \lambda_{d,0}(\sigma_2)\mathbf{1}_n\mathbf{1}_{N_2}^\top/\sqrt{d}\|_{\mathrm{op}} \\ &= O_{\mathbb{P}}(\exp(C\sqrt{\log d})),\end{aligned} \tag{I.16}$$

where the last equality in (I.16) follows by Lemma C.5 in (Mei and Montanari, 2022). Moreover, for any fixed $\lambda > 0$, it also determinstically holds that

$$\|(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\mathbf{J}^\top\|_{\mathrm{op}} \le 2/\sqrt{\lambda}, \quad \|(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\|_{\mathrm{op}} \le 1/\lambda.$$

Now recall that

$$\begin{aligned}K_{11} =& \psi_3\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\mathbf{\Lambda}_{d,0}\mathbf{1}_N, \\ K_{12} =& \mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\mathbf{J}^\top\mathbf{1}_n/\sqrt{d}, \\ K_{22} =& \mathbf{1}_n^\top\mathbf{J}(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\mathbf{J}^\top\mathbf{1}_n/n, \\ G_{11} =& \psi_3\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\mathbf{\Gamma}(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\mathbf{\Lambda}_{d,0}\mathbf{1}_N, \\ G_{12} =& \mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\mathbf{\Gamma}(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\mathbf{J}^\top\mathbf{1}_n/\sqrt{d}, \\ G_{22} =& \mathbf{1}_n^\top\mathbf{J}(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\mathbf{\Gamma}(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\mathbf{J}^\top\mathbf{1}_n/n.\end{aligned}$$

Therefore we deterministically have

$$|K_{12}| \le \|(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\mathbf{J}^\top\|_{\mathrm{op}}\|\mathbf{1}_n\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}/\sqrt{d}\|_{\mathrm{op}} = O_d(\sqrt{d/\lambda}). \tag{I.17}$$

For K_{22}, by its definition, it is clear that $K_{22} > 0$. Moreover, we have

$$\begin{aligned}K_{22} &\le \lambda_{\max}(\mathbf{J}(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}\mathbf{J})\mathrm{tr}(\mathbf{1}_n\mathbf{1}_n^\top/n) \\ &= \lambda_{\max}(\mathbf{I}_N - \lambda(\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}) = 1 - \frac{\lambda}{\|\mathbf{J}^\top\mathbf{J}\|_{\mathrm{op}} + \lambda}.\end{aligned}$$

Therefore we have

$$0 < K_{22} \le 1 - \frac{\lambda}{\|\mathbf{J}^\top\mathbf{J}\|_{\mathrm{op}} + \lambda}. \tag{I.18}$$

For K_{11}, the condition $\mu_{1,0}^2 + \mu_{2,0}^2 > 0$ ensures that there exists $j \in \{1, 2\}$ such that $\mu_{j,0}^2 > 0$. By Lemma I.6 (note that $B(d, 0) = 1$), we have $\lambda_{d,0}^2(\sigma_j) \to \mu_{j,0}^2$ as $d \to +\infty$. Therefore for large enough d, we have $\lambda_{d,0}(\sigma_j) > \mu_{j,0}/2 > 0$, and

$$\begin{aligned}K_{11} &\ge \psi_3\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}^2\mathbf{1}_N\lambda_{\min}((\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}) \\ &\ge \psi_3 \cdot (\mu_{j,0}^2/4) \cdot N_j \cdot \lambda_{\min}((\mathbf{J}^\top\mathbf{J} + \lambda\mathbf{I}_N)^{-1}) \\ &= \frac{\Omega_d(d)}{\|\mathbf{J}^\top\mathbf{J}\|_{\mathrm{op}} + \lambda}.\end{aligned} \tag{I.19}$$

Plugging (I.17), (I.18), (I.19) into (I.14) then gives

$$|\mathcal{B}| = \frac{\left|G_{22}(1 + K_{12})^2 + G_{11}(1 - K_{22})^2 + 2G_{12}(1 + K_{12})(1 - K_{22})\right|}{\left[(1 + K_{12})^2 + K_{11} \cdot (1 - K_{22})\right]^2}$$

$$\leq \frac{\left|G_{22}(1+K_{12})^2+G_{11}(1-K_{22})^2+2G_{12}(1+K_{12})(1-K_{22})\right|}{\left[K_{11}\cdot(1-K_{22})\right]^2}$$
$$\leq O_d(1)\cdot\frac{|G_{22}|\cdot d+|G_{12}|\cdot\sqrt{d}+|G_{11}|}{d^2/(\lambda+\|\mathbf{J}\mathbf{J}^\top\|_{\rm op})^4},$$

where we utilize the upper and lower bounds in (I.17), (I.18), (I.19) to obtain the last inequality. For G_{11}, G_{12} and G_{22}, we have

$$\mathbb{E}\big[|G_{11}|^k\big]^{1/k}\leq \psi_3\big\|(\mathbf{J}^\top\mathbf{J}+\lambda\mathbf{I}_N)^{-1}\big\|_{\rm op}\big[\mathbb{E}\|\mathbf{\Gamma}\|_{\rm op}^k\big]^{1/k}\big\|(\mathbf{J}^\top\mathbf{J}+\lambda\mathbf{I}_N)^{-1}\big\|_{\rm op}\big\|\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\big\|_{\rm op}=O_d(d),$$
$$\mathbb{E}\big[|G_{12}|^k\big]^{1/k}\leq \big\|(\mathbf{J}^\top\mathbf{J}+\lambda\mathbf{I}_N)^{-1}\big\|_{\rm op}\big[\mathbb{E}\|\mathbf{\Gamma}\|_{\rm op}^k\big]^{1/k}\big\|(\mathbf{J}^\top\mathbf{J}+\lambda\mathbf{I}_N)^{-1}\mathbf{J}^\top\big\|_{\rm op}\big\|\mathbf{1}_n\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}/\sqrt{d}\big\|_{\rm op}=O_d(\sqrt{d}),$$
$$\mathbb{E}\big[|G_{22}|^k\big]^{1/k}\leq \big\|(\mathbf{J}^\top\mathbf{J}+\lambda\mathbf{I}_N)^{-1}\big\|_{\rm op}\big[\mathbb{E}\|\mathbf{\Gamma}\|_{\rm op}^k\big]^{1/k}\big\|(\mathbf{J}^\top\mathbf{J}+\lambda\mathbf{I}_N)^{-1}\mathbf{J}^\top\mathbf{J}\big\|_{\rm op}\mathrm{tr}\big(\mathbf{1}_n\mathbf{1}_n^\top/n\big)=O_d(1).$$

Thus by the bounds above and the triangle inequality of the L_k-norm $\mathbb{E}[|\cdot|^k]^{1/k}$, we have

$$\begin{aligned}\big(\mathbb{E}|\mathcal{B}|^k\big)^{1/k}&\leq O_d(1)\cdot\frac{\mathbb{E}\big[|G_{22}|^k\big]^{1/k}\cdot d+\mathbb{E}\big[|G_{12}|^k\big]^{1/k}\cdot\sqrt{d}+\mathbb{E}\big[|G_{11}|^k\big]^{1/k}}{d^2/(\lambda+\|\mathbf{J}\mathbf{J}^\top\|_{\rm op})^4}\\&=O_d(1)\cdot\frac{d}{d^2/(\lambda+\|\mathbf{J}\mathbf{J}^\top\|_{\rm op})^4}=O_d\bigg(\frac{\big(\lambda+\|\mathbf{J}\mathbf{J}^\top\|_{\rm op}\big)^4}{d}\bigg)=O_d\bigg(\frac{\exp(C\sqrt{\log d})}{d}\bigg)\\&=o_d(1),\end{aligned}$$

and

$$\mathbb{E}|\mathcal{C}|=O_d\bigg(\frac{\big(\lambda+\|\mathbf{J}\mathbf{J}^\top\|_{\rm op}\big)^2}{d}\bigg)=O_d\bigg(\frac{\exp(C\sqrt{\log d})}{d}\bigg)=o_d(1).$$

This completes the proof.

I.2.2 Estimation for $\mathcal{D}$

The proof is similar to the calculations for $\mathcal{B}$ and $\mathcal{C}$ in the previous section. Also we use a set of similar notations as previously which may however have slightly different values. Let

$$\begin{aligned}&\mathbf{J}=\mathbf{Z}-\mathbf{1}_n\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}/\sqrt{d}, && \mathbf{E}_0=\mathbf{J}\mathbf{J}^\top+\lambda\mathbf{I}_n,\\&\mathbf{T}_1=\mathbf{J}\mathbf{\Lambda}_{d,0}\mathbf{1}_N/\sqrt{d}, && \mathbf{T}_2=\mathbf{1}_n,\\&K_{11}=\mathbf{T}_1^\top\mathbf{E}_0^{-1}\mathbf{T}_1, && K_{12}=\mathbf{T}_1^\top\mathbf{E}_0^{-1}\mathbf{T}_2, && K_{22}=\mathbf{T}_2^\top\mathbf{E}_0^{-1}\mathbf{T}_2,\\&G_{11}=\mathbf{T}_1^\top\mathbf{E}_0^{-1}\mathbf{\Gamma}\mathbf{E}_0^{-1}\mathbf{T}_1, && G_{12}=\mathbf{T}_1^\top\mathbf{E}_0^{-1}\mathbf{\Gamma}\mathbf{E}_0^{-1}\mathbf{T}_2, && G_{22}=\mathbf{T}_2^\top\mathbf{E}_0^{-1}\mathbf{\Gamma}\mathbf{E}_0^{-1}\mathbf{T}_2,\end{aligned}$$

where $\mathbf{\Gamma}\in\mathbb{R}^{n\times n}$ is a symmetric matrix. We express $\mathcal{D}$ with the terms defined above. Recall that $\mathbf{\Upsilon}=(\mathbf{Z}^\top\mathbf{Z}+\lambda\mathbf{I}_N)^{-1}$and further define $\mathbf{\Xi}=(\mathbf{Z}\mathbf{Z}^\top+\lambda\mathbf{I}_n)^{-1}$. Clearly, $\mathbf{Z}\mathbf{\Upsilon}=\mathbf{\Xi}\mathbf{Z}$. Therefore we have

$$\mathcal{D}=\frac{1}{d}\mathrm{tr}(\mathbf{Z}\mathbf{\Upsilon}\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\mathbf{\Upsilon}\mathbf{Z}^\top\mathbf{\Gamma})=\frac{1}{d}\mathrm{tr}(\mathbf{\Xi}\mathbf{Z}\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\mathbf{Z}^\top\mathbf{\Xi}\mathbf{\Gamma}).\tag{I.20}$$

We proceed to calculate $\mathbf{\Xi}$ and $\mathbf{Z}\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\mathbf{Z}^\top$, respectively. Define $c=\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}^2\mathbf{1}_N/d=\Theta(1)$, $\mathbf{F}_1=[\mathbf{T}_1,\mathbf{T}_2,\mathbf{T}_2]\in\mathbb{R}^{n\times3}$, $\mathbf{F}_2=[\mathbf{T}_2,\mathbf{T}_1,c\mathbf{T}_2]\in\mathbb{R}^{n\times3}$. Then we have

$$\mathbf{\Xi}=\Big(\big(\mathbf{J}+\mathbf{1}_n\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}/\sqrt{d}\big)\big(\mathbf{J}+\mathbf{1}_n\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}/\sqrt{d}\big)^\top+\lambda\mathbf{I}_n\Big)^{-1}$$

$$=\big(\mathbf{E}_0+\mathbf{F}_1\mathbf{F}_2^\top\big)^{-1}=\ \mathbf{E}_0^{-1}-\mathbf{E}_0^{-1}\mathbf{F}_1(\mathbf{I}_3+\mathbf{F}_2^\top\mathbf{E}_0^{-1}\mathbf{F}_1)^{-1}\mathbf{F}_2^T\mathbf{E}_0^{-1}, \tag{I.21}$$

where the last equality follows from the Sherman-Morrison-Woodbury formula. Moreover, we have

$$\begin{aligned}\mathbf{Z}\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\mathbf{Z}^\top=&\Big(\mathbf{J}+\mathbf{1}_n\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}/\sqrt{d}\Big)\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\Big(\mathbf{J}+\mathbf{1}_n\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}/\sqrt{d}\Big)^\top\\=&d\big(\mathbf{T}_1\mathbf{T}_1^\top+c\big(\mathbf{T}_2\mathbf{T}_1^\top+\mathbf{T}_1\mathbf{T}_2^\top\big)+c^2\mathbf{T}_2\mathbf{T}_2^\top\big)\\=&d\cdot\big(\mathbf{T}_1+c\mathbf{T}_2\big)\big(\mathbf{T}_1+c\mathbf{T}_2\big)^\top.\end{aligned} \tag{I.22}$$

Plugging (I.21) and (I.22) into (I.20), we obtain

$$\begin{aligned}\mathcal{D}=&\mathrm{tr}\Big(\big(\mathbf{T}_1+c\mathbf{T}_2\big)^\top\big(\mathbf{E}_0^{-1}-\mathbf{E}_0^{-1}\mathbf{F}_1(\mathbf{I}_3+\mathbf{F}_2^\top\mathbf{E}_0^{-1}\mathbf{F}_1)^{-1}\mathbf{F}_2^T\mathbf{E}_0^{-1}\big)\\&\cdot\mathbf{\Gamma}\big(\mathbf{E}_0^{-1}-\mathbf{E}_0^{-1}\mathbf{F}_1(\mathbf{I}_3+\mathbf{F}_2^\top\mathbf{E}_0^{-1}\mathbf{F}_1)^{-1}\mathbf{F}_2^T\mathbf{E}_0^{-1}\big)\big(\mathbf{T}_1+c\mathbf{T}_2\big)\Big).\end{aligned}$$

With similar calculation as in the proof of Lemma I.3, we obtain that

$$\mathcal{D}=\frac{G_{11}(1+K_{12})^2+G_{22}(c-K_{11})^2+2G_{12}(1+K_{12})(c-K_{11})}{\big(1+2K_{12}+K_{12}^2+cK_{22}-K_{11}K_{22}\big)^2}. \tag{I.23}$$

We then estimate the order for K_{11}, K_{12}, K_{22}, G_{11}, G_{12} and G_{22}, respectively. For K_{11}, apparently we have $K_{11}>0$. Moreover,

$$\begin{aligned}c-K_{11}=&\frac{1}{d}\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\big(\mathbf{I}_N-\mathbf{J}^\top(\mathbf{J}\mathbf{J}^\top+\lambda\mathbf{I}_n)^{-1}\mathbf{J}\big)\mathbf{\Lambda}_{d,0}\mathbf{1}_N\\\geq&c\big(1-\lambda_{\max}(\mathbf{J}^\top(\mathbf{J}\mathbf{J}^\top+\lambda\mathbf{I}_n)^{-1}\mathbf{J})\big)=\frac{c\lambda}{\lambda+\|\mathbf{J}\mathbf{J}^\top\|_{\mathrm{op}}}>0.\end{aligned}$$

Therefore we have

$$c\geq c-K_{11}\geq\frac{c\lambda}{\lambda+\|\mathbf{J}\mathbf{J}^\top\|_{\mathrm{op}}}>0. \tag{I.24}$$

Similarly, for K_{12} and K_{22} we have

$$|K_{12}|\leq\|(\mathbf{J}\mathbf{J}^\top+\lambda\mathbf{I}_n)^{-1}\mathbf{J}^\top\|_{\mathrm{op}}\|\mathbf{1}_n\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}/\sqrt{d}\|_{\mathrm{op}}=O_d(\sqrt{d/\lambda}), \tag{I.25}$$

$$K_{22}\geq n\lambda_{\min}\big(\big(\mathbf{J}\mathbf{J}^\top+\lambda\mathbf{I}_n\big)^{-1}\big)=\Omega(d)/(\|\mathbf{J}\mathbf{J}^\top\|_{\mathrm{op}}+\lambda). \tag{I.26}$$

Plugging (I.24), (I.25), (I.26) into (I.23) then gives

$$\begin{aligned}|\mathcal{D}|=&\frac{\big|G_{11}(1+K_{12})^2+G_{22}(c-K_{11})^2+2G_{12}(1+K_{12})(c-K_{11})\big|}{\big[(1+K_{12})^2+K_{22}\cdot(c-K_{11})\big]^2}\\\leq&\frac{\big|G_{11}(1+K_{12})^2+G_{22}(c-K_{11})^2+2G_{12}(1+K_{12})(c-K_{11})\big|}{\big[K_{22}\cdot(c-K_{11})\big]^2}\\\leq&O_d(1)\cdot\frac{|G_{11}|\cdot d+|G_{12}|\cdot\sqrt{d}+|G_{22}|\cdot c^2}{d^2/(\lambda+\|\mathbf{J}\mathbf{J}^\top\|_{\mathrm{op}})^4},\end{aligned}$$

where we utilize the upper and lower bounds in (I.24), (I.25), (I.26) to obtain the last inequality. Now recall that

$$\begin{aligned}
G_{11} =& \mathbf{1}_N^\top \mathbf{\Lambda}_{d,0} \mathbf{J}^\top (\mathbf{J}\mathbf{J}^\top + \lambda \mathbf{I}_n)^{-1} \mathbf{\Gamma} (\mathbf{J}\mathbf{J}^\top + \lambda \mathbf{I}_n)^{-1} \mathbf{J} \mathbf{\Lambda}_{d,0} \mathbf{1}_N / d, \\
G_{12} =& \mathbf{1}_N^\top \mathbf{\Lambda}_{d,0} \mathbf{J}^\top (\mathbf{J}\mathbf{J}^\top + \lambda \mathbf{I}_n)^{-1} \mathbf{\Gamma} (\mathbf{J}\mathbf{J}^\top + \lambda \mathbf{I}_n)^{-1} \mathbf{1}_n / \sqrt{d}, \\
G_{22} =& \mathbf{1}_n (\mathbf{J}\mathbf{J}^n + \lambda \mathbf{I}_n)^{-1} \mathbf{\Gamma} (\mathbf{J}\mathbf{J}^n + \lambda \mathbf{I}_n)^{-1} \mathbf{1}_n.
\end{aligned}$$

Therefore we have

$$\mathbb{E}\big[|G_{11}|^k\big]^{1/k} \le c \cdot O_d(\lambda^{-1}) \cdot \big[\mathbb{E}\|\mathbf{\Gamma}\|_{\rm op}^k\big]^{1/k} = O_d(1), \tag{I.27}$$

$$\mathbb{E}\big[|G_{12}|^k\big]^{1/k} \le O_d(\lambda^{-3/2}) \cdot \big[\mathbb{E}\|\mathbf{\Gamma}\|_{\rm op}^k\big]^{1/k} \cdot \|\mathbf{1}_n \mathbf{1}_N^\top \mathbf{\Lambda}_{d,0} / \sqrt{d}\|_{\rm op} = O_d(\sqrt{d}), \tag{I.28}$$

$$\mathbb{E}\big[|G_{22}|^k\big]^{1/k} \le O_d(\lambda^{-2}) \cdot \big[\mathbb{E}\|\mathbf{\Gamma}\|_{\rm op}^k\big]^{1/k} \cdot \|\mathbf{1}_n \mathbf{1}_n^\top\|_{\rm op} = O_d(d). \tag{I.29}$$

By the triangle inequality of the L_k-norm $\mathbb{E}[|\cdot|^k]^{1/k}$, we have

$$\begin{aligned}
\big(\mathbb{E}|\mathcal{D}|^k\big)^{1/k} \le& O_d(1) \cdot \frac{\mathbb{E}\big[|G_{11}|^k\big]^{1/k} \cdot d + \mathbb{E}\big[|G_{12}|^k\big]^{1/k} \cdot \sqrt{d} + \mathbb{E}\big[|G_{22}|^k\big]^{1/k} \cdot c^2}{d^2/(\lambda + \|\mathbf{J}\mathbf{J}^\top\|_{\rm op})^4} \\
=& O_d(1) \cdot \frac{d}{d^2/(\lambda + \|\mathbf{J}\mathbf{J}^\top\|_{\rm op})^4} \\
=& O_d\bigg(\frac{\big(\lambda + \|\mathbf{J}\mathbf{J}^\top\|_{\rm op}\big)^4}{d}\bigg) = O_d\bigg(\frac{\exp(C\sqrt{\log d})}{d}\bigg) = o_d(1),
\end{aligned}$$

where the first equality follows by (I.27), (I.28) and (I.29). This completes the proof.

I.3 Proof of Lemma I.4

The first result follows by the rotation invariance of the learning problem. For any $\boldsymbol{\beta}_d = [F_0, \boldsymbol{\beta}_{1,d}^\top]^\top$ and $\widetilde{\boldsymbol{\beta}}_d = [F_0, \widetilde{\boldsymbol{\beta}}_{1,d}^\top]^\top$ with $\boldsymbol{\beta}_{1,d}, \widetilde{\boldsymbol{\beta}}_{1,d} \in F_{1,d} \cdot \mathbb{S}^{d-1}$, there exists an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}\boldsymbol{\beta}_{1,d} = \widetilde{\boldsymbol{\beta}}_{1,d}$. Then by definition, we have

$$R_d(\mathbf{X}\mathbf{P}, \mathbf{\Theta}\mathbf{P}, \lambda, \boldsymbol{\beta}_d, \varepsilon) = R_d(\mathbf{X}, \mathbf{\Theta}, \lambda, \widetilde{\boldsymbol{\beta}}_d, \varepsilon).$$

Moreover, it is easy to check that

$$\overline{R}_d(\mathbf{X}\mathbf{P}, \mathbf{\Theta}\mathbf{P}, \lambda, F_{1,d}, \tau) = \overline{R}_d(\mathbf{X}, \mathbf{\Theta}, \lambda, F_{1,d}, \tau).$$

Since $(\mathbf{X}\mathbf{P}, \mathbf{\Theta}\mathbf{P}) \overset{\rm d}{=} (\mathbf{X}, \mathbf{\Theta})$, we see that conditional to $\boldsymbol{\beta}_d, \widetilde{\boldsymbol{\beta}}_d$, we have

$$R_d(\mathbf{X}, \mathbf{\Theta}, \lambda, \boldsymbol{\beta}_d, \varepsilon) - \overline{R}_d(\mathbf{X}, \mathbf{\Theta}, \lambda, F_{1,d}, \tau) \overset{\rm d}{=} R_d(\mathbf{X}, \mathbf{\Theta}, \lambda, \widetilde{\boldsymbol{\beta}}_d, \varepsilon) - \overline{R}_d(\mathbf{X}, \mathbf{\Theta}, \lambda, F_{1,d}, \tau).$$

This implies the first result in Lemma I.4.

If we assume that $\widetilde{\boldsymbol{\beta}}_{1,d} \sim \mathrm{N}(\mathbf{0}, [F_{1,d}^2/d]\mathbf{I}_d)$, then $F_{1,d} \cdot \widetilde{\boldsymbol{\beta}}_{1,d}/\|\widetilde{\boldsymbol{\beta}}_{1,d}\|_2 \sim F_{1,d} \cdot \mathrm{Unif}(\mathbb{S}^{d-1})$. The proof of the second result in Lemma I.4 from Gaussian $\widetilde{\boldsymbol{\beta}}_{1,d}$ to spherical $\widetilde{\boldsymbol{\beta}}_{1,d}$ differs by the factor $\|\widetilde{\boldsymbol{\beta}}_{1,d}\|_2/F_{1,d}$. Note that in high dimensions, the norm of Gaussian $\widetilde{\boldsymbol{\beta}}_{1,d}$ ($\|\widetilde{\boldsymbol{\beta}}_{1,d}\|_2$) is tightly concentrated on $F_{1,d}$. Therefore, it is not hard to translate the proof from Gaussian version to spherical version.

Based on the analysis above, without loss of generality we could assume $\widetilde{\boldsymbol{\beta}}_{1,d} \sim \mathrm{N}(\mathbf{0}, [F_{1,d}^2/d]\mathbf{I}_d)$ in the following of the proof. The lemma below helps us further handle the quadratic form of the variance which appears later.

Lemma I.7. *Let $\mathbf{A} \in \mathbb{R}^{d\times d}$, and define the random vector $\mathbf{h} \sim \mathrm{N}(0, (F_{1,d}^2/d)\mathbf{I}_d)$. Then we have*

$$\mathrm{Var}_{\mathbf{h}}(\mathbf{h}^\top \mathbf{A}\mathbf{h}) = \frac{F_{1,d}^4}{d^2}\big(\|\mathbf{A}\|_F^2 + \mathrm{tr}(\mathbf{A}^2)\big).$$

The proof of Lemma I.7 is given at the end of this section. With this lemma, we are well-prepared to prove the second result in Lemma I.4. Recall the definitions

$$\boldsymbol{\sigma}(\mathbf{x}) = \big(\sigma_1(\mathbf{x}^\top \boldsymbol{\Theta}_1^\top/\sqrt{d}), \sigma_2(\mathbf{x}^\top \boldsymbol{\Theta}_2^\top/\sqrt{d})\big)^\top \in \mathbb{R}^N, \qquad \boldsymbol{\Upsilon} = (\mathbf{Z}^\top\mathbf{Z} + \lambda \mathbf{I}_N)^{-1},$$

and $\widetilde{\mathbf{V}} = \mathbb{E}_{\mathbf{x}}\boldsymbol{\sigma}(\mathbf{x})(\mathbf{x}^\top\widetilde{\boldsymbol{\beta}}_{1,d}+F_0)$, $\mathbf{U} = \mathbb{E}_{\mathbf{x}}\boldsymbol{\sigma}(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})^\top$. By the definition of the risk $R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \widetilde{\boldsymbol{\beta}}_{1,d}, \boldsymbol{\varepsilon})$, we have

$$\begin{aligned} R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \widetilde{\boldsymbol{\beta}}_{1,d}, \boldsymbol{\varepsilon}) &= \mathbb{E}_{\mathbf{x}}\big(\mathbf{x}^\top\widetilde{\boldsymbol{\beta}}_d + F_0 - \widehat{\mathbf{a}}(\lambda)^\top\boldsymbol{\sigma}(\mathbf{x})\big)^2 \\ &= F_0^2 + F_{1,d}^2 - 2\Gamma_1 + \Gamma_2 + \Gamma_3 - 2\Gamma_4 + 2\Gamma_5, \end{aligned} \tag{I.30}$$

where

$$\begin{aligned} &\boldsymbol{f} = \mathbf{X}\widetilde{\boldsymbol{\beta}}_{1,d} + \mathbf{1}_n F_0, \qquad \Gamma_1 = \boldsymbol{f}^\top\mathbf{Z}\boldsymbol{\Upsilon}\widetilde{\mathbf{V}}/\sqrt{d}, \qquad \Gamma_2 = \boldsymbol{f}^\top\big[\mathbf{U}\big]_{\mathbf{Z}}\boldsymbol{f}/d, \\ &\Gamma_2 = \boldsymbol{\varepsilon}^\top\big[\mathbf{U}\big]_{\mathbf{Z}}\boldsymbol{\varepsilon}/d, \qquad \Gamma_4 = \boldsymbol{\varepsilon}^\top\mathbf{Z}\boldsymbol{\Upsilon}\widetilde{\mathbf{V}}/\sqrt{d}, \qquad \Gamma_5 = \boldsymbol{\varepsilon}^\top\big[\mathbf{U}\big]_{\mathbf{Z}}\boldsymbol{f}/d. \end{aligned}$$

Note that the terms F_0^2 and $F_{1,d}^2$ in (I.30) are constants, and therefore do not contribute to the variance of $R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \widetilde{\boldsymbol{\beta}}_{1,d}, \boldsymbol{\varepsilon})$. In the following, we aim to show that $\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\big[\mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d},\boldsymbol{\varepsilon}}(\Gamma_k)\big] = o_d(1)$ for $k \in [5]$. Consider first the variance of Γ_1. From Lemma I.2, we have $\widetilde{\mathbf{V}} = \boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\widetilde{\boldsymbol{\beta}}_{1,d} + \boldsymbol{\Lambda}_{d,0}\mathbf{1}_N F_0$. Then

$$\begin{aligned} \mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d}}(\Gamma_1) &= \mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d}}\Big((\mathbf{X}\widetilde{\boldsymbol{\beta}}_{1,d} + \mathbf{1}_n F_0)^\top\mathbf{Z}\boldsymbol{\Upsilon}\big(\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\widetilde{\boldsymbol{\beta}}_{1,d} + \boldsymbol{\Lambda}_{d,0}\mathbf{1}_N F_0\big)/\sqrt{d}\Big) \\ &= \frac{1}{d}\mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d}}\Big(\widetilde{\boldsymbol{\beta}}_{1,d}^\top\mathbf{X}^T\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\widetilde{\boldsymbol{\beta}}_{1,d} + \widetilde{\boldsymbol{\beta}}_{1,d}^\top\mathbf{X}\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N F_0 \\ &\quad + F_0\mathbf{1}_n^\top\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\widetilde{\boldsymbol{\beta}}_{1,d} + F_0\mathbf{1}_n^T\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N F_0\Big) \\ &\le \frac{4}{d}\mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d}}\Big(\widetilde{\boldsymbol{\beta}}_{1,d}^\top\mathbf{X}^T\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\widetilde{\boldsymbol{\beta}}_{1,d}\Big) + \frac{4}{d}\mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d}}\Big(\widetilde{\boldsymbol{\beta}}_{1,d}^\top\mathbf{X}\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N F_0\Big) \\ &\quad + \frac{4}{d}\mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d}}\Big(F_0\mathbf{1}_n^\top\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\widetilde{\boldsymbol{\beta}}_{1,d}\Big) + 0 \\ &\le 4F_{1,d}^4\cdot\underbrace{\frac{1}{d^3}\Big(\|\mathbf{X}^T\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\|_F^2 + \mathrm{tr}\big(\mathbf{X}^T\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\mathbf{X}^T\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\big)\Big)}_{I_1} \\ &\quad + 4F_{1,d}^2F_0^2\cdot\underbrace{\Big(\frac{1}{d}\mathrm{tr}\Big(\frac{\mathbf{X}\mathbf{X}^\top}{d}\big[\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}\big]_{\mathbf{Z}}\Big) + \frac{1}{d}\mathrm{tr}\Big(\mathbf{1}_n\mathbf{1}_n^\top\Big[\boldsymbol{\Lambda}_{d,1}\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\top}{d}\boldsymbol{\Lambda}_{d,1}\Big]_{\mathbf{Z}}\Big)\Big)}_{I_2}. \end{aligned}$$

The first inequality holds from $\mathrm{Var}(a+b) \le 2\,\mathrm{Var}(a) + 2\,\mathrm{Var}(b)$, I_1 comes from Lemma I.7 and I_2 comes from $\mathrm{Var}(a) \le \mathbb{E}a^2$. Note that $\|\boldsymbol{\Lambda}_{d,1}\|_{\mathrm{op}} = O_d(1/\sqrt{d})$ and $\|\mathbf{Z}\boldsymbol{\Upsilon}\|_{\mathrm{op}} \le 1/(2\sqrt{\lambda})$, we conclude

that

$$\begin{aligned}\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}|I_1| &\le \Big|\frac{1}{d^3}\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\mathrm{tr}(\mathbf{X}^T\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\boldsymbol{\Theta}^\top\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Upsilon}\mathbf{Z}^\top\mathbf{X})\Big| + \Big|\frac{1}{d^3}\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\mathrm{tr}\big(\mathbf{X}^T\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\mathbf{X}^T\mathbf{Z}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\big)\Big| \\ &\le \frac{1}{4\lambda}\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\Big\|\boldsymbol{\Lambda}_{d,1}\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\top}{d}\boldsymbol{\Lambda}_{d,1}\Big\|_{\mathrm{op}}\cdot\Big\|\frac{\mathbf{X}\mathbf{X}^\top}{d}\Big\|_{\mathrm{op}} + \frac{1}{4\lambda}\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\Big\|\frac{\boldsymbol{\Lambda}_{d,1}\boldsymbol{\Theta}\mathbf{X}^\top}{d}\Big\|_{\mathrm{op}}^2 = o_d(1).\end{aligned}$$

Furthermore from Lemma I.3, we have

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}|I_2| = \mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\Big|\frac{1}{d}\mathrm{tr}\Big(\frac{\mathbf{X}\mathbf{X}^\top}{d}\big[\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}\big]_{\mathbf{Z}}\Big) + \frac{1}{d}\mathrm{tr}\Big(\mathbf{1}_n\mathbf{1}_n^\top\Big[\boldsymbol{\Lambda}_{d,1}\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\top}{d}\boldsymbol{\Lambda}_{d,1}\Big]_{\mathbf{Z}}\Big)\Big| = o_d(1).$$

Thus we obtain $\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta}}\big(\mathrm{Var}_{\boldsymbol{\beta}_d}(\Gamma_1)\big) = o_d(1)$. Similarly, we have for Γ_2,

$$\begin{aligned}\mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d}}(\Gamma_2) =& \mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d}}\Big((\mathbf{X}\widetilde{\boldsymbol{\beta}}_{1,d}+\mathbf{1}_nF_0)^\top[\mathbf{U}]_{\mathbf{Z}}(\mathbf{X}\widetilde{\boldsymbol{\beta}}_{1,d}+\mathbf{1}_nF_0)/d\Big) \\ =& \frac{1}{d^2}\mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d}}\Big(\widetilde{\boldsymbol{\beta}}_{1,d}^\top\mathbf{X}^T[\mathbf{U}]_{\mathbf{Z}}\mathbf{X}\widetilde{\boldsymbol{\beta}}_{1,d} + \widetilde{\boldsymbol{\beta}}_{1,d}^\top\mathbf{X}^\top[\mathbf{U}]_{\mathbf{Z}}\mathbf{1}_nF_0 \\ &+ F_0\mathbf{1}_n^\top[\mathbf{U}]_{\mathbf{Z}}\mathbf{X}\widetilde{\boldsymbol{\beta}}_{1,d} + F_0\mathbf{1}_n^T[\mathbf{U}]_{\mathbf{Z}}\mathbf{1}_nF_0\Big) \\ \le& \frac{4}{d^2}\mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d}}\Big(\widetilde{\boldsymbol{\beta}}_{1,d}^\top\mathbf{X}^T[\mathbf{U}]_{\mathbf{Z}}\mathbf{X}\widetilde{\boldsymbol{\beta}}_{1,d}\Big) + \frac{8}{d^2}\mathrm{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d}}\Big(\widetilde{\boldsymbol{\beta}}_{1,d}^\top\mathbf{X}^\top[\mathbf{U}]_{\mathbf{Z}}\mathbf{1}_nF_0\Big) \\ \le& 8F_{1,d}^4\cdot\underbrace{\frac{1}{d^4}\mathrm{tr}\Big(\mathbf{X}^T[\mathbf{U}]_{\mathbf{Z}}\mathbf{X}\mathbf{X}^T[\mathbf{U}]_{\mathbf{Z}}\mathbf{X}\Big)}_{I_3} + 8F_{1,d}^2F_0^2\cdot\underbrace{\frac{1}{d^2}\mathrm{tr}\Big([\mathbf{U}]_{\mathbf{Z}}\mathbf{1}_n\mathbf{1}_n^\top[\mathbf{U}]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\Big)}_{I_4}.\end{aligned}$$

The first inequality holds from $\mathrm{Var}(a+b)\le 2\,\mathrm{Var}(a)+2\,\mathrm{Var}(b)$, I_3 comes from Lemma I.7 and the symmetric of $\mathbf{X}^T[\mathbf{U}]_{\mathbf{Z}}\mathbf{X}$, and I_4 comes from $\mathrm{Var}(a)\le\mathbb{E}a^2$. Define $\boldsymbol{\Gamma}_{\mathbf{U}} = \mathbf{U} - \boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}$, from Lemma I.2, $\mathbb{E}\|\boldsymbol{\Gamma}_{\mathbf{U}}\|_{\mathrm{op}}^2 = O_d(1)$. By replacing $\mathbf{U}$ by $\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}+\boldsymbol{\Gamma}_{\mathbf{U}}$ in the terms I_3 and I_4, we obtain the following equalities:

$$\begin{aligned}I_3 =& \frac{1}{d^2}\mathrm{tr}\Big(\big[\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}+\boldsymbol{\Gamma}_{\mathbf{U}}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\big[\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}+\boldsymbol{\Gamma}_{\mathbf{U}}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\Big) \\ =& \underbrace{\frac{1}{d^2}\mathrm{tr}\Big(\big[\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\big[\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\Big)}_{J_1} \\ &+ \underbrace{\frac{2}{d^2}\mathrm{tr}\Big(\big[\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\big[\boldsymbol{\Gamma}_{\mathbf{U}}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\Big)}_{J_2} + \underbrace{\frac{1}{d^2}\mathrm{tr}\Big(\big[\boldsymbol{\Gamma}_{\mathbf{U}}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\big[\boldsymbol{\Gamma}_{\mathbf{U}}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\Big)}_{J_3}, \\ I_4 =& \frac{1}{d^2}\mathrm{tr}\Big(\big[\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}+\boldsymbol{\Gamma}_{\mathbf{U}}\big]_{\mathbf{Z}}\mathbf{1}_n\mathbf{1}_n^\top\big[\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}+\boldsymbol{\Gamma}_{\mathbf{U}}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\Big) \\ =& \underbrace{\frac{1}{d^2}\mathrm{tr}\Big(\big[\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}\big]_{\mathbf{Z}}\mathbf{1}_n\mathbf{1}_n^\top\big[\boldsymbol{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{\Lambda}_{d,0}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\Big)}_{K_1}\end{aligned}$$

$$+\underbrace{\frac{2}{d^2}\mathrm{tr}\left(\left[\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\right]_{\mathbf{Z}}\mathbf{1}_n\mathbf{1}_n^\top\left[\mathbf{\Gamma_U}\right]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right)}_{K_2}+\underbrace{\frac{1}{d^2}\mathrm{tr}\left(\left[\mathbf{\Gamma_U}\right]_{\mathbf{Z}}\mathbf{1}_n\mathbf{1}_n^\top\left[\mathbf{\Gamma_U}\right]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right)}_{K_3}.$$

We investigate the terms K_i, $i=1,2,3$. The investigation of terms J_i, $i=1,2,3$ are quite similar, we omit the proof for J_i for brevity. Consider first the term K_2. Due to $\mathbb{E}\|\mathbf{\Gamma_U}\|_{\mathrm{op}}^2 = O_d(1)$, it is true that

$$\begin{aligned}\left(\mathbb{E}\left\|\frac{\mathbf{1}_n\mathbf{1}_n^\top}{d}\left[\mathbf{\Gamma_U}\right]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right\|_{\mathrm{op}}^2\right)^{1/2} =& O_d(1)\cdot\left(\mathbb{E}\left\|\mathbf{\Gamma_U}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right\|_{\mathrm{op}}^2\right)^{1/2}\\ =& O_d(1)\cdot\left(\mathbb{E}\left\|\mathbf{\Gamma_U}\right\|_{\mathrm{op}}^2\right)^{1/2}\cdot\left(\mathbb{E}\left\|\frac{\mathbf{X}\mathbf{X}^\top}{d}\right\|_{\mathrm{op}}^2\right)^{1/2} = O_d(1).\end{aligned}$$

The second equality comes from the independence of $\mathbf{\Gamma_U}$ and $\mathbf{X}$. Note that for any rank 1 matrix $\mathbf{A}$, $|\mathrm{tr}\mathbf{A}| = \|\mathbf{A}\|_{\mathrm{op}}$, the term K_2 has the property

$$\begin{aligned}\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}|K_2| &= \mathbb{E}_{\mathbf{X},\mathbf{\Theta}}\left|\frac{2}{d^2}\mathrm{tr}\left(\left[\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\right]_{\mathbf{Z}}\mathbf{1}_n\mathbf{1}_n^\top\left[\mathbf{\Gamma_U}\right]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right)\right|\\ &\le \mathbb{E}_{\mathbf{X},\mathbf{\Theta}}\left(\frac{2}{d}\left\|\left[\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\right]_{\mathbf{Z}}\right\|_{\mathrm{op}}\cdot\left\|\frac{\mathbf{1}_n\mathbf{1}_n^\top}{d}\left[\mathbf{\Gamma_U}\right]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right\|_{\mathrm{op}}\right)\\ &\le \frac{2}{d}\left(\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}\left\|\left[\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\right]_{\mathbf{Z}}\right\|_{\mathrm{op}}^2\cdot\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}\left\|\frac{\mathbf{1}_n\mathbf{1}_n^\top}{d}\left[\mathbf{\Gamma_U}\right]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right\|_{\mathrm{op}}^2\right)^{1/2}\\ &= o_d(1)\cdot O_d(1) = o_d(1).\end{aligned}$$

The equality comes from the estimation of $\mathcal{D}$ in Lemma I.3 . For the term K_1, it is true from the estimation of $\mathcal{D}$ in Lemma I.3 that

$$\begin{aligned}&\frac{1}{d}\left(\mathbb{E}\left\|\frac{\mathbf{1}_n\mathbf{1}_n^\top}{d}\left[\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\right]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right\|_{\mathrm{op}}^2\right)^{1/2}\\ &\quad\le\frac{1}{d}\left\|\frac{\mathbf{1}_n\mathbf{1}_n^\top}{d}\right\|_{\mathrm{op}}\left(\mathbb{E}\left\|\left[\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\right]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right\|_{op}^2\right)^{1/2} = O_d(1)\cdot o_d(1) = o_d(1).\end{aligned}$$

By repeating the arguments used previously for the term K_2 but for the consideration of K_1, we have

$$\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}|K_1| = \frac{1}{d^2}\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}\left|\mathrm{tr}\left(\left[\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\right]_{\mathbf{Z}}\mathbf{1}_n\mathbf{1}_n^\top\left[\mathbf{\Lambda}_{d,0}\mathbf{1}_N\mathbf{1}_N^\top\mathbf{\Lambda}_{d,0}\right]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right)\right| = o_d(1).$$

For the term K_3, similarly we have

$$\begin{aligned}\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}|K_3| &= \mathbb{E}_{\mathbf{X},\mathbf{\Theta}}\left|\frac{1}{d^2}\mathrm{tr}\left(\left[\mathbf{\Gamma_U}\right]_{\mathbf{Z}}\mathbf{1}_n\mathbf{1}_n^\top\left[\mathbf{\Gamma_U}\right]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right)\right|\\ &\le \frac{1}{d^2}\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}\left(\left\|\left[\mathbf{\Gamma_U}\right]_{\mathbf{Z}}\mathbf{1}_n\mathbf{1}_n^\top\right\|_{\mathrm{op}}\cdot\left\|\left[\mathbf{\Gamma_U}\right]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right\|_{\mathrm{op}}\right)\end{aligned}$$

$$
\begin{aligned}
&\le \frac{1}{d^2}\bigg(\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}\big\|\big[\mathbf{\Gamma_U}\big]_{\mathbf{Z}}\mathbf{1}_n\mathbf{1}_n^\top\big\|_{\rm op}^2\cdot\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}\bigg\|\big[\mathbf{\Gamma_U}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\bigg\|_{\rm op}^2\bigg)^{1/2}\\
&\le \frac{1}{d^2}\big(\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}\big\|\big[\mathbf{\Gamma_U}\big]_{\mathbf{Z}}\mathbf{1}_n\mathbf{1}_n^\top\big\|_{\rm op}^2\big)^{1/2}\cdot O_d(1)\\
&= o_d(1)\cdot O_d(1) = o_d(1).
\end{aligned}
$$

Now we conclude that K_1, K_2 and K_3 are all small terms under the expectation over $\mathbf{X}$ and $\mathbf{\Theta}$, we immediately get that

$$\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}|I_4| = o_d(1).$$

Similarly we get $\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}|I_3| = o_d(1)$, thus we conclude that $\mathbb{E}_{\mathbf{X},\mathbf{\Theta}}\operatorname{Var}_{\widetilde{\boldsymbol{\beta}}_{1,d}}(\Gamma_2) = o_d(1)$. We omit the other terms for brevity. The proof of Lemma I.4 is complete.

Proof of Lemma I.7

We have

$$\mathbb{E}[\mathbf{h}^\top\mathbf{A}\mathbf{h}] = \mathbb{E}\mathrm{tr}(\mathbf{A}\mathbf{h}\mathbf{h}^\top) = \frac{F_{1,d}^2}{d}\mathrm{tr}(\mathbf{A}).$$

Hence we have

$$
\begin{aligned}
\operatorname{Var}(\mathbf{h}^\top\mathbf{A}\mathbf{h}) &= \sum_{i_1,i_2,i_3,i_4}\mathbb{E}\Big[\mathbf{h}_{i_1}\mathbf{A}_{i_1,i_2}\mathbf{h}_{i_2}\mathbf{h}_{i_3}\mathbf{A}_{i_3,i_4}\mathbf{h}_{i_4}\Big] - \frac{F_{1,d}^4}{d^2}\mathrm{tr}(\mathbf{A})^2\\
&= \Bigg\{\Bigg(\sum_{\substack{i_1=i_2,i_3=i_4\\ i_1=i_3}} + \sum_{\substack{i_1=i_2,i_3=i_4\\ i_1\neq i_3}} + \sum_{\substack{i_1=i_3,i_2=i_4\\ i_1\neq i_2}} + \sum_{\substack{i_1=i_4,i_2=i_3\\ i_1\neq i_2}}\Bigg)\mathbb{E}\Big[\mathbf{h}_{i_1}\mathbf{A}_{i_1,i_2}\mathbf{h}_{i_2}\mathbf{h}_{i_3}\mathbf{A}_{i_3,i_4}\mathbf{h}_{i_4}\Big]\Bigg\}\\
&\quad - \frac{F_{1,d}^4}{d^2}\mathrm{tr}(\mathbf{A})^2\\
&= \frac{F_{1,d}^4}{d^2}\bigg(\sum_{i=1}^d\mathbf{A}_{i,i}^2\cdot\frac{d^2}{F_{1,d}^4}(\mathbb{E}h_i^4) + \sum_{i\neq j}\mathbf{A}_{i,i}\mathbf{A}_{j,j} + \sum_{i\neq j}(\mathbf{A}_{i,j}\mathbf{A}_{i,j}+\mathbf{A}_{i,j}\mathbf{A}_{j,i}) - \mathrm{tr}(\mathbf{A})^2\bigg)\\
&= \frac{F_{1,d}^4}{d^2}\bigg(\sum_{i=1}^d 3\mathbf{A}_{i,i}^2 + \sum_{i\neq j}(\mathbf{A}_{i,j}\mathbf{A}_{i,j}+\mathbf{A}_{i,j}\mathbf{A}_{j,i}) + \sum_{i\neq j}\mathbf{A}_{i,i}\mathbf{A}_{j,j} - \mathrm{tr}(\mathbf{A})^2\bigg)\\
&= \frac{F_{1,d}^4}{d^2}\bigg(\sum_{i,j}(\mathbf{A}_{i,j}\mathbf{A}_{i,j}+\mathbf{A}_{i,j}\mathbf{A}_{j,i}) + \sum_{i,j}\mathbf{A}_{i,i}\mathbf{A}_{j,j} - \mathrm{tr}(\mathbf{A})^2\bigg) = \frac{F_{1,d}^4}{d^2}\big(\|\mathbf{A}\|_F^2 + \mathrm{tr}(\mathbf{A}^2)\big).
\end{aligned}
$$

This proves Lemma I.7.

II. Proof of Proposition A.4

We first recall the following matrix differential rules:

$$\frac{\partial\det(\mathbf{Y})}{\partial x} = \det(\mathbf{Y})\cdot\mathrm{tr}\Big(\mathbf{Y}^{-1}\cdot\frac{\partial\mathbf{Y}}{\partial x}\Big),\quad \frac{\partial\mathbf{Y}^{-1}}{\partial x} = -\mathbf{Y}^{-1}\frac{\partial\mathbf{Y}}{\partial x}\mathbf{Y}^{-1}. \tag{II.1}$$

Let q_i, q_j be the elements in the vector $\mathbf{q}$. Now the matrix $\mathbf{A} = \mathbf{A}(\mathbf{q}, \boldsymbol{\mu})$ (see Definition A.3) is linear in $\mathbf{q}$, thus $\frac{\partial^2 \mathbf{A}}{\partial q_i \partial q_j} = 0$. Therefore by the definition $G_d(\xi) = \frac{1}{d}\log\det\big(\mathbf{A} - \xi\mathbf{I}\big)$ and the matrix derivative rules in (II.1), we have

$$\frac{\partial G_d}{\partial q_i} = \frac{1}{d}\mathrm{tr}\Big(\big(\mathbf{A} - \xi\mathbf{I}\big)^{-1}\frac{\partial \mathbf{A}}{\partial q_i}\Big), \tag{II.2}$$

$$\frac{\partial^2 G_d}{\partial q_i \partial q_j} = \frac{1}{d}\mathrm{tr}\left(\frac{\partial\big(\mathbf{A} - \xi\mathbf{I}\big)^{-1}}{\partial q_j}\frac{\partial \mathbf{A}}{\partial q_i}\right) = -\frac{1}{d}\mathrm{tr}\Big(\big(\mathbf{A} - \xi\mathbf{I}\big)^{-1}\frac{\partial \mathbf{A}}{\partial q_j}\big(\mathbf{A} - \xi\mathbf{I}\big)^{-1}\frac{\partial \mathbf{A}}{\partial q_i}\Big). \tag{II.3}$$

By the Schur complement formula, we further have

$$(\mathbf{A}(\mathbf{0}, \boldsymbol{\mu}) - \xi\mathbf{I}_P)^{-1} = \begin{bmatrix} -\xi\mathbf{I}_N & \mathbf{Z}^\top \\ \mathbf{Z} & -\xi\mathbf{I}_n \end{bmatrix}^{-1} = \begin{bmatrix} * & \mathbf{Z}^\top(\mathbf{Z}\mathbf{Z}^\top - \xi^2\mathbf{I}_n)^{-1} \\ (\mathbf{Z}\mathbf{Z}^\top - \xi^2\mathbf{I}_n)^{-1}\mathbf{Z} & * \end{bmatrix},$$

where we use $*$ to hide the irrelevant blocks in the matrix inverse. Moreover, by definition, it holds that

$$\frac{\partial \mathbf{A}(\mathbf{q}, \boldsymbol{\mu})}{\partial q_1} = \begin{bmatrix} \mathbf{0} & \frac{1}{d}\mathbf{M}_1\boldsymbol{\Theta}\mathbf{X}^\top \\ \frac{1}{d}\mathbf{X}\boldsymbol{\Theta}^\top\mathbf{M}_1 & \mathbf{0} \end{bmatrix}, \quad \frac{\partial \mathbf{A}(\mathbf{q}, \boldsymbol{\mu})}{\partial q_4} = \begin{bmatrix} \mathbf{M}_1\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\top}{d}\mathbf{M}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

$$\frac{\partial \mathbf{A}(\mathbf{q}, \boldsymbol{\mu})}{\partial q_2} = \begin{bmatrix} \mathbf{M}_*^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \frac{\partial \mathbf{A}(\mathbf{q}, \boldsymbol{\mu})}{\partial q_3} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix}, \quad \frac{\partial \mathbf{A}(\mathbf{q}, \boldsymbol{\mu})}{\partial q_5} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\mathbf{X}\mathbf{X}^\top}{d} \end{bmatrix}.$$

By plugging these derivatives into (II.2) and (II.3), we continue the calculation with $\xi = \xi^*$. Note that we have the identity $(\mathbf{Z}\mathbf{Z}^\top - (\xi^*)^2\mathbf{I}_n)^{-1}\mathbf{Z} = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{I}_N)^{-1} = \mathbf{Z}\boldsymbol{\Upsilon}$. Then by (II.2), we have

$$\begin{aligned}\frac{\partial G_d(\xi^*; \mathbf{q}, \boldsymbol{\mu})}{\partial q_1}\bigg|_{\mathbf{q}=\mathbf{0}} =& \frac{1}{d}\mathrm{tr}\left(\begin{bmatrix} * & \mathbf{Z}^\top(\mathbf{Z}\mathbf{Z}^\top - (\xi^*)^2\mathbf{I}_n)^{-1} \\ (\mathbf{Z}\mathbf{Z}^\top - (\xi^*)^2\mathbf{I}_n)^{-1}\mathbf{Z} & * \end{bmatrix}\begin{bmatrix} \mathbf{0} & \frac{1}{d}\mathbf{M}_1\boldsymbol{\Theta}\mathbf{X}^\top \\ \frac{1}{d}\mathbf{X}\boldsymbol{\Theta}^\top\mathbf{M}_1 & \mathbf{0} \end{bmatrix}\right) \\ =& \frac{1}{d}\mathrm{tr}\left(\begin{bmatrix} * & \boldsymbol{\Upsilon}\mathbf{Z}^\top \\ \mathbf{Z}\boldsymbol{\Upsilon} & * \end{bmatrix}\begin{bmatrix} \mathbf{0} & \frac{1}{d}\mathbf{M}_1\boldsymbol{\Theta}\mathbf{X}^\top \\ \frac{1}{d}\mathbf{X}\boldsymbol{\Theta}^\top\mathbf{M}_1 & \mathbf{0} \end{bmatrix}\right) = \frac{2}{d}\mathrm{tr}\mathbf{M}_1\frac{\boldsymbol{\Theta}\mathbf{X}^\top}{d}\mathbf{Z}\boldsymbol{\Upsilon},\end{aligned}$$

Similarly, by (II.3), we have

$$\begin{aligned}-\frac{\partial^2 G_d(\xi^*; \mathbf{q}, \boldsymbol{\mu})}{\partial q_4 \partial q_5}\bigg|_{\mathbf{q}=\mathbf{0}} =& \mathrm{tr}\left(\begin{bmatrix} * & \boldsymbol{\Upsilon}\mathbf{Z}^\top \\ \mathbf{Z}\boldsymbol{\Upsilon} & * \end{bmatrix}\begin{bmatrix} \mathbf{M}_1\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\top}{d}\mathbf{M}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} * & \boldsymbol{\Upsilon}\mathbf{Z}^\top \\ \mathbf{Z}\boldsymbol{\Upsilon} & * \end{bmatrix}\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\mathbf{X}\mathbf{X}^\top}{d} \end{bmatrix}\right) \\ =& \frac{1}{d}\mathrm{tr}\mathbf{Z}\boldsymbol{\Upsilon}\mathbf{M}_1\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\top}{d}\mathbf{M}_1\boldsymbol{\Upsilon}\mathbf{Z}^\top\frac{\mathbf{X}\mathbf{X}^\top}{d}, \\ -\frac{\partial^2 G_d(\xi^*; \mathbf{q}, \boldsymbol{\mu})}{\partial q_2 \partial q_5}\bigg|_{\mathbf{q}=\mathbf{0}} =& \mathrm{tr}\left(\begin{bmatrix} * & \boldsymbol{\Upsilon}\mathbf{Z}^\top \\ \mathbf{Z}\boldsymbol{\Upsilon} & * \end{bmatrix}\begin{bmatrix} \mathbf{M}_*\mathbf{M}_* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} * & \boldsymbol{\Upsilon}\mathbf{Z}^\top \\ \mathbf{Z}\boldsymbol{\Upsilon} & * \end{bmatrix}\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\mathbf{X}\mathbf{X}^\top}{d} \end{bmatrix}\right) \\ =& \frac{1}{d}\mathrm{tr}\mathbf{Z}\boldsymbol{\Upsilon}\mathbf{M}_*\mathbf{M}_*\boldsymbol{\Upsilon}\mathbf{Z}^\top\frac{\mathbf{X}\mathbf{X}^\top}{d}, \\ -\frac{\partial^2 G_d(\xi^*; \mathbf{q}, \boldsymbol{\mu})}{\partial q_3 \partial q_4}\bigg|_{\mathbf{q}=\mathbf{0}} =& \mathrm{tr}\left(\begin{bmatrix} * & \boldsymbol{\Upsilon}\mathbf{Z}^\top \\ \mathbf{Z}\boldsymbol{\Upsilon} & * \end{bmatrix}\begin{bmatrix} \mathbf{M}_1\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\top}{d}\mathbf{M}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} * & \boldsymbol{\Upsilon}\mathbf{Z}^\top \\ \mathbf{Z}\boldsymbol{\Upsilon} & * \end{bmatrix}\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix}\right) \\ =& \frac{1}{d}\mathrm{tr}\mathbf{Z}\boldsymbol{\Upsilon}\mathbf{M}_1\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\top}{d}\mathbf{M}_1\boldsymbol{\Upsilon}\mathbf{Z}^\top,\end{aligned}$$

$$-\frac{\partial^2 G_d(\xi^*;\mathbf{q},\boldsymbol{\mu})}{\partial q_2 \partial q_3}\bigg|_{\mathbf{q}=\mathbf{0}} = \mathrm{tr}\left(\begin{bmatrix} * & \boldsymbol{\Upsilon}\mathbf{Z}^\top \\ \mathbf{Z}\boldsymbol{\Upsilon} & * \end{bmatrix}\begin{bmatrix} \mathbf{M}_*\mathbf{M}_* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} * & \boldsymbol{\Upsilon}\mathbf{Z}^\top \\ \mathbf{Z}\boldsymbol{\Upsilon} & * \end{bmatrix}\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix}\right)$$
$$= \frac{1}{d}\mathrm{tr}\mathbf{Z}\boldsymbol{\Upsilon}\mathbf{M}_*\mathbf{M}_*\boldsymbol{\Upsilon}\mathbf{Z}^\top.$$

The above equations complete the proof of Proposition A.4.

III. Properties of the fixed point equation

In this section, we justify the definition of $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$ below Definition A.5, by proving that there exists a constant $\xi_0 > 0$, such that the fixed point equation (A.2) has a unique solution defined on $\{\xi : \Im(\xi) > \xi_0\}$ satisfying $|m_j(\xi)| \le 2\psi_j/\xi_0$ for $j = 1, 2, 3$. The result is given in the following lemma.

Lemma III.1. *Let $\mathsf{F}(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})$, $\mathbf{q} \in \mathcal{Q}$ be defined in Definition A.5, and $\mathbb{D}(r) = \{z : |z| < r\}$ be the disk of radius r in the complex plane. There exists $\xi_0 > 0$ such that, for any $\xi \in \mathbb{C}_+$ with $\Im(\xi) > \xi_0$, $\mathsf{F}(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})$ is $1/2$-Lipschitz continuous with respect to the ℓ_2 norm, and the map $\mathbf{m} \mapsto \mathsf{F}(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})$ admits a unique fixed point in $\mathbb{D}(2\psi_1/\xi_0) \times \mathbb{D}(2\psi_2/\xi_0) \times \mathbb{D}(2\psi_3/\xi_0)$.*

Lemma III.1 demonstrates that our definition of $\mathbf{m}$ in Subsection A.3 as the unique fixed point of F is valid.

Proof [Proof of Lemma III.1] We prove the existence and uniqueness of the solution by the Banach fixed point theorem when $\Im(\xi) \ge \xi_0$ for some sufficiently large ξ_0. To do so, we want to show that

1. $\mathsf{F}(\cdot;\mathbf{q},\boldsymbol{\mu})$ maps domain $\mathbb{D}(2\psi_1/\xi_0) \times \mathbb{D}(2\psi_2/\xi_0) \times \mathbb{D}(2\psi_3/\xi_0)$ into itself.
2. $\mathsf{F}(\cdot;\mathbf{q},\boldsymbol{\mu})$ is Lipschitz continuous with a Lipschitz constant smaller than 1.

For $\mathsf{F}_1(\cdot;\mathbf{q},\boldsymbol{\mu})$, by Definition A.5, we have

$$\mathsf{F}_1(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu}) = \frac{\psi_1}{-\xi + q_2\mu_{1,*}^2 + H_1(\mathbf{m};\mathbf{q},\boldsymbol{\mu})},$$

where

$$H_1(\mathbf{m};\mathbf{q},\boldsymbol{\mu}) = -\mu_{1,*}^2 m_3 + \frac{1}{m_1 + \frac{-\mu_{2,1}^2(1+q_1)^2 m_2 m_3 + (1+\mu_{2,1}^2 m_2 q_4)(1+m_3 q_5)}{\mu_{1,1}^2 q_4(1+m_3 q_5) - \mu_{1,1}^2(1+q_1)^2 m_3}}. \tag{III.1}$$

Note that $q_4, q_5 \le (1+q_1)/2$. Thus for small enough r_0, we have for any $\mathbf{m} \in \mathbb{D}(r_0)^3$

$$|H_1(\mathbf{m};\mathbf{q},\boldsymbol{\mu})| \le 2 + 2|q_4|\mu_{1,1}^2. \tag{III.2}$$

Now as long as $\xi_0 \ge 4 + 4|q_4|\mu_{1,1}^2$, it is clear that for ξ with $\Im(\xi) \ge \xi_0$ we have

$$\Im(\xi) \ge \xi_0/2 + \xi_0/2 \ge \xi_0/2 + 2 + 2|q_4|\mu_{1,1}^2 \ge \xi_0/2 + |H_1(\mathbf{m};\mathbf{q},\boldsymbol{\mu})|. \tag{III.3}$$

Therefore,

$$|\mathsf{F}_1(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})| \le \frac{\psi_1}{|\Im\big(\xi - q_2\mu_{1,*}^2 - H_1(\mathbf{m};\mathbf{q},\boldsymbol{\mu})\big)|}$$

$$\leq \frac{\psi_1}{\Im(\xi) - |H_1(\mathbf{m};\mathbf{q},\boldsymbol{\mu})|} \leq \frac{2\psi_1}{\xi_0},$$

where the last inequality follows from (III.3).

Similarly, for F_2 and F_3 we show that $|\mathsf{F}_2(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})| \leq 2\psi_2/\xi_0$ provided $\Im(\xi) \geq \xi_0 \geq 4 + 4|q_4|\mu_{2,1}^2$, and $|\mathsf{F}_3(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})| \leq 2\psi_3/\xi_0$ provided $\Im(\xi) \geq \xi_0 \geq 4 + 4|q_5|$. Therefore if ξ_0 satisfies $2\max\{\psi_1,\psi_2,\psi_3\}/\xi_0 \leq r_0$ and $\xi_0 \geq 4 + 4\max\{|q_4|\mu_{1,1}^2, |q_4|\mu_{2,1}^2, |q_5|\}$, $\mathbf{F}$ maps domain $\mathbb{D}(2\psi_1/\xi_0) \times \mathbb{D}(2\psi_2/\xi_0) \times \mathbb{D}(2\psi_3/\xi_0)$ into itself.

As for the Lipschitz continuity of $\mathbf{F}(\cdot;\mathbf{q},\boldsymbol{\mu})$, note that

$$\nabla_{\mathbf{m}}\mathsf{F}_1(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu}) = -\frac{\psi_1}{(-\xi + q_2\mu_{1,*}^2 + H_1(\mathbf{m};\mathbf{q},\boldsymbol{\mu}))^2} \cdot \nabla_{\mathbf{m}} H_1(\mathbf{m};\mathbf{q},\boldsymbol{\mu}).$$

With the same calculation as above, it is easy to see that when ξ_0 is sufficiently large, $\|\nabla_{\mathbf{m}} H_1(\mathbf{m};\mathbf{q},\boldsymbol{\mu})\|_2 \leq C(\mathbf{q},\boldsymbol{\mu})$ for all $\mathbf{m} \in \mathbb{D}(2\psi_1/\xi_0) \times \mathbb{D}(2\psi_2/\xi_0) \times \mathbb{D}(2\psi_3/\xi_0)$, where $C(\mathbf{q},\boldsymbol{\mu})$ is a constant that only depends on $\mathbf{q}$ and $\boldsymbol{\mu}$. Thus for such ξ_0 and ξ with $\Im(\xi) \geq \xi_0$,

$$\left\|\nabla_{\mathbf{m}}\mathsf{F}_1(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})\right\|_2 \leq \frac{C(\mathbf{q},\boldsymbol{\mu})\cdot\psi_1}{\Im(\xi) - |H_1(\mathbf{m};\mathbf{q},\boldsymbol{\mu})|} \leq \frac{4C(\mathbf{q},\boldsymbol{\mu})\cdot\psi_1}{\xi_0} \leq \frac{1}{4},$$

where we again utilize (III.3). We can apply the same argument for F_2 and F_3, and conclude that $\mathbf{F}$ is 1/2-Lipschitz on $\mathbf{m} \in \mathbb{D}(2\psi_1/\xi_0) \times \mathbb{D}(2\psi_2/\xi_0) \times \mathbb{D}(2\psi_3/\xi_0)$. Therefore by Banach fixed point theorem, there exists a unique fixed point of $\mathbf{F}$. Thus the solution of the implicit equations defined in Definition A.5 exists and is unique. ■

IV. Proof of Proposition A.6

The proof of Proposition A.6 is split into several sections. In Sections IV.1 and IV.2, we give some useful preliminary results. In Section IV.3, we show that $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$ is analytic on $\{\xi : \Im(\xi) \geq \xi_0\}$, and then prove the first and second conclusions of Proposition A.6. In Section IV.4, we prove the point convergence of $M_d(\xi;\mathbf{q},\boldsymbol{\mu})$ to $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$ under the additional assumption that $\sigma_j(x)$, $j = 1, 2$ are polynomials. In Section IV.5, we extend this point convergence result to general activation functions satisfying Assumption 3.2. In Section IV.6, we conclude the proof by showing the uniform convergence of $M_d(\xi;\mathbf{q},\boldsymbol{\mu})$ to $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$ on compact sets.

IV.1 Equivalence between Gaussian and spherical versions

The first step in the proof of Proposition A.6 is to relate the Stieltjes transform $M_d(\xi;\mathbf{q},\boldsymbol{\mu})$ to the Stieltjes transform corresponding to Gaussian data and Gaussian random features.

Definition IV.1. *Let $(\overline{\boldsymbol{\theta}}_a)_{a\in[N]}$ be i.i.d. standard Gaussian random vectors distributed as $\mathrm{N}(\mathbf{0},\mathbf{I}_d)$, and $\overline{\boldsymbol{\Theta}} \in \mathbb{R}^{N\times d}$ be the matrix whose a^{th} row is given by $\overline{\boldsymbol{\theta}}_a$. Similarly, we denote $(\overline{\mathbf{x}}_i)_{i\in[n]} \sim_{\text{iid}} \mathrm{N}(\mathbf{0},\mathbf{I}_d)$, and let $\overline{\mathbf{X}} \in \mathbb{R}^{n\times d}$ be the matrix whose a^{th} row is $\overline{\mathbf{x}}_i$.* □

Given these definitions, our original data inputs and random feature parameters which are distributed uniformly on the sphere $\sqrt{d}\cdot\mathbb{S}^{d-1}$ can be represented as

$$\mathbf{x}_i=\sqrt{d}\cdot\frac{\overline{\mathbf{x}}_i}{\|\overline{\mathbf{x}}_i\|_2}\sim\mathrm{Unif}(\sqrt{d}\cdot\mathbb{S}^{d-1}),\ \text{and}\ \boldsymbol{\theta}_a=\sqrt{d}\cdot\frac{\overline{\boldsymbol{\theta}}_a}{\|\overline{\boldsymbol{\theta}}_a\|_2}\sim\mathrm{Unif}(\sqrt{d}\cdot\mathbb{S}^{d-1}) \tag{IV.1}$$

for all $i\in[n]$ and $a\in[N]$. We can now consider the "Gaussian version" of the learning problem, where the data inputs are $(\overline{\mathbf{x}}_i)_{i\in[n]}$, and the double random feature model uses random parameters $(\overline{\boldsymbol{\theta}}_a)_{a\in[N]}$ and activation functions

$$\phi_j(x)=\sigma_j(x)-\mathbb{E}_{G\sim\mathrm{N}(0,1)}[\sigma_j(G)],\quad j=1,2. \tag{IV.2}$$

For this version of the learning problem, we can similarly construct the linear pencil matrix $\overline{\mathbf{A}}(\mathbf{q},\boldsymbol{\mu})$, which is the counterpart of the linear pencil matrix $\mathbf{A}(\mathbf{q},\boldsymbol{\mu})$ defined in Definition A.3.

Definition IV.2. *The linear pencil matrix* $\overline{\mathbf{A}}(\mathbf{q},\boldsymbol{\mu})\in\mathbb{R}^{P\times P}$ *($P=N+n$) is defined as*

$$\overline{\mathbf{A}}(\mathbf{q},\boldsymbol{\mu})=\begin{bmatrix}q_2\mu_{1,*}^2\mathbf{I}_{N_1}+q_4\mu_{1,1}^2\frac{\overline{\boldsymbol{\Theta}}_1\overline{\boldsymbol{\Theta}}_1^\top}{d} & q_4\mu_{1,1}\mu_{2,1}\frac{\overline{\boldsymbol{\Theta}}_1\overline{\boldsymbol{\Theta}}_2^\top}{d} & \mathbf{J}_1^\top+q_1\widetilde{\mathbf{J}}_1^\top\\ q_4\mu_{1,1}\mu_{2,1}\frac{\overline{\boldsymbol{\Theta}}_2\overline{\boldsymbol{\Theta}}_1^\top}{d} & q_2\mu_{2,*}^2\mathbf{I}_{N_2}+q_4\mu_{2,1}^2\frac{\overline{\boldsymbol{\Theta}}_2\overline{\boldsymbol{\Theta}}_2^\top}{d} & \mathbf{J}_2^\top+q_1\widetilde{\mathbf{J}}_2^\top\\ \mathbf{J}_1+q_1\widetilde{\mathbf{J}}_1 & \mathbf{J}_2+q_1\widetilde{\mathbf{J}}_2 & q_3\mathbf{I}_n+q_5\frac{\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top}{d}\end{bmatrix},$$

where $\mathbf{J}_j=\phi_j\big(\overline{\mathbf{X}}\,\overline{\boldsymbol{\Theta}}_j^\top/\sqrt{d}\big)/\sqrt{d}$, $\widetilde{\mathbf{J}}_j=\frac{\mu_{j,1}}{d}\overline{\mathbf{X}}\,\overline{\boldsymbol{\Theta}}_j^\top$, $j=1,2$. □

We also define $\overline{M}_d(\xi;\mathbf{q},\boldsymbol{\mu})=\frac{1}{d}\mathrm{tr}\big[(\overline{\mathbf{A}}(\mathbf{q},\boldsymbol{\mu})-\xi\mathbf{I}_P)^{-1}\big]$ as the counterpart of the Stieltjes transform $M_d(\xi;\mathbf{q},\boldsymbol{\mu})$. The following lemma establishes the equivalence between the two versions of the Stieltjes transforms.

Lemma IV.3. *Suppose that* $\sigma_j(x)$, $j=1,2$, *are polynomials. Then for any fixed* $\mathbf{q}$ *and* $\xi\in\mathbb{C}_+$, *we have*

$$\mathbb{E}\big|\overline{M}_d(\xi;\mathbf{q},\boldsymbol{\mu})-M_d(\xi;\mathbf{q},\boldsymbol{\mu})\big|=o_d(1).$$

Proof [Proof of Lemma IV.3] Define

$$\boldsymbol{\Delta}(\mathbf{A},\overline{\mathbf{A}},\xi)=M_d(\xi;\mathbf{q},\boldsymbol{\mu})-\overline{M}_d(\xi;\mathbf{q},\boldsymbol{\mu}),$$

and write $M_d(\xi;\mathbf{q},\boldsymbol{\mu})$ and $\overline{M}_d(\xi;\mathbf{q},\boldsymbol{\mu})$ as $M_d(\xi)$ and $\overline{M}_d(\xi)$ to simplify the notation. Then by definition we have

$$\begin{aligned}|\boldsymbol{\Delta}(\mathbf{A},\overline{\mathbf{A}},\xi)|=&\big|\mathrm{tr}\big[(\mathbf{A}-\xi\mathbf{I})^{-1}(\mathbf{A}-\overline{\mathbf{A}})(\overline{\mathbf{A}}-\xi\mathbf{I})^{-1}\big]\big|/d\\ \le&\|(\mathbf{A}-\xi\mathbf{I})^{-1}(\overline{\mathbf{A}}-\xi\mathbf{I})^{-1}\|_{\mathrm{op}}\|\mathbf{A}-\overline{\mathbf{A}}\|_\star/d\\ \le&\|\mathbf{A}-\overline{\mathbf{A}}\|_\star\cdot\frac{1}{d}\cdot\frac{1}{(\Im(\xi))^2},\end{aligned} \tag{IV.3}$$

where $\|\cdot\|_\star$ is the nuclear norm, the first inequality follows from the fact that $\mathrm{tr}(\mathbf{UV})\le\|\mathbf{U}\|_{\mathrm{op}}\|\mathbf{V}\|_\star$ for all $\mathbf{U}\in\mathbb{C}^{N\times N}$ and Hermite $\mathbf{V}\in\mathbb{C}^{N\times N}$, and the second inequality follows from the fact that $\mathbf{A}$ and $\overline{\mathbf{A}}$ are real matrices. Because

$$|M_d(\xi)|=\frac{1}{d}\Big|\mathrm{tr}\big(\mathbf{A}-\xi\mathbf{I}\big)^{-1}\Big|\le\frac{P}{d}\big\|\mathbf{A}-\xi\mathbf{I}\big\|_{\mathrm{op}}\le P/(d\cdot\Im(\xi)),$$

$$|\overline{M}_d(\xi)| = \frac{1}{d}\left|\mathrm{tr}\left(\overline{\mathbf{A}} - \xi\mathbf{I}\right)^{-1}\right| \leq \frac{P}{d}\left\|\overline{\mathbf{A}} - \xi\mathbf{I}\right\|_{\mathrm{op}} \leq P/(d\cdot\Im(\xi)),$$

$|\boldsymbol{\Delta}(\mathbf{A}, \overline{\mathbf{A}}, \xi)|$ is deterministically upper bounded:

$$|\boldsymbol{\Delta}(\mathbf{A}, \overline{\mathbf{A}}, \xi)| \leq |M_d(\xi)| + |\overline{M}_d(\xi)| \leq 2P/(d\cdot\Im(\xi)). \tag{IV.4}$$

Therefore, if we can prove $\|\mathbf{A} - \overline{\mathbf{A}}\|_\star/d = o_{\mathbb{P}}(1)$, then according to (IV.3) and (IV.4), we can conclude that $\mathbb{E}|\boldsymbol{\Delta}(\mathbf{A}, \overline{\mathbf{A}}, \xi)| = o_d(1)$ by the dominated convergence theorem. To this end, we first recall the notations in Definitions A.1 and A.3 that for $j = 1, 2$,

$$\mathbf{Z}_j = \sigma_j\left(\mathbf{X}\boldsymbol{\Theta}_j^\top/\sqrt{d}\right)/\sqrt{d} \in \mathbb{R}^{n\times N_j}, \qquad \widetilde{\mathbf{Z}}_j = \frac{\mu_{j,1}}{d}\mathbf{X}\boldsymbol{\Theta}_j^\top.$$

We also remind readers that $\mathbf{J}_j = \phi_j\big(\overline{\mathbf{X}}\ \overline{\boldsymbol{\Theta}}_j^\top/\sqrt{d}\big)/\sqrt{d}$, $\widetilde{\mathbf{J}}_j = \frac{\mu_{j,1}}{d}\overline{\mathbf{X}}\ \overline{\boldsymbol{\Theta}}_j^\top$ are the "Gaussian version" counterparts of $\mathbf{Z}_j$ and $\widetilde{\mathbf{Z}}_j$ respectively. We further denote $\mathbf{Z}_{j,0} = \mu_{j,0}\mathbf{1}_n\mathbf{1}_{N_j}/\sqrt{d}$ and let $\mathbf{Z}_{j,\star} = \mathbf{Z}_j - \mathbf{Z}_{j,0}$ for $j = 1, 2$. Then by the definition of the functions ϕ_1, ϕ_2, clearly we have $\mathbf{Z}_{j,\star} = \phi_j(\mathbf{X}\boldsymbol{\Theta}^\top/\sqrt{d})/\sqrt{d}$ for $j = 1, 2$. With these notations, we can rewrite $\mathbf{A} - \overline{\mathbf{A}}$ as follows:

$$\begin{aligned}
\mathbf{A} - \overline{\mathbf{A}} =& q_5\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \frac{\mathbf{X}\mathbf{X}^\top - \overline{\mathbf{X}}\ \overline{\mathbf{X}}^\top}{d}\end{bmatrix} + q_4\begin{bmatrix}\frac{\mathbf{M}_1\boldsymbol{\Theta}\boldsymbol{\Theta}^\top\mathbf{M}_1 - \mathbf{M}_1\overline{\boldsymbol{\Theta}}\ \overline{\boldsymbol{\Theta}}^\top\mathbf{M}_1}{d} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\\
&+ q_1\begin{bmatrix}\mathbf{0} & [\widetilde{\mathbf{Z}}_1, \widetilde{\mathbf{Z}}_2]^\top - [\widetilde{\mathbf{J}}_1, \widetilde{\mathbf{J}}_2]^\top\\ [\widetilde{\mathbf{Z}}_1, \widetilde{\mathbf{Z}}_2]^\top - [\widetilde{\mathbf{J}}_1, \widetilde{\mathbf{J}}_2] & \mathbf{0}\end{bmatrix} + \begin{bmatrix}\mathbf{0} & [\mathbf{Z}_{1,0}, \mathbf{Z}_{2,0}]^\top\\ [\mathbf{Z}_{1,0}, \mathbf{Z}_{2,0}] & \mathbf{0}\end{bmatrix}\\
&+ \begin{bmatrix}\mathbf{0} & [\mathbf{Z}_{1,\star}, \mathbf{Z}_{2,\star}]^\top - [\mathbf{J}_1, \mathbf{J}_2]^\top\\ [\mathbf{Z}_{1,\star}, \mathbf{Z}_{2,\star}] - [\mathbf{J}_1, \mathbf{J}_2] & \mathbf{0}\end{bmatrix}.
\end{aligned}$$

Then by the triangle inequality and Cauchy-Schwarz inequality, we have

$$\frac{\|\mathbf{A} - \overline{\mathbf{A}}\|_\star}{d} = O_{\mathbb{P}}(I_1 + I_2 + I_3 + I_4 + I_5),$$

where

$$\begin{aligned}
I_1 &= \frac{1}{\sqrt{d}}\left\|\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \frac{\mathbf{X}\mathbf{X}^\top - \overline{\mathbf{X}}\ \overline{\mathbf{X}}^\top}{d}\end{bmatrix}\right\|_F,\\
I_2 &= \frac{1}{\sqrt{d}}\left\|\begin{bmatrix}\frac{\mathbf{M}_1\boldsymbol{\Theta}\boldsymbol{\Theta}^\top\mathbf{M}_1 - \mathbf{M}_1\overline{\boldsymbol{\Theta}}\ \overline{\boldsymbol{\Theta}}^\top\mathbf{M}_1}{d} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right\|_F,\\
I_3 &= \frac{1}{\sqrt{d}}\left\|\begin{bmatrix}\mathbf{0} & [\widetilde{\mathbf{Z}}_1, \widetilde{\mathbf{Z}}_2]^\top - [\widetilde{\mathbf{J}}_1, \widetilde{\mathbf{J}}_2]^\top\\ [\widetilde{\mathbf{Z}}_1, \widetilde{\mathbf{Z}}_2]^\top - [\widetilde{\mathbf{J}}_1, \widetilde{\mathbf{J}}_2] & \mathbf{0}\end{bmatrix}\right\|_F,\\
I_4 &= \frac{1}{d}\left\|\begin{bmatrix}\mathbf{0} & [\mathbf{Z}_{1,0}, \mathbf{Z}_{2,0}]^\top\\ [\mathbf{Z}_{1,0}, \mathbf{Z}_{2,0}] & \mathbf{0}\end{bmatrix}\right\|_\star,\\
I_5 &= \frac{1}{\sqrt{d}}\left\|\begin{bmatrix}\mathbf{0} & [\mathbf{Z}_{1,\star}, \mathbf{Z}_{2,\star}]^\top - [\mathbf{J}_1, \mathbf{J}_2]^\top\\ [\mathbf{Z}_{1,\star}, \mathbf{Z}_{2,\star}] - [\mathbf{J}_1, \mathbf{J}_2] & \mathbf{0}\end{bmatrix}\right\|_F.
\end{aligned}$$

In the following, we bound the terms $I_1, \ldots, I_5$ separately. For I_1, let $\mathbf{D}_{\mathbf{x}} = \mathrm{diag}(\sqrt{d}/\|\overline{\mathbf{x}}_1\|_2, ..., \sqrt{d}/\|\overline{\mathbf{x}}_n\|_2)$. Then we have $\mathbf{X} = \mathbf{D}_{\mathbf{x}}\overline{\mathbf{X}}$ by (IV.1), and

$$\begin{aligned}
I_1 =& \frac{1}{\sqrt{d}}\Big\|\frac{\mathbf{X}\mathbf{X}^\top - \overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top}{d}\Big\|_F \le \Big\|\frac{\mathbf{X}\mathbf{X}^\top - \overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top}{d}\Big\|_{\mathrm{op}} = \Big\|\frac{\mathbf{D}_{\mathbf{x}}\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top\mathbf{D}_{\mathbf{x}} - \overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top}{d}\Big\|_{\mathrm{op}} \\
=& \Big\|\frac{(\mathbf{D}_{\mathbf{x}} - \mathbf{I}_n)\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top(\mathbf{D}_{\mathbf{x}} + \mathbf{I}_n) + \overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top\mathbf{D}_{\mathbf{x}} - \mathbf{D}_{\mathbf{x}}\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top}{d}\Big\|_{\mathrm{op}} \\
\le& \|\mathbf{D}_{\mathbf{x}} - \mathbf{I}_n\|_{\mathrm{op}} \cdot \Big\|\frac{\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top}{d}\Big\|_{\mathrm{op}} \cdot \big(1 + \|\mathbf{D}_{\mathbf{x}}\|_{\mathrm{op}}\big) + \Big\|\frac{\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top\mathbf{D}_{\mathbf{x}} - \mathbf{D}_{\mathbf{x}}\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top}{d}\Big\|_{\mathrm{op}} \\
=& \Big\|\frac{\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top\mathbf{D}_{\mathbf{x}} - \mathbf{D}_{\mathbf{x}}\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top}{d}\Big\|_{\mathrm{op}} + o_{\mathbb{P}}(1),
\end{aligned} \tag{IV.5}$$

where the first inequality holds since the average of the d squared eigenvalues of $(\mathbf{X}\mathbf{X}^\top - \overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top)/d$ is bounded by the largest one of them, and the last equality follows from $\|\mathbf{D}_{\mathbf{x}} - \mathbf{I}_n\|_{\mathrm{op}} = O_{\mathbb{P}}\Big(\sqrt{\frac{\log d}{d}}\Big)$ and $\|\mathbf{D}_{\mathbf{x}}\|_{\mathrm{op}} = O_{\mathbb{P}}(1)$, which are direct consequences of the definition of $\mathbf{D}_{\mathbf{x}}$. We further let $\widetilde{\mathbf{D}}_{\mathbf{x}}$ be the matrix whose elements $\big(\widetilde{\mathbf{D}}_{\mathbf{x}}\big)_{ij}$ satisfy $\big(\widetilde{\mathbf{D}}_{\mathbf{x}}\big)_{ij} = \big(\mathbf{D}_{\mathbf{x}}\big)_{jj} - \big(\mathbf{D}_{\mathbf{x}}\big)_{ii}$ for $i, j \in [n]$. Then we have $\|\widetilde{\mathbf{D}}_{\mathbf{x}}\|_{\max} = o_{\mathbb{P}}(1)$, and

$$\Big\|\frac{\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top\mathbf{D}_{\mathbf{x}} - \mathbf{D}_{\mathbf{x}}\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top}{d}\Big\|_{\mathrm{op}} = \Big\|\widetilde{\mathbf{D}}_{\mathbf{x}} \odot \frac{\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top}{d}\Big\|_{\mathrm{op}} \le \|\widetilde{\mathbf{D}}_{\mathbf{x}}\|_{\max} \cdot \Big\|\frac{\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top}{d}\Big\|_{\mathrm{op}} = o_{\mathbb{P}}(1). \tag{IV.6}$$

Plugging (IV.6) into (IV.5) completes the proof of $I_1 = o_{\mathbb{P}}(1)$. Similarly, it can be shown that I_2 and I_3 are both $o_{\mathbb{P}}(1)$. For I_4, by the definition that $\mathbf{Z}_{j,0} = \mu_{j,0}\mathbf{1}_n\mathbf{1}_{N_j}/\sqrt{d}$, $j = 1, 2$, it is clear that $\mathbf{Z}_{j,0}$ is rank-one and $\|\mathbf{Z}_{j,0}\|_{\mathrm{op}} = O_d(\sqrt{d})$. Therefore we have

$$I_4 = \frac{1}{d}\left\|\begin{bmatrix}\mathbf{0} & [\mathbf{Z}_{1,0}, \mathbf{Z}_{2,0}]^\top \\ [\mathbf{Z}_{1,0}, \mathbf{Z}_{2,0}] & \mathbf{0}\end{bmatrix}\right\|_\star = o_d(1).$$

Finally, to prove $I_5 = o_d(1)$, it clearly suffices to show that

$$\frac{1}{\sqrt{d}}\|\mathbf{Z}_{j,\star} - \mathbf{J}_j\|_F = o_{\mathbb{P}}(1), \quad j = 1, 2.$$

Define $\overline{\mathbf{Z}}_{j,\star} = \phi_j(\mathbf{X}\,\overline{\boldsymbol{\Theta}}_j^\top/\sqrt{d})/\sqrt{d} = \phi_j(\mathbf{D}_{\mathbf{x}}\overline{\mathbf{X}}\,\overline{\boldsymbol{\Theta}}_j^\top/\sqrt{d})/\sqrt{d}$, $j \in [N]$ and $r_i = \sqrt{d}/\|\overline{\mathbf{x}}_i\|_2$, $i \in [n]$. By the mean value theorem, for $j = 1, 2$, $a \in [N_j]$ and $i \in [n]$, there exists ζ_{ia} between r_i and 1, such that

$$\begin{aligned}
\overline{\mathbf{Z}}_{j,\star} - \mathbf{J}_j =& \big[\phi_j(r_i\langle\overline{\mathbf{x}}_i, \overline{\boldsymbol{\theta}}_a\rangle/\sqrt{d})/\sqrt{d} - \phi_j(\langle\overline{\mathbf{x}}_i, \overline{\boldsymbol{\theta}}_a\rangle/\sqrt{d})/\sqrt{d}\big]_{i\in[n], a\in[N_j]} \\
=& \big[(r_i - 1)(\langle\overline{\mathbf{x}}_i, \overline{\boldsymbol{\theta}}_a\rangle/\sqrt{d})\phi_j'(\zeta_{ij}\langle\overline{\mathbf{x}}_i, \overline{\boldsymbol{\theta}}_a\rangle/\sqrt{d})/\sqrt{d}\big]_{i\in[n], a\in[N_j]} \\
=& (\mathbf{D}_{\mathbf{x}} - \mathbf{I}_n)\overline{\phi}_j(\boldsymbol{\zeta} \odot (\overline{\mathbf{X}}\,\overline{\boldsymbol{\Theta}}_j^\top/\sqrt{d}))/\sqrt{d},
\end{aligned}$$

where $\boldsymbol{\zeta} = (\zeta_{ij})_{i\in[n], a\in[N_j]}$ and $\overline{\phi}_j(x) = x\phi_j'(x)$. By Bernstein-type concentration inequalities (Vershynin, 2010), we have

$$\|\mathbf{D}_{\mathbf{x}} - \mathbf{I}_n\|_{\mathrm{op}} = O_{\mathbb{P}}\Big(\sqrt{\frac{\log d}{d}}\Big), \ \|\boldsymbol{\zeta}\|_{\max} = O_{\mathbb{P}}(1), \ \|\overline{\mathbf{X}}\,\overline{\boldsymbol{\Theta}}_j^\top/\sqrt{d}\|_{\max} = O_{\mathbb{P}}(\sqrt{\log d}).$$

Moreover, note that we currently assume that the activation functions σ_j, $j = 1, 2$ are fixed polynomials, which implies that ϕ_j are also fixed polynomials. Therefore, there exists a constant $M_0 \in \mathbb{N}$ such that

$$||\overline{\mathbf{Z}}_{j,\star} - \mathbf{J}_j||_F/\sqrt{d} \leq \|\mathbf{D}_{\mathbf{x}} - \mathbf{I}_n\|_{\text{op}}||\overline{\phi}_j(\boldsymbol{\zeta} \odot (\overline{\mathbf{X}}\ \overline{\boldsymbol{\Theta}}_j^\top/\sqrt{d}))||_F/d = O_{\mathbb{P}}((\log d)^{M_0}/\sqrt{d}) = o_{\mathbb{P}}(1).$$

With exactly the same argument, it can be shown that $||\overline{\mathbf{Z}}_{j,\star} - \mathbf{Z}_{j,\star}||_F/\sqrt{d} = o_{\mathbb{P}}(1)$ (recall $\mathbf{Z}_{j,\star} = \phi_j(\mathbf{X}\boldsymbol{\Theta}^\top/\sqrt{d})/\sqrt{d}$). Therefore we have

$$\frac{1}{\sqrt{d}}\|\mathbf{Z}_{j,\star} - \mathbf{J}_j\|_F \leq \frac{1}{\sqrt{d}}\|\overline{\mathbf{Z}}_{j,\star} - \mathbf{Z}_{j,\star}\|_F + \frac{1}{\sqrt{d}}\|\overline{\mathbf{Z}}_{j,\star} - \mathbf{J}_j\|_F = o_{\mathbb{P}}(1)$$

for $j = 1, 2$. Finally $I_5 = o_p(1)$ and the proof of Lemma IV.3 is complete. ■

IV.2 Calculation of the resolvent equations

Lemma IV.3 and its proof show the readers that the Stieltjes transforms of the empirical eigenvalue distributions of $\mathbf{A}$ and $\overline{\mathbf{A}}$ share the same asymptotics. Based on this result, we can equivalently consider the "Gaussian version" counterpart of the learning problem. Therefore, throughout Section IV.2, we directly consider the matrices $\overline{\mathbf{X}}$ and $\overline{\boldsymbol{\Theta}}$, whose elements are independently generated from standard normal $\mathrm{N}(0, 1)$. In addition, the activation functions for the two types of random features are $\phi_j(x) = \sigma_j(x) - \mu_{j,0}$, $j = 1, 2$, and the linear pencil matrix is $\overline{\mathbf{A}}(\mathbf{q}, \boldsymbol{\mu})$ is given in Definition IV.2.

Moreover, for $j = 1, 2$, let $\Phi_j(x) = \phi_j(x) + q_1\mu_{j,1}x$, and it is easy to see $\mathbf{J}_j^\top + q_1\widetilde{\mathbf{J}}_j^\top = \Phi_j\left(\overline{\mathbf{X}}\ \overline{\boldsymbol{\Theta}}_j^\top/\sqrt{d}\right)/\sqrt{d}$. We further denote $\phi_{j,0} \triangleq \mathbb{E}_{G\sim\mathrm{N}(0,1)}\{\Phi_j(G)\}$, $\phi_{j,1} \triangleq \mathbb{E}_{G\sim\mathrm{N}(0,1)}\{G\Phi_j(G)\}$, $\phi_{j,*} \triangleq \mathbb{E}_{G\sim\mathrm{N}(0,1)}\ \{\Phi_j(G)^2\} - \phi_{j,0}^2 - \phi_{j,1}^2$. By these definitions, it is easy to see that $\phi_{j,0} = 0$, $\phi_{j,1}^2 = \mu_{j,1}^2(1 + q_1)^2$, $\phi_{j,*}^2 = \mu_{j,*}^2$. Importantly, the property that $\mathbb{E}_{G\sim\mathrm{N}(0,1)}\{\Phi_j(G)\} = \phi_{j,0} = 0$ enables the application of the following lemma, which is summarized from Section 4.3, Step 2 in Cheng and Singer (2013).

Lemma IV.4. *Suppose that Φ is a polynomial satisfying $\mathbb{E}_{G\sim\mathrm{N}(0,1)}\{\Phi(G)\} = 0$, $\mathbb{E}_{G\sim\mathrm{N}(0,1)}\{G\Phi(G)\} = 0$ and $\overline{\mathbf{x}}_i, \overline{\boldsymbol{\theta}}_a \in \mathbb{R}^d$, $i \in [n]$, $a \in [N]$ are standard Gaussian vectors. Define matrix $\mathbf{E} \in \mathbb{R}^{n\times N}$ elementwisely as*

$$(\mathbf{E}_j)_{i,a} = \frac{1}{\sqrt{d}}\left[\Phi\left(\frac{1}{\sqrt{d}}\langle\overline{\mathbf{x}}_i, \overline{\boldsymbol{\theta}}_a\rangle\right) - \Phi\left(\frac{1}{\sqrt{d}}\langle(\overline{\mathbf{x}}_i)_{[1:d-1]}, (\overline{\boldsymbol{\theta}}_a)_{[1:d-1]}\rangle\right)\right]$$

for $i \in [n]$, $a \in [N]$. Then $\|\mathbf{E}\|_{\text{op}} = o_{\mathbb{P}}(1)$.

Lemma IV.4 formally shows the intuitive result that under the setting where d, n grows proportionally, removing one entry in the random vectors does not change the asymptotic limit of polynomials. This enables us to apply the standard leave-one-out argument in random matrix theory.

Our goal in this part of the proof is to calculate the resolvent equations of the Stieltjes transforms corresponding to the pencil matrix $\overline{\mathbf{A}}(\mathbf{q}, \boldsymbol{\mu})$. To do so, we define the following terms:

$$\overline{m}_{1,d}(\xi; \mathbf{q}, \boldsymbol{\mu}) = \mathbb{E}\big[\overline{M}_{1,d}(\xi; \mathbf{q}, \boldsymbol{\mu})\big], \quad \overline{M}_{1,d}(\xi; \mathbf{q}, \boldsymbol{\mu}) = \frac{1}{d}\mathrm{tr}_{[1:N_1]}\big[(\overline{\mathbf{A}}(\mathbf{q}, \boldsymbol{\mu}) - \xi \mathbf{I}_P)^{-1}\big],$$
$$\overline{m}_{2,d}(\xi; \mathbf{q}, \boldsymbol{\mu}) = \mathbb{E}\big[\overline{M}_{2,d}(\xi; \mathbf{q}, \boldsymbol{\mu})\big], \quad \overline{M}_{2,d}(\xi; \mathbf{q}, \boldsymbol{\mu}) = \frac{1}{d}\mathrm{tr}_{[N_1+1:N]}\big[(\overline{\mathbf{A}}(\mathbf{q}, \boldsymbol{\mu}) - \xi \mathbf{I}_P)^{-1}\big],$$
$$\overline{m}_{3,d}(\xi; \mathbf{q}, \boldsymbol{\mu}) = \mathbb{E}\big[\overline{M}_{3,d}(\xi; \mathbf{q}, \boldsymbol{\mu})\big], \quad \overline{M}_{3,d}(\xi; \mathbf{q}, \boldsymbol{\mu}) = \frac{1}{d}\mathrm{tr}_{[N+1:P]}\big[(\overline{\mathbf{A}}(\mathbf{q}, \boldsymbol{\mu}) - \xi \mathbf{I}_P)^{-1}\big].$$

Standard argument in random matrix theory then gives us the concentration result

$$\mathbb{E}\big|\overline{M}_{i,d}(\xi; \mathbf{q}, \boldsymbol{\mu}) - \overline{m}_{i,d}(\xi; \mathbf{q}, \boldsymbol{\mu})\big| = o_d(1)$$

for any fixed $\xi \in \mathbb{C}_+$. Therefore, denoting $\overline{m}_d(\xi) = \sum_{i=1}^{3} \overline{m}_{i,d}(\xi)$ and $\overline{M}_d(\xi) = \sum_{i=1}^{3} \overline{M}_{i,d}(\xi)$, (we drop the argument $\mathbf{q}, \boldsymbol{\mu}$ for simplicity), we have

$$\mathbb{E}\big|\overline{M}_d(\xi) - \overline{m}_d(\xi)\big| = o_d(1) \tag{IV.7}$$

for any fixed $\xi \in \mathbb{C}_+$. A proof of this concentration can be found in Hastie et al. (2022); Mei and Montanari (2022). Based on (IV.7), to study $\overline{M}_d(\xi; \mathbf{q}, \boldsymbol{\mu})$, which is the Stieltjes transform of the empirical eigenvalue distribution of $\overline{A}(\mathbf{q}, \boldsymbol{\mu})$, it suffices to derive the resolvent equations for $\overline{m}_d(\xi; \mathbf{q}, \boldsymbol{\mu})$. This is done in the following lemma.

Lemma IV.5. *Let $\overline{\mathbf{m}}_d(\xi) = [\overline{m}_{1,d}(\xi), \overline{m}_{2,d}(\xi), \overline{m}_{3,d}(\xi)]^\top$. Then for any fixed $\xi \in \mathbb{C}_+$, the following property holds:*

$$\|\overline{\mathbf{m}}_d(\xi) - \mathbf{F}(\overline{\mathbf{m}}_d(\xi))\|_2 = o_d(1).$$

Proof [Proof of Lemma IV.5] Since $\overline{\mathbf{m}}_d(\xi), \mathbf{F}(\overline{\mathbf{m}}_d(\xi)) \in \mathbb{C}^3$, Lemma IV.5 essentially contains three results showing that the first, second, and third elements of $\overline{\mathbf{m}}_d(\xi) - \mathbf{F}(\overline{\mathbf{m}}_d(\xi))$ are all asymptotically zero. Since the proofs of the three results are almost the same, we mainly focus on the proof of the first result. The proof consists of three steps. The first step is to use the Schur complement formula to calculate $\overline{m}_{1,d}$. The second step is to simplify the formula of $\overline{m}_{1,d}$. The third step is to give the recursive equations of $\overline{m}_{1,d}$ based on the result of step 2.

Step 1. We first use a leave-one-out argument to calculate $\overline{m}_{1,d}$. Let $\overline{\mathbf{A}}_{\cdot,N_1} \in \mathbb{R}^{P-1}$ be the N_1^{th} column of $\overline{\mathbf{A}}$, with the N_1^{th} entry removed. We further denote by $\overline{\mathbf{B}} \in \mathbb{R}^{(P-1)\times(P-1)}$ the sub-matrix of $\overline{\mathbf{A}}$ obtained by removing the N_1^{th} row and N_1^{th} column in $\overline{\mathbf{A}}$. We can then treat $\overline{\mathbf{A}}$ as a 2×2 block matrix formed by $\overline{\mathbf{A}}_{\cdot,N_1}$, $\overline{\mathbf{A}}_{\cdot,N_1}^\top$, $\overline{\mathbf{B}}$, and $\overline{\mathbf{A}}_{N_1,N_1} = q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2\|\overline{\boldsymbol{\theta}}_{N_1}\|_2^2/d$. Then by the Schur complement formula, we get

$$\overline{m}_{1,d} = \psi_1 \mathbb{E}\left(-\xi + q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2\|\overline{\boldsymbol{\theta}}_{N_1}\|_2^2/d - \overline{\mathbf{A}}_{\cdot,N_1}^\top(\overline{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\overline{\mathbf{A}}_{\cdot,N_1}\right)^{-1}. \tag{IV.8}$$

We decompose the vectors $\overline{\boldsymbol{\theta}}_a$, $a \in [N]$ and $\overline{\mathbf{x}}_i$, $i \in [n]$ into components along the direction of $\overline{\boldsymbol{\theta}}_{N_1}$ and other orthogonal directions:

$$\begin{aligned}
\overline{\boldsymbol{\theta}}_a &= \eta_a \frac{\overline{\boldsymbol{\theta}}_{N_1}}{\|\overline{\boldsymbol{\theta}}_{N_1}\|} + \widetilde{\boldsymbol{\theta}}_a,\ \langle \overline{\boldsymbol{\theta}}_{N_1}, \widetilde{\boldsymbol{\theta}}_a \rangle = 0,\ a \in [N]\backslash\{N_1\}, \\
\overline{\mathbf{x}}_i &= u_i \frac{\overline{\boldsymbol{\theta}}_{N_1}}{\|\overline{\boldsymbol{\theta}}_{N_1}\|} + \widetilde{\mathbf{x}}_i,\ \langle \overline{\boldsymbol{\theta}}_{N_1}, \widetilde{\mathbf{x}}_i \rangle = 0,\ i \in [n].
\end{aligned} \tag{IV.9}$$

Note that for any $a \in [N]\backslash\{N_1\}$ and $i \in [n]$, η_a, u_i are standard Gaussian and are independent of $\widetilde{\boldsymbol{\theta}}_a$ and $\widetilde{\mathbf{x}}_i$. Moreover, $\widetilde{\boldsymbol{\theta}}_a$ and $\widetilde{\mathbf{x}}_i$ are conditionally independent on each other given $\overline{\boldsymbol{\theta}}_{N_1}$, with $\widetilde{\boldsymbol{\theta}}_a, \widetilde{\mathbf{x}}_i \sim N(0, P_\perp)$, where $P_\perp$ is the projector orthogonal to $\overline{\boldsymbol{\theta}}_{N_1}$. We can then use the coefficients η_a, $a \in [N]\backslash\{N_1\}$ and u_i, $i \in [n]$ to represent the entries of $\overline{\mathbf{A}}_{\cdot,N_1}$. We have $\overline{\mathbf{A}}_{\cdot,N_1} = [\overline{\mathbf{A}}_{1,N_1}, ..., \overline{\mathbf{A}}_{P-1,N_1}]^\top \in \mathbb{R}^{P-1}$ with

$$\overline{\mathbf{A}}_{i,N_1} = \begin{cases} \dfrac{q_4 \mu_{1,1}^2 \eta_i}{d} \|\overline{\boldsymbol{\theta}}_{N_1}\|_2, & \text{if } i \in [1, N_1 - 1], \\ \dfrac{q_4 \mu_{1,1} \mu_{2,1} \eta_{i+1}}{d} \|\overline{\boldsymbol{\theta}}_{N_1}\|_2, & \text{if } i \in [N_1, N - 1], \\ \dfrac{1}{\sqrt{d}} \Phi_1\Big(\dfrac{1}{\sqrt{d}} u_{i-N+1} \|\overline{\boldsymbol{\theta}}_{N_1}\|_2\Big), & \text{if } i \geq N. \end{cases} \tag{IV.10}$$

To calculate the resolvent equations, we need to further represent the matrix $\overline{\mathbf{B}}$ in (IV.8) with η_a, $\widetilde{\boldsymbol{\theta}}_a$, u_i, and $\widetilde{\mathbf{x}}_i$ for $a \in [N]\backslash\{N_1\}$ and $i \in [n]$. Below we first list some additional notations for easier reference. We write $\boldsymbol{\eta}_1 = [\eta_1, ..., \eta_{N_1-1}] \in \mathbb{R}^{N_1-1}$, $\boldsymbol{\eta}_2 = [\eta_{N_1+1}, ..., \eta_N] \in \mathbb{R}^{N_2}$, $\boldsymbol{\eta} = [\boldsymbol{\eta}_1^\top, \boldsymbol{\eta}_2^\top]^\top \in \mathbb{R}^{N-1}$, $\mathbf{u} = [u_1, ..., u_n]^\top \in \mathbb{R}^n$, $\widetilde{\boldsymbol{\Theta}}_1 = [\widetilde{\boldsymbol{\theta}}_1, ..., \widetilde{\boldsymbol{\theta}}_{N_1-1}]^\top$, $\widetilde{\boldsymbol{\Theta}}_2 = [\widetilde{\boldsymbol{\theta}}_{N_1+1}, ..., \widetilde{\boldsymbol{\theta}}_N]^\top$, $\widetilde{\boldsymbol{\Theta}} = \begin{bmatrix} \widetilde{\boldsymbol{\Theta}}_1 \\ \widetilde{\boldsymbol{\Theta}}_2 \end{bmatrix} \in \mathbb{R}^{(N-1)\times d}$, $\widetilde{\mathbf{M}}_1 = \begin{bmatrix} \mu_{1,1}\mathbf{I}_{N_1-1} & \\ & \mu_{2,1}\mathbf{I}_{N_2} \end{bmatrix}$ and $\widetilde{\mathbf{M}}_* = \begin{bmatrix} \mu_{1,*}\mathbf{I}_{N_1-1} & \\ & \mu_{2,*}\mathbf{I}_{N_2} \end{bmatrix}$. Now with (IV.9) and the notations above, we can decompose $\overline{\mathbf{B}}_{[1:N-1],[1:N-1]}$ as follows:

$$\overline{\mathbf{B}}_{[1:N-1],[1:N-1]} = q_2 \widetilde{\mathbf{M}}_* \widetilde{\mathbf{M}}_* + \frac{q_4}{d} \widetilde{\mathbf{M}}_1 \widetilde{\boldsymbol{\Theta}} \widetilde{\boldsymbol{\Theta}}^\top \widetilde{\mathbf{M}}_1 + \frac{q_4}{d} \widetilde{\mathbf{M}}_1 \boldsymbol{\eta}\boldsymbol{\eta}^\top \widetilde{\mathbf{M}}_1. \tag{IV.11}$$

Moreover, for $i, j \in [n]$ we define

$$\big(\widetilde{\mathbf{H}}\big)_{ij} = \frac{1}{d} \langle \widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}_j \rangle.$$

Then we can decompose $\overline{\mathbf{B}}_{[N:P-1],[N:P-1]}$ into

$$\overline{\mathbf{B}}_{[N:P-1],[N:P-1]} = q_3 \mathbf{I}_n + q_5 \widetilde{\mathbf{H}} + \frac{q_5}{d} \mathbf{u}\mathbf{u}^\top. \tag{IV.12}$$

For $\overline{\mathbf{B}}_{[N:P-1],[1:N-1]}$, by definition we see that the elements in $\overline{\mathbf{B}}_{[N:P-1],[1:N-1]}$ are $(\mathbf{Z})_{i,a}$ for $a \in [N]\backslash\{N_1\}$ and $i \in [n]$. Therefore, we have

$$\begin{aligned}
(\mathbf{Z})_{i,a} &= \frac{1}{\sqrt{d}} \Phi_j\Big(\frac{1}{\sqrt{d}} \langle \overline{\mathbf{x}}_i, \overline{\boldsymbol{\theta}}_a \rangle\Big) = \frac{1}{\sqrt{d}} \Phi_j\Big(\frac{1}{\sqrt{d}} \langle \widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a \rangle + \frac{1}{d} u_i \eta_a\Big) \\
&= \frac{1}{\sqrt{d}} \Phi_j\Big(\frac{1}{\sqrt{d}} \langle \widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a \rangle\Big) + \frac{\phi_{j,1}}{d} u_i \eta_a + \frac{1}{\sqrt{d}} \Big[\Phi_{j,\perp}\Big(\frac{1}{\sqrt{d}} \langle \widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a \rangle + \frac{1}{\sqrt{d}} u_i \eta_a\Big) - \Phi_{j,\perp}\Big(\frac{1}{\sqrt{d}} \langle \widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a \rangle\Big)\Big],
\end{aligned}$$

where $\Phi_{j,\perp}(x) = \Phi_j(x) - \phi_{j,1}x$, $j = 1$ when $a \le N_1 - 1$ and $j = 2$ when $a \ge N_1 + 1$. By the symmetry of $\overline{\mathbf{B}}$, we can then decompose $\overline{\mathbf{B}}_{[N:P-1],[1:N-1]}$ and $\overline{\mathbf{B}}^\top_{[1:N-1],[N:P-1]}$ into

$$\overline{\mathbf{B}}_{[N:P-1],[1:N-1]} = \overline{\mathbf{B}}^\top_{[1:N-1],[N:P-1]} = \widetilde{\mathbf{Z}} + \frac{1}{d}\mathbf{u}\boldsymbol{\eta}\mathbf{M}_\phi + [\mathbf{E}_1, \mathbf{E}_2], \tag{IV.13}$$

where we define

$$\widetilde{\mathbf{Z}} = [\widetilde{\mathbf{Z}}_1, \widetilde{\mathbf{Z}}_2], \quad (\widetilde{\mathbf{Z}}_1)_{i,a} = \frac{1}{\sqrt{d}}\Phi_1\Big(\frac{1}{\sqrt{d}}\langle\widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a\rangle\Big), \quad (\widetilde{\mathbf{Z}}_2)_{i,a} = \frac{1}{\sqrt{d}}\Phi_2\Big(\frac{1}{\sqrt{d}}\langle\widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a\rangle\Big),$$
$$\mathbf{M}_\phi = \begin{bmatrix}\phi_{1,1}\mathbf{I}_{N_1-1} & \\ & \phi_{2,1}\mathbf{I}_{N_2}\end{bmatrix}, \quad (\mathbf{E}_j)_{i,a} = \frac{1}{\sqrt{d}}\Big[\Phi_{j,\perp}\Big(\frac{1}{\sqrt{d}}\langle\widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a\rangle + \frac{1}{\sqrt{d}}u_i\eta_a\Big) - \Phi_{j,\perp}\Big(\frac{1}{\sqrt{d}}\langle\widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a\rangle\Big)\Big]$$

for $a \in [N]\backslash\{N_1\}$, $i \in [n]$ and $j \in \{1, 2\}$. Combining (IV.11), (IV.12) and (IV.13), we decompose $\overline{\mathbf{B}}$ into

$$\overline{\mathbf{B}} = \widetilde{\mathbf{B}} + \boldsymbol{\Delta} + \mathbf{E} \in \mathbb{R}^{(P-1)\times(P-1)}, \tag{IV.14}$$

where

$$\widetilde{\mathbf{B}} = \begin{bmatrix} q_2\widetilde{\mathbf{M}}_*\widetilde{\mathbf{M}}_* + \frac{q_4}{d}\widetilde{\mathbf{M}}_1\widetilde{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Theta}}^\top\widetilde{\mathbf{M}}_1 & \widetilde{\mathbf{Z}}^\top \\ \widetilde{\mathbf{Z}} & q_3\mathbf{I}_n + q_5\widetilde{\mathbf{H}}\end{bmatrix}$$
$$= \begin{bmatrix} q_2\mu_{1,*}^2\mathbf{I}_{N_1-1} + \frac{q_4\mu_{1,1}^2}{d}\widetilde{\boldsymbol{\Theta}}_1\widetilde{\boldsymbol{\Theta}}_1^\top & \frac{q_4\mu_{1,1}\mu_{2,1}}{d}\widetilde{\boldsymbol{\Theta}}_1\widetilde{\boldsymbol{\Theta}}_2^\top & \widetilde{\mathbf{Z}}_1^\top \\ \frac{q_4\mu_{1,1}\mu_{2,1}}{d}\widetilde{\boldsymbol{\Theta}}_2\widetilde{\boldsymbol{\Theta}}_1^\top & q_2\mu_{2,*}^2\mathbf{I}_{N_2} + \frac{q_4\mu_{2,1}^2}{d}\widetilde{\boldsymbol{\Theta}}_2\widetilde{\boldsymbol{\Theta}}_2^\top & \widetilde{\mathbf{Z}}_2^\top \\ \widetilde{\mathbf{Z}}_1 & \widetilde{\mathbf{Z}}_2 & q_3\mathbf{I}_n + q_5\widetilde{\mathbf{H}}\end{bmatrix},$$
$$\boldsymbol{\Delta} = \begin{bmatrix}\frac{q_4}{d}\widetilde{\mathbf{M}}_1\boldsymbol{\eta}\boldsymbol{\eta}^\top\widetilde{\mathbf{M}}_1 & \frac{1}{d}\mathbf{M}_\phi\boldsymbol{\eta}\mathbf{u}^\top \\ \frac{1}{d}\mathbf{u}\boldsymbol{\eta}\mathbf{M}_\phi & \frac{q_5}{d}\mathbf{u}\mathbf{u}^\top\end{bmatrix} = \begin{bmatrix}\frac{q_4\mu_{1,1}^2}{d}\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top & \frac{q_4\mu_{1,1}\mu_{2,1}}{d}\boldsymbol{\eta}_1\boldsymbol{\eta}_2^\top & \frac{\phi_{1,1}}{d}\boldsymbol{\eta}_1\mathbf{u}^\top \\ \frac{q_4\mu_{1,1}\mu_{2,1}}{d}\boldsymbol{\eta}_2\boldsymbol{\eta}_1^\top & \frac{q_4\mu_{2,1}^2}{d}\boldsymbol{\eta}_2\boldsymbol{\eta}_2^\top & \frac{\phi_{2,1}}{d}\boldsymbol{\eta}_2\mathbf{u}^\top \\ \frac{\phi_{1,1}}{d}\mathbf{u}\boldsymbol{\eta}_1^\top & \frac{\phi_{2,1}}{d}\mathbf{u}\boldsymbol{\eta}_2^\top & \frac{q_5}{d}\mathbf{u}\mathbf{u}^\top\end{bmatrix}, \ \mathbf{E} = \begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{E}_1^\top \\ \mathbf{0} & \mathbf{0} & \mathbf{E}_2^\top \\ \mathbf{E}_1 & \mathbf{E}_2 & \mathbf{0}\end{bmatrix}.$$

Clearly, by the definition of $\widetilde{\mathbf{B}}$, the Stieltjes transform corresponding to $\widetilde{\mathbf{B}}$ shares the same asymptotics as the Stieltjes transform corresponding to $\overline{\mathbf{A}}$.

Step 2. According to our analysis in **Step 1**, we can then calculate $\overline{m}_{1,d}$ by (IV.8), in which the terms $\overline{\mathbf{A}}_{\cdot,N_1}$ and $\overline{\mathbf{B}}$ have the decompositions (IV.10) and (IV.14) respectively. In this step, we aim to further simplify the calculation by getting rid of the terms $\|\overline{\boldsymbol{\theta}}_{N_1}\|_2^2/d$ in (IV.8) and $\mathbf{E}$ in (IV.14). Define

$$w_0 = \Big(-\xi + q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2 - \overline{\mathbf{A}}_{\cdot,N_1}^\top(\overline{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\overline{\mathbf{A}}_{\cdot,N_1}\Big)^{-1},$$
$$w_1 = \Big(-\xi + q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2\|\overline{\boldsymbol{\theta}}_{N_1}\|_2^2/d - \overline{\mathbf{A}}_{\cdot,N_1}^\top(\overline{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\overline{\mathbf{A}}_{\cdot,N_1}\Big)^{-1},$$
$$w_2 = \Big(-\xi + q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2 - \overline{\mathbf{A}}_{\cdot,N_1}^\top(\widetilde{\mathbf{B}} + \boldsymbol{\Delta} - \xi\mathbf{I}_{P-1})^{-1}\overline{\mathbf{A}}_{\cdot,N_1}\Big)^{-1}.$$

Then by (IV.8), we have $\overline{m}_{1,d} = \psi_1\mathbb{E}w_1$. We now give an upper bound of $|w_1 - w_2|$. Recall that we consider a fixed $\xi \in \mathbb{C}_+$. Since $\overline{\mathbf{B}}$ is a real symmetric matrix, by diagonalizing $\overline{\mathbf{B}}$, it is easy to see that $\Im\big(\overline{\mathbf{A}}_{\cdot,N_1}^\top(\overline{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\overline{\mathbf{A}}_{\cdot,N_1}\big) \ge 0$. Therefore, we deterministically have

$$\Im(-w_1^{-1}) = \Im(\xi) + \Im\big(\overline{\mathbf{A}}_{\cdot,N_1}^\top(\overline{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\overline{\mathbf{A}}_{\cdot,N_1}\big) \ge \Im(\xi).$$

Thus we have $|w_1| \le 1/\Im(\xi)$. Using a similar argument, we have $\max\{|w_0|, |w_1|, |w_2|\} \le 1/\Im(\xi)$, which indicates that $|w_1 - w_2| \le 2/\Im(\xi)$. Moreover, we have

$$\begin{aligned}
|w_1 - w_2| &\le |w_1 - w_0| + |w_0 - w_2| \\
&\le q_4\mu_{1,1}^2 \big| w_1 \big(\overline{\boldsymbol{\theta}}_{N_1}\|_2^2/d - 1\big) w_0 \big| + \big| w_1 w_2 \overline{\mathbf{A}}_{\cdot,N_1}^\top \big((\overline{\mathbf{B}} - \xi \mathbf{I}_{P-1})^{-1} - (\widetilde{\mathbf{B}} + \boldsymbol{\Delta} - \xi \mathbf{I}_{P-1})^{-1} \big) \overline{\mathbf{A}}_{\cdot,N_1} \big| \\
&\le q_4\mu_{1,1}^2 |\|\overline{\boldsymbol{\theta}}_{N_1}\|_2^2/d - 1| / \Im^2(\xi) + 2\|\overline{\mathbf{A}}_{\cdot,N_1}\|_2^2 \|\mathbf{E}\|_{\mathrm{op}} / \Im^4(\xi)
\end{aligned}$$

By Lemma IV.4, we have $\|\mathbf{E}_1\|_{\mathrm{op}} = o_{\mathbb{P}}(1)$, $\|\mathbf{E}_2\|_{\mathrm{op}} = o_{\mathbb{P}}(1)$. It is also easy to see that $\|\overline{\mathbf{A}}_{\cdot,N_1}\|_2^2 = O_{\mathbb{P}}(1)$ and $\|\overline{\boldsymbol{\theta}}_{N_1}\|_2^2/d - 1 = o_{\mathbb{P}}(1)$. Therefore we have

$$|w_1 - w_2| = o_{\mathbb{P}}(1).$$

Combining with the fact that $|w_1 - w_2|$ is deterministically bounded by $2/\Im(\xi)$, by the dominated convergence theorem, we have

$$\mathbb{E}|w_1 - w_2| = o_d(1).$$

Therefore $\overline{m}_{1,d} = \psi_1 \mathbb{E} w_2 + o_d(1)$, and the derivation of the resolvent equations reduces to the calculation of $\mathbb{E} w_2$.

Step 3. We calculate $\mathbb{E} w_2$ to get the resolvent equations. For simplicity, we give some notations which will be used later. Let

$$\mathbf{v} = \overline{\mathbf{A}}_{\cdot,N_1}, \quad \mathbf{v}_i = \overline{\mathbf{A}}_{i,N_1} = \begin{cases} \dfrac{q_4\mu_{1,1}^2\eta_i}{d}\|\overline{\boldsymbol{\theta}}_{N_1}\|_2, & \text{if } i \in [1, N_1 - 1], \\ \dfrac{q_4\mu_{1,1}\mu_{2,1}\eta_{i+1}}{d}\|\overline{\boldsymbol{\theta}}_{N_1}\|_2, & \text{if } i \in [N_1, N-1], \\ \dfrac{1}{\sqrt{d}}\Phi_1\Big(\dfrac{1}{\sqrt{d}} u_{i-N+1}\|\overline{\boldsymbol{\theta}}_{N_1}\|_2\Big), & \text{if } i \ge N, \end{cases}$$

and

$$\mathbf{U} = \frac{1}{\sqrt{d}} \begin{bmatrix} \boldsymbol{\eta}_1 & & \\ & \boldsymbol{\eta}_2 & \\ & & \mathbf{u} \end{bmatrix} \in \mathbb{R}^{(P-1)\times 3}, \quad \mathbf{M} = \begin{bmatrix} q_4\mu_{1,1}^2 & q_4\mu_{1,1}\mu_{2,1} & \phi_{1,1} \\ q_4\mu_{1,1}\mu_{2,1} & q_4\mu_{2,1}^2 & \phi_{2,1} \\ \phi_{1,1} & \phi_{2,1} & q_5 \end{bmatrix}.$$

By direct verification, we have

$$\boldsymbol{\Delta} = \mathbf{U}\mathbf{M}\mathbf{U}^\top.$$

We now decompose w_2 into the terms related with $\widetilde{\mathbf{B}}$, $\mathbf{v}$ and $\mathbf{U}$. By Schur complement formula, we have

$$\begin{aligned}
(\widetilde{\mathbf{B}} + \mathbf{U}\mathbf{M}\mathbf{U}^\top - \xi\mathbf{I}_{P-1})^{-1} &= (\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1} \\
&\quad - (\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{U}[\mathbf{M}^{-1} + \mathbf{U}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{U}]^{-1}\mathbf{U}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}. \quad \text{(IV.15)}
\end{aligned}$$

Then w_2 can be rewritten as

$$\begin{aligned}
w_2 =& \Big(-\xi + q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2 - \mathbf{v}^\top(\widetilde{\mathbf{B}} + \mathbf{U}\mathbf{M}\mathbf{U}^\top - \xi\mathbf{I}_{P-1})^{-1}\mathbf{v} \Big)^{-1} \\
=& \big[-\xi + q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2 - \mathbf{v}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{v} \\
&+ \mathbf{v}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{U}(\mathbf{M}^{-1} + \mathbf{U}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{U})^{-1}\mathbf{U}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{v} \big]^{-1}, \quad \text{(IV.16)}
\end{aligned}$$

where the first equation is the definition of w_2, and the second equation follows by (IV.15). To continue the calculation, we study the terms $\mathbf{v}^\top(\widetilde{\mathbf{B}}-\xi\mathbf{I}_{P-1})^{-1}\mathbf{v}$, $\mathbf{v}^\top(\widetilde{\mathbf{B}}-\xi\mathbf{I}_{P-1})^{-1}\mathbf{U}$ and $\mathbf{U}^\top(\widetilde{\mathbf{B}}-\xi\mathbf{I}_{P-1})^{-1}\mathbf{U}$ in the denominator of (IV.16). To do so, we note that $\widetilde{\mathbf{B}}$ is independent of $\mathbf{v}$ and $\mathbf{U}$. Moreover, by the leave-one-out argument, the Stieltjes transform corresponding to $\widetilde{\mathbf{B}}$ shares the same asymptotics as the Stieltjes transform corresponding to $\overline{\mathbf{A}}$. Notice that η_i is independent on $\widetilde{\mathbf{B}}$ conditioned on $\overline{\boldsymbol{\theta}}_{N_1}$, and $\widetilde{\mathbf{B}}$ is independent on $\overline{\boldsymbol{\theta}}_{N_1}$. We have

$$\begin{aligned}
&\mathbb{E}\mathbf{v}^\top(\widetilde{\mathbf{B}}-\xi\mathbf{I}_{P-1})^{-1}\mathbf{v}=\mathbb{E}\mathrm{tr}(\widetilde{\mathbf{B}}-\xi\mathbf{I}_{P-1})^{-1}\mathbf{v}\mathbf{v}^\top=\mathrm{tr}\Big(\mathbb{E}(\widetilde{\mathbf{B}}-\xi\mathbf{I}_{P-1})^{-1}\mathbb{E}\mathbf{v}\mathbf{v}^\top\Big)\\
=&\mathrm{tr}\Bigg(\begin{bmatrix}\frac{d\overline{m}_{1,d}}{N_1}\mathbf{I}_{N_1-1} & * & *\\ * & \frac{d\overline{m}_{2,d}}{N_2}\mathbf{I}_{N_2} & *\\ * & * & \frac{d\overline{m}_{3,d}}{n}\mathbf{I}_n\end{bmatrix}\\
&\cdot\frac{1}{d}\begin{bmatrix}(q_4^2\mu_{1,1}^4+o_d(1))\mathbf{I}_{N_1-1} & & \\ & (q_4^2\mu_{1,1}^2\mu_{2,1}^2+o_d(1))\mathbf{I}_{N_2} & \\ & & (\phi_{1,1}^2+\phi_{1,*}^2+o_d(1))\mathbf{I}_n\end{bmatrix}\Bigg)\\
=&q_4^2\mu_{1,1}^2(\mu_{1,1}^2\overline{m}_{1,d}+\mu_{2,1}^2\overline{m}_{2,d})+(\phi_{1,1}^2+\phi_{1,*}^2)\overline{m}_{3,d}+o_d(1),
\end{aligned}$$

where the second equality follows from the fact that $\mathbb{E}\Phi_1^2\Big(\frac{1}{\sqrt{d}}u_{i-N+1}\|\overline{\boldsymbol{\theta}}_{N_1}\|_2\Big)=\phi_{1,1}^2+\phi_{1,*}^2+o_d(1)$, and we have denoted by '*' the blocks that are irrelevant to the calculation. By a concentration measure argument (see in Tao (2012) Section 2.4.3), we have

$$\mathbf{v}^\top(\widetilde{\mathbf{B}}-\xi\mathbf{I}_{P-1})^{-1}\mathbf{v}=q_4^2\mu_{1,1}^2(\mu_{1,1}^2\overline{m}_{1,d}+\mu_{2,1}^2\overline{m}_{2,d})+(\phi_{1,1}^2+\phi_{1,*}^2)\overline{m}_{3,d}+o_{\mathbb{P}}(1). \tag{IV.17}$$

After direct calculation with the same argument, we obtain that

$$\mathbf{v}^\top(\widetilde{\mathbf{B}}-\xi\mathbf{I}_{P-1})^{-1}\mathbf{U}=\begin{bmatrix}q_4\mu_{1,1}^2\overline{m}_{1,d}\\ q_4\mu_{1,1}\mu_{2,1}\overline{m}_{2,d}\\ \phi_{1,1}\overline{m}_{3,d}\end{bmatrix}^\top+o_{\mathbb{P}}(1), \tag{IV.18}$$

$$\mathbf{U}^\top(\widetilde{\mathbf{B}}-\xi\mathbf{I}_{P-1})^{-1}\mathbf{U}=\begin{bmatrix}\overline{m}_{1,d} & & \\ & \overline{m}_{2,d} & \\ & & \overline{m}_{3,d}\end{bmatrix}+o_{\mathbb{P}}(1). \tag{IV.19}$$

Now since $|w_2|\le 1/\xi_0$ is deterministically bounded, by dominated convergence theorem, we have the L_1 convergence of w_2 by plugging the main terms of (IV.17), (IV.18) and (IV.19) into (IV.16). Further note that equation (IV.16) has a part in the form of $\left(\mathbf{A}^{-1}+\mathbf{M}^{-1}\right)^{-1}$, where

$$\mathbf{A}=\begin{bmatrix}1/\overline{m}_{1,d} & & \\ & 1/\overline{m}_{2,d} & \\ & & 1/\overline{m}_{3,d}\end{bmatrix},\quad \mathbf{M}=\begin{bmatrix}q_4\mu_{1,1}^2 & q_4\mu_{1,1}\mu_{2,1} & \phi_{1,1}\\ q_4\mu_{1,1}\mu_{2,1} & q_4\mu_{2,1}^2 & \phi_{2,1}\\ \phi_{1,1} & \phi_{2,1} & q_5\end{bmatrix}.$$

By the formula $\left(\mathbf{A}^{-1}+\mathbf{M}^{-1}\right)^{-1}=\mathbf{A}-\mathbf{A}\left(\mathbf{A}+\mathbf{M}\right)^{-1}\mathbf{A}$, we have

$$\left(\mathbf{M}^{-1}+\mathbf{A}^{-1}\right)^{-1}=\begin{bmatrix}1/\overline{m}_{1,d} & & \\ & 1/\overline{m}_{2,d} & \\ & & 1/\overline{m}_{3,d}\end{bmatrix}-\mathbf{A}\begin{bmatrix}q_4\mu_{1,1}^2+1/\overline{m}_{1,d} & q_4\mu_{1,1}\mu_{2,1} & \phi_{1,1}\\ q_4\mu_{1,1}\mu_{2,1} & q_4\mu_{2,1}^2+1/\overline{m}_{2,d} & \phi_{2,1}\\ \phi_{1,1} & \phi_{2,1} & q_5+1/\overline{m}_{3,d}\end{bmatrix}^{-1}\mathbf{A}.$$

Denote $\boldsymbol{l} = [q_4\mu_{1,1}^2\overline{m}_{1,d} \quad q_4\mu_{1,1}\mu_{2,1}\overline{m}_{2,d} \quad \phi_{1,1}\overline{m}_{3,d}]^\top$. Then by plugging the equation above into (IV.16), and combing it with (IV.17), (IV.18) and (IV.19), we finally get

$$\begin{aligned}
\overline{m}_{1,d} =&\psi_1\mathbb{E}w_2\\
=&\psi_1\Big\{ -\xi + q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2 - q_4^2\mu_{1,1}^2(\mu_{1,1}^2\overline{m}_{1,d} + \mu_{2,1}^2\overline{m}_{2,d})\\
&- (\phi_{1,1}^2 + \phi_{1,*}^2)\overline{m}_{3,d} + \boldsymbol{l}^\top\mathbf{A}\boldsymbol{l} - \boldsymbol{l}^\top\mathbf{A}(\mathbf{A}+\mathbf{M})^{-1}\mathbf{A}\boldsymbol{l}\Big\}^{-1} + o_d(1)\\
=&\psi_1\left\{ -\xi + q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2 - \phi_{1,*}^2\overline{m}_{3,d} - \begin{bmatrix} q_4\mu_{1,1}^2\\ q_4\mu_{1,1}\mu_{2,1}\\ \phi_{1,1}\end{bmatrix}^\top (\mathbf{A}+\mathbf{M})^{-1}\begin{bmatrix} q_4\mu_{1,1}^2\\ q_4\mu_{1,1}\mu_{2,1}\\ \phi_{1,1}\end{bmatrix}\right\}^{-1}\\
&+ o_d(1).
\end{aligned}$$

Now note that $\phi_{j,1}^2 = \mu_{j,1}^2(1+q_1)^2$, $\phi_{j,*}^2 = \mu_{j,*}^2$, $j = 1, 2$. Therefore with direct calculation, we have

$$\overline{m}_{1,d} =\psi_1\left\{ -\xi + q_2\mu_{1,*}^2 - \mu_{1,*}^2\overline{m}_{3,d} + \frac{H_{1,d}}{H_{D,d}}\right\}^{-1} + o_d(1), \tag{IV.20}$$

where

$$\begin{aligned}
H_{1,d} =&\mu_{1,1}^2 q_4(1+\overline{m}_{3,d}q_5) - \mu_{1,1}^2(1+q_1)^2\overline{m}_{3,d},\\
H_{D,d} =&(1+\mu_{1,1}^2\overline{m}_{1,d}q_4 + \mu_{2,1}^2\overline{m}_{2,d}q_4)(1+\overline{m}_{3,d}q_5) - \mu_{2,1}^2(1+q_1)^2\overline{m}_{2,d}\overline{m}_{3,d}\\
&- \mu_{1,1}^2(1+q_1)^2\overline{m}_{1,d}\overline{m}_{3,d}.
\end{aligned}$$

The equation above shows that the magnitude of the first element of $\overline{\mathbf{m}}_d(\xi) - \mathbf{F}(\overline{\mathbf{m}}_d(\xi))$ is $o_d(1)$. With exactly the same proof, we also have

$$\begin{aligned}
\overline{m}_{2,d} =&\psi_2\left\{ -\xi + q_2\mu_{2,*}^2 - \mu_{2,*}^2\overline{m}_{3,d} + \frac{H_{2,d}}{H_{D,d}}\right\}^{-1} + o_d(1),\\
\overline{m}_{3,d} =&\psi_3\left\{ -\xi + q_3 - \mu_{1,*}^2\overline{m}_{1,d} - \mu_{2,*}^2\overline{m}_{2,d} + \frac{H_{3,d}}{H_{D,d}}\right\}^{-1} + o_d(1),
\end{aligned} \tag{IV.21}$$

where

$$\begin{aligned}
H_{2,d} =&\mu_{2,1}^2 q_4(1+\overline{m}_{3,d}q_5) - \mu_{2,1}^2(1+q_1)^2\overline{m}_{3,d},\\
H_{3,d} =&q_5(1+\mu_{1,1}^2\overline{m}_{1,d}q_4 + \mu_{2,1}^2\overline{m}_{2,d}q_4) - \mu_{2,1}^2(1+q_1)^2\overline{m}_{2,d} - \mu_{1,1}^2(1+q_1)^2\overline{m}_{1,d}.
\end{aligned}$$

This completes the proof of Lemma IV.5. ■

IV.3 Proof for conclusions 1 and 2 in Proposition A.6

We first introduce an important lemma about the property of Stieltjes transforms, which is given in Hastie et al. (2022).

Lemma IV.6 (Lemma 7 in Hastie et al. (2022)). *The functions $\xi \to \overline{m}_{i,d}(\xi)$, $i = 1, 2, 3$, have the following properties:*

1. $\overline{m}_{i,d}$, $i = 1, 2, 3$ are analytical on $\mathbb{C}_+$, and map $\mathbb{C}_+$ into $\mathbb{C}_+$.

2. *Let $\Omega \subset \mathbb{C}_+$ be a set with an accumulation point. If $\overline{m}_{i,d} \to m_i(\xi)$ for all $\xi \in \Omega$, then $m_i(\xi)$ has an unique analytic continuation to $\mathbb{C}_+$ and $\overline{m}_{i,d} \to m_i(\xi)$ for all $\xi \in \mathbb{C}_+$. Moreover, the convergence is uniform over compact sets $\Omega \subset \mathbb{C}_+$.*

We now give the proof of the conclusions in Proposition A.6 that $\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu})$ is analytic on $\{\xi : \Im(\xi) > \xi_0\}$ for some sufficiently large ξ_0, has unique analytic continuation to $\mathbb{C}_+$ and maps $\mathbb{C}_+$ to $\mathbb{C}_+^3$. Denote $\overline{\mathbf{m}}_d = \overline{\mathbf{m}}_d(\xi) = [\overline{m}_{1,d}(\xi), \overline{m}_{2,d}(\xi), \overline{m}_{3,d}(\xi)]^\top$. Then for any fixed $\xi \in \mathbb{C}_+$, Lemma IV.5 gives

$$||\overline{\mathbf{m}}_d - \mathbf{F}(\overline{\mathbf{m}}_d)||_2 = o_d(1). \tag{IV.22}$$

By Lemma III.1, there exists a $\xi_0 > 0$, such that for all ξ with $\Im(\xi) \geq \xi_0$, $\mathbf{F}(\cdot)$ is 1/2-Lipschitz with respect to ℓ_2 norm. Moreover, for all ξ with $\Im(\xi) \geq \xi_0$ we have

$$\begin{aligned}\|\overline{\mathbf{m}}_d - \mathbf{m}\|_2 &= \|\overline{\mathbf{m}}_d - \mathbf{F}(\mathbf{m})\|_2 \\ &\leq \|\overline{\mathbf{m}}_d - \mathbf{F}(\overline{\mathbf{m}}_d)\|_2 + \|\mathbf{F}(\overline{\mathbf{m}}_d) - \mathbf{F}(\mathbf{m})\|_2 \\ &\leq o_d(1) + \frac{1}{2} \cdot \|\overline{\mathbf{m}}_d - \mathbf{m}\|_2,\end{aligned}$$

where the equality is by the definition of $\mathbf{m}$ as the unique fixed point of $\mathbf{F}(\cdot)$, the first inequality is by triangle inequality, the second inequality is by (IV.22) and the fact that $\mathbf{F}(\cdot)$ is 1/2-Lipschitz with respect to ℓ_2 norm. Therefore we have $\|\overline{\mathbf{m}}_d(\xi) - \mathbf{m}(\xi)\|_2 = o_d(1)$ for all ξ with $\Im(\xi) \geq \xi_0$. The properties of Stieljes transforms (see Lemma IV.6) then imply that $\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu})$ is analytic in $\{\xi : \Im(\xi) > \xi_0\}$, and has a unique analytic continuation to $\mathbb{C}_+$. Moreover, the extended $\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu})$ satisfies

$$||\overline{\mathbf{m}}_d(\xi) - \mathbf{m}(\xi)||_2 = o_d(1) \tag{IV.23}$$

for any fixed $\xi \in \mathbb{C}_+$. This implies that $\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu})$ is $\mathbb{C}_+ \to \mathbb{C}_+^3$ by the definition of $\overline{\mathbf{m}}_d$. The proof of Conclusion 1 is complete.

To prove Conclusion 2, we first prove that $\mathbf{m}(\xi)$ is a continuity point of $\mathbf{F}(\cdot)$ for any fixed $\xi \in \mathbb{C}_+$. For any fixed $\xi \in \mathbb{C}_+$, assume that $\mathbf{m}(\xi)$ is not a continuity point of $\mathbf{F}(\cdot)$, by the definition of $\mathbf{F}(\cdot)$ we have $\|\mathbf{F}(\mathbf{m}(\xi))\|_2 = +\infty$. Therefore, for any $M > 0$, there exsits $\delta(\xi, M) > 0$ ($\xi \in \mathbb{C}_+$ is fixed here), as long as $\|\overline{\mathbf{m}}_d(\xi) - \mathbf{m}(\xi)\| < \delta(\xi, M)$, the inequality $\mathbf{F}(\overline{\mathbf{m}}_d(\xi)) > M$ holds. Moreover, for the $\delta(\xi, M)$, there always exists d_0 such that $\|\overline{\mathbf{m}}_d(\xi) - \mathbf{m}(\xi)\| < \delta(\xi, M)$ for all $d > d_0$. That is: for any fixed $\xi \in \mathbb{C}_+$, and any large constant $M > 0$, there always exists d_0 such that $\mathbf{F}(\overline{\mathbf{m}}_d(\xi)) > M$ for $d > d_0$. Combined with (IV.22), there exists $d_1 > 0$, such that $||\overline{\mathbf{m}}_d(\xi) - \mathbf{F}(\overline{\mathbf{m}}_d(\xi))||_2 < 1$ for all $d > d_1$. Then for $d > \max\{d_0, d_1\}$, we have $\mathbf{F}(\overline{\mathbf{m}}_d(\xi)) > M$ and $||\overline{\mathbf{m}}_d(\xi) - \mathbf{F}(\overline{\mathbf{m}}_d(\xi))||_2 < 1$. We have $\|\overline{\mathbf{m}}_d(\xi)\|_2 > M - 1$ for $d > \max\{d_0, d_1\}$. On the other hand, $\|\overline{\mathbf{m}}_d(\xi)\|_2 \leq 2(\psi_1 + \psi_2 + \psi_3)/\Im(\xi)$ from the definition of $\overline{\mathbf{m}}_d(\xi)$. Note that $\xi \in \mathbb{C}_+$ is fixed here. Enlarging M leads to a contradiction. Therefore, $\mathbf{m}(\xi)$ is the continuity point of $\mathbf{F}(\cdot)$ for any fixed $\xi \in \mathbb{C}_+$.

For any fixed $\xi \in \mathbb{C}_+$, note that $\mathbf{m}(\xi)$ is the continuity point of $\mathbf{F}(\cdot)$. Let $d \to +\infty$, (IV.22) and (IV.23) give us

$$||\mathbf{m} - \mathbf{F}(\mathbf{m})||_2 = 0.$$

This means that $\mathbf{F}(\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu})) = \mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu})$ for any fixed $\xi \in \mathbb{C}_+$. The proof of Conclusion 2 is complete.

IV.4 Point convergence for polynomial activation functions

We now give the proof of point convergence under the additional assumption that the activation functions are polynomials. We remind readers the "Gaussian version" of the problem defined in Sections IV.1 and IV.2, where the data inputs $\overline{\mathbf{x}}_i$, $i \in [n]$ and $\overline{\boldsymbol{\theta}}_a$, $a \in [N]$ are defined in Definition IV.1 and the activation functions $\phi_1(x), \phi_2(x)$ are given in (IV.2). We also remind readers that the "Gaussian version" and "spherical version" Stieltjes transforms of the empirical eigenvalue distributions of linear pencil matrices are denoted as $\overline{M}_d(\xi)$ and $M_d(\xi)$, respectively. Importantly, the expectation of $\overline{M}_d(\xi)$ is denoted as $\overline{m}_d$, while $m(\xi; \mathbf{q}, \boldsymbol{\mu}) = \sum_{i=1}^3 m_i(\xi)$, where $\mathbf{m} = \mathbf{m}(\xi) = (m_1(\xi), m_2(\xi), m_3(\xi))^\top$ is defined as the solution of (A.2) on $\{\xi : \Im(\xi) \geq \xi_0\}$ and then extended to $\mathbb{C}_+$ by analytic continuation.

By (IV.7), for all fixed $\xi \in \mathbb{C}_+$, we have

$$\mathbb{E}\left|\overline{M}_d(\xi) - \sum_{i=1}^3 \overline{m}_{i,d}(\xi)\right| = o_d(1). \tag{IV.24}$$

In addition, by Lemma IV.3, when the activation functions are polynomials, we have

$$\mathbb{E}\big|M_d(\xi) - \overline{M}_d(\xi)\big| = o_d(1). \tag{IV.25}$$

Combining (IV.23) (IV.24) and (IV.25) gives

$$\mathbb{E}\big|M_d(\xi) - m(\xi)\big| = o_d(1)$$

for any fixed $\xi \in \mathbb{C}_+$, which completes the proof of the point convergence for polynomial activation functions.

IV.5 Point convergence for general activation functions satisfying Assumption 3.2

We now extend the result for polynomial activation functions to general activation functions satisfying Assumption 3.2. Let τ_d be the marginal distribution of $\langle \mathbf{x}, \boldsymbol{\theta}\rangle/\sqrt{d}$ for $\mathbf{x}, \boldsymbol{\theta} \sim_{\text{iid}} \text{Unif}(\sqrt{d}\cdot\mathbb{S}^{d-1})$, and $\overline{\tau}_d$ the marginal distribution of $\langle \overline{\mathbf{x}}, \overline{\boldsymbol{\theta}}\rangle/\sqrt{d}$ for $\overline{\mathbf{x}}, \overline{\boldsymbol{\theta}} \sim_{\text{iid}} \text{N}(0, \mathbf{I}_d)$. For $j = 1, 2$, suppose that σ_j are activation functions satisfying Assumption 3.2. The idea here is to construct polynomial activation functions $\widetilde{\sigma}_j$ to approximate σ_j. To do so, we recall that $\mathbf{m} = \mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu})$ solves the implicit equations

$$\mathbf{m} = \mathbf{F}(\mathbf{m}; \xi, \mathbf{q}, \boldsymbol{\mu}),$$

where $\mathbf{F}(\cdot; \xi, \mathbf{q}, \boldsymbol{\mu})$ is defined in Definition A.5. When $\Im(\xi) > \xi_0$ for some large enough ξ_0, by the continuity of the solution of the fixed point equation with respect to $\boldsymbol{\mu}$, we have

$$\lim_{\widetilde{\boldsymbol{\mu}} \to \boldsymbol{\mu}} \mathbf{m}(\xi; \mathbf{q}, \widetilde{\boldsymbol{\mu}}) = \mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu}).$$

According to our proof in Section IV.3, we can extend the definition of $\mathbf{m}$ to $\xi \in \mathbb{C}_+$ with analytic continuation. Then with the same proof as in Mei and Montanari (2022) (see equation (10.56) in Mei and Montanari (2022)), for any fixed $\xi \in \mathbb{C}_+$ and any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon, \xi, \mathbf{q}, \boldsymbol{\mu}) > 0$

such that

$$\|\mathbf{m}(\xi;\mathbf{q},\widetilde{\boldsymbol{\mu}})-\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})\|_2\le\varepsilon \tag{IV.26}$$

for all $\widetilde{\boldsymbol{\mu}}$ with $\|\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}}\|_2\le\delta$. Now by Assumption 3.2, for any fixed $\varepsilon>0$, we can choose a sufficiently large integer $\bar{k}$ and construct

$$\widetilde{\sigma}_j(x)=\sum_{k=0}^{\bar{k}}\frac{\mu_{j,k}}{k!}H_k(x),$$

such that for $G\sim N(0,1)$,

$$\mathbb{E}[\sigma_j(G)-\widetilde{\sigma}_j(G)]^2\le\varepsilon^2, \tag{IV.27}$$
$$|\mathbb{E}\{\widetilde{\sigma}_j(G)^2\}-\mathbb{E}\{\widetilde{\sigma}_j(G)^2\}|\le\delta^2/2. \tag{IV.28}$$

Here, $\{H_k(x)\}$ are the family of Hermite polynomials. Then by (IV.27) and Lemma 5 in Ghorbani et al. (2021), we have

$$\|\sigma_j-\widetilde{\sigma}_j\|_{L^2(\tau_d)}\le\varepsilon \tag{IV.29}$$

for $j=1,2$ and sufficiently large d, where we denote $\|\sigma_j-\widetilde{\sigma}_j\|_{L^2(\nu)}=\int(\sigma_j(x)-\widetilde{\sigma}_j(x))^2\nu(dx)$.

Following Definition 3.1, we can also define the parameters $\widetilde{\mu}_{j,0},\widetilde{\mu}_{j,1},\widetilde{\mu}_{j,*}$ corresponding to the polynomial activation functions $\widetilde{\sigma}_j$ by

$$\widetilde{\mu}_{j,0}\triangleq\mathbb{E}\{\widetilde{\sigma}_j(G)\},\ \widetilde{\mu}_{j,1}\triangleq\mathbb{E}\{G\widetilde{\sigma}_j(G)\},\ \widetilde{\mu}_{j,*}^2\triangleq\mathbb{E}\{\widetilde{\sigma}_j(G)^2\}-\widetilde{\mu}_{j,0}^2-\widetilde{\mu}_{j,1}^2.$$

Then by the definition of $\widetilde{\sigma}_j$, we have $\mu_{j,0}=\widetilde{\mu}_{j,0}$, $\mu_{j,1}=\widetilde{\mu}_{j,1}$ for $j=1,2$. Moreover, we also have

$$|\mu_{j,*}-\widetilde{\mu}_{j,*}|\le\sqrt{|\mu_{j,*}^2-\widetilde{\mu}_{j,*}^2|}=\sqrt{|\mathbb{E}\{\sigma_j(G)^2-\widetilde{\sigma}_j(G)^2\}|}\le\delta/\sqrt{2}$$

for $j=1,2$, where the first inequality follows from $|a-b|\le\sqrt{|a^2-b^2|}$ for all $a,b>0$, the equality follows by $\mu_{j,0}=\widetilde{\mu}_{j,0}$, $\mu_{j,1}=\widetilde{\mu}_{j,1}$ for $j=1,2$, and the last inequality follows by (IV.28). Therefore we have $\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2\le\delta$.

Let $\widetilde{\mathbf{m}}(\xi)=[\widetilde{m}_1(\xi),\widetilde{m}_2(\xi),\widetilde{m}_3(\xi)]^\top$ be the solution of the implicit equations

$$\widetilde{\mathbf{m}}=\mathbf{F}(\widetilde{\mathbf{m}};\xi,\mathbf{q},\widetilde{\boldsymbol{\mu}}),$$

and let $\widetilde{m}(\xi)=\widetilde{m}_1(\xi)+\widetilde{m}_2(\xi)+\widetilde{m}_3(\xi)$, where we drop the arguments $\mathbf{q},\widetilde{\boldsymbol{\mu}}$ in $\widetilde{\mathbf{m}}(\xi;\mathbf{q},\widetilde{\boldsymbol{\mu}})$ for notation simplification. Then by (IV.26), we have

$$\big|\widetilde{m}(\xi)-m(\xi)\big|\le 3\varepsilon. \tag{IV.30}$$

Let $\widetilde{\mathbf{A}}$ be the linear pencil matrix corresponding to $\widetilde{\sigma}$ in Definition A.3, and define $\widetilde{M}_d(\xi)=(1/d)\cdot\mathrm{tr}[(\widetilde{\mathbf{A}}-\xi\mathbf{I})^{-1}]$. Then we have

$$\mathbb{E}\big[|M_d(\xi)-\widetilde{M}_d(\xi)|\big]=\frac{1}{d}\mathbb{E}\big[\big|\mathrm{tr}[(\mathbf{A}-\xi\mathbf{I})^{-1}(\widetilde{\mathbf{A}}-\mathbf{A})(\widetilde{\mathbf{A}}-\xi\mathbf{I})^{-1}]\big|\big]$$

$$
\begin{aligned}
\leq & \frac{1}{d}\mathbb{E}\big[\|(\mathbf{A}-\xi\mathbf{I})^{-1}(\widetilde{\mathbf{A}}-\xi\mathbf{I})^{-1}\|_{\mathrm{op}}\|\widetilde{\mathbf{A}}-\mathbf{A}\|_{\star}\big] \\
\leq & \big[1/(\Im(\xi)^2\big]\cdot P^{-1/2}\cdot\mathbb{E}\{\|\widetilde{\mathbf{A}}-\mathbf{A}\|_F^2\}^{1/2} \\
\leq & C'(\xi,\boldsymbol{\psi})\cdot\big[1/(\Im(\xi)^2\big]\cdot d^{-1/2}\cdot\mathbb{E}\{\|\widetilde{\mathbf{A}}-\mathbf{A}\|_F^2\}^{1/2} \\
\leq & C''(\xi,\mathbf{q})\cdot\big(\|\sigma_1-\widetilde{\sigma}_1\|_{L^2(\tau_d)}+\|\sigma_2-\widetilde{\sigma}_2\|_{L^2(\tau_d)}\big),
\end{aligned}
\tag{IV.31}
$$

where $C'(\xi,\boldsymbol{\psi})>0$ is a constant only depending on ξ and $\boldsymbol{\psi}$, and $C''(\xi,\mathbf{q})>0$ only depends on ξ, $\mathbf{q}$ and $\boldsymbol{\psi}$. Here the second inequality above follows by Cauchy-Schwarz inequality, the third inequality follows by $P=N_1+N_2+n$ and the assumption that N_1,N_2,n,d goes to infinity proportionally, and the last inequality follows by the definitions of $\widetilde{\mathbf{A}}$ and $\mathbf{A}$. Therefore, by (IV.29) and (IV.31), we have

$$
\mathbb{E}|M_d(\xi)-\widetilde{M}_d(\xi)|\leq 2C''(\xi,\mathbf{q})\cdot\varepsilon \tag{IV.32}
$$

for sufficiently large d. Moreover, since $\widetilde{\sigma}_j$, $j=1,2$ are polynomial activation functions, by the results in Appendix IV.4, we have

$$
\mathbb{E}\big|\widetilde{M}_d(\xi)-\widetilde{m}(\xi)\big|=o_d(1). \tag{IV.33}
$$

Combining (IV.30), (IV.32) and (IV.33) and taking $d\to\infty$, we have

$$
\limsup_{d\to+\infty}\mathbb{E}\big|M_d(\xi)-m(\xi)\big|\leq(2C''(\xi,\mathbf{q})+3)\cdot\varepsilon
$$

for all fixed $\xi\in\mathbb{C}_+$. Taking $\varepsilon\to 0^+$, we conclude that $\lim_{d\to\infty}\mathbb{E}\big|\widetilde{M}_d(\xi)-\widetilde{m}(\xi)\big|=0$, which proves the point convergence for general activation functions.

IV.6 Uniform convergence on compact sets

In this section, we aim to prove that on compact sets the point convergence established above could be extended to uniform convergence. Consider a compact set $\Omega\subset\mathbb{C}_+$. From the proof above we have

$$
\lim_{d\to+\infty}|\mathbb{E}M_d(\xi;\mathbf{q},\boldsymbol{\mu})-m(\xi;\mathbf{q},\boldsymbol{\mu})|=0.
$$

Then from Lemma IV.6, we have

$$
\lim_{d\to+\infty}\sup_{\xi\in\Omega}|\mathbb{E}M_d(\xi;\mathbf{q},\boldsymbol{\mu})-m(\xi;\mathbf{q},\boldsymbol{\mu})|=0. \tag{IV.34}
$$

Moreover, by the definition of $M_d(\xi;\mathbf{q},\boldsymbol{\mu})$, we have

$$
\begin{aligned}
|M_d(\xi_1;\mathbf{q},\boldsymbol{\mu})-M_d(\xi_2;\mathbf{q},\boldsymbol{\mu})| &= \frac{1}{d}\big|\mathrm{tr}\big((\mathbf{A}-\xi_1\mathbf{I})^{-1}(\xi_1-\xi_2)(\mathbf{A}-\xi_2\mathbf{I})^{-1}\big)\big| \\
&\leq \frac{P}{d\cdot\Im(\xi_1)\Im(\xi_2)}\cdot|\xi_1-\xi_2|.
\end{aligned}
$$

Since P is proportional to d, there exists a constant L_0 that only depends on ψ_1,ψ_2,ψ_3 and Ω, such that $M_d(\xi;\mathbf{q},\boldsymbol{\mu})$ is L_0-Lipschitz for all $d\in\mathbb{N}$. Then by the compactness of Ω, for any $\varepsilon>0$, there

exists a finite set $\mathcal{N}_\varepsilon(\Omega) \subset \mathbb{C}_+$, that is an ε/L_0 covering of the compact set Ω. Specifically, for any $\xi \in \Omega$, there exists a $\xi_* \in \mathcal{N}_\varepsilon(\Omega)$ such that $|\xi - \xi_*| < \varepsilon/L_0$. Therefore

$$\begin{aligned}&\sup_{\xi\in\Omega}\inf_{\xi_*\in\mathcal{N}_\varepsilon(\Omega)}|M_d(\xi;\mathbf{q},\boldsymbol{\mu})-M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})|\le\varepsilon,\\&\sup_{\xi\in\Omega}\inf_{\xi_*\in\mathcal{N}_\varepsilon(\Omega)}|\mathbb{E}M_d(\xi;\mathbf{q},\boldsymbol{\mu})-\mathbb{E}M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})|\le\varepsilon\end{aligned}\tag{IV.35}$$

for all $d \in \mathbb{N}$. Moreover, since $\mathcal{N}_\varepsilon(\Omega)$ is finite, the number of ξ_* is finite. Similar to the proof of (IV.7), we have

$$\sup_{\xi_*\in\mathcal{N}_\varepsilon(\Omega)}|M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})-\mathbb{E}M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})|=o_{\mathbb{P}}(1).$$

Now since $|M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})| \le P/(d\cdot\Im(\xi_*)) \le P/(d\cdot\inf_{\xi\in\Omega}\Im(\xi))$, $|M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})|$ is bounded by some constant. By the dominated convergence theorem, we have

$$\mathbb{E}\sup_{\xi_*\in\mathcal{N}_\varepsilon(\Omega)}|M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})-\mathbb{E}M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})|=o_d(1).\tag{IV.36}$$

Combining (IV.34), (IV.35) and (IV.36), we obtain

$$\begin{aligned}&\mathbb{E}\big[\sup_{\xi\in\Omega}\big|M_d(\xi;\mathbf{q},\boldsymbol{\mu})-m(\xi;\mathbf{q},\boldsymbol{\mu})\big|\big]\\&=\mathbb{E}\Big\{\sup_{\xi\in\Omega}\inf_{\xi_*\in\mathcal{N}_\varepsilon(\Omega)}\big|M_d(\xi;\mathbf{q},\boldsymbol{\mu})+M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})-M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})+\mathbb{E}M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})\\&\quad-\mathbb{E}M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})+\mathbb{E}M_d(\xi;\mathbf{q},\boldsymbol{\mu})-\mathbb{E}M_d(\xi;\mathbf{q},\boldsymbol{\mu})-m(\xi;\mathbf{q},\boldsymbol{\mu})\big|\Big\}\\&\le\mathbb{E}\Big\{\sup_{\xi\in\Omega}\inf_{\xi_*\in\mathcal{N}_\varepsilon(\Omega)}\big[|M_d(\xi;\mathbf{q},\boldsymbol{\mu})-M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})|+|\mathbb{E}M_d(\xi;\mathbf{q},\boldsymbol{\mu})-\mathbb{E}M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})|\\&\quad+|M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})-\mathbb{E}M_d(\xi_*;\mathbf{q},\boldsymbol{\mu})|+|\mathbb{E}M_d(\xi;\mathbf{q},\boldsymbol{\mu})-m(\xi;\mathbf{q},\boldsymbol{\mu})|\big]\Big\}\\&\le 2\varepsilon+o_d(1).\end{aligned}$$

Taking $d\to+\infty$, we have

$$\lim_{d\to+\infty}\mathbb{E}\Big[\sup_{\xi\in\Omega}\big|M_d(\xi;\mathbf{q},\boldsymbol{\mu})-m(\xi;\mathbf{q},\boldsymbol{\mu})\big|\Big]\le 2\varepsilon.$$

Therefore, taking $\varepsilon\to 0^+$ proves Conclusion 3 in Proposition A.6. The proof of Proposition A.6 is complete.

V. Proof of Proposition A.7

We first present some lemmas in Section V.1, and then complete the proof in Section V.2. Recall that we assume $\mathbf{q}\in\mathcal{Q}$ (see Definition A.3).

V.1 Preliminary lemmas

The lemma below presents some additional properties of the function $\mathbf{m}(\xi) = [m_1(\xi), m_2(\xi), m_3(\xi)]^\top$ defined in Proposition A.6.

Lemma V.1. *Let $\mathbf{m}(\xi) = [m_1(\xi), m_2(\xi), m_3(\xi)]^\top$ defined on $\xi \in \mathbb{C}_+$ be the analytic continuation of the solution of the implicit equations $\mathbf{m} = \mathsf{F}(\mathbf{m}; \xi, \mathbf{q}, \boldsymbol{\mu})$ defined in Proposition A.6. Then for any fixed $\xi_r \in \mathbb{R}$ and $j = 1, 2, 3$, we have*

$$\lim_{u \to +\infty} |m_j(\xi_r + \mathrm{i}u) \cdot (\xi_r + \mathrm{i}u) + \psi_j| = 0,$$

Proof [Proof of Lemma V.1] We denote $\xi_u = \xi_r + \mathrm{i}u$ for $u > 0$, and use the same definition of $H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu})$ as in (III.1) that

$$H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu}) = -\mu_{1,*}^2 m_3 + \frac{1}{m_1 + \frac{-\mu_{2,1}^2(1+q_1)^2 m_2 m_3 + (1+\mu_{2,1}^2 m_2 q_4)(1+m_3 q_5)}{\mu_{1,1}^2 q_4 (1+m_3 q_5) - \mu_{1,1}^2 (1+q_1)^2 m_3}}.$$

Then we have

$$\mathsf{F}_1(\mathbf{m}; \xi, \mathbf{q}, \boldsymbol{\mu}) = \frac{\psi_1}{-\xi + q_2 \mu_{1,*}^2 + H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu})}. \tag{V.1}$$

Moreover, define $\check{m}_1(\xi) = -\psi_1/\xi$, $\check{m}_2(\xi) = -\psi_2/\xi$, and $\check{m}_3 = -\psi_3/\xi$, and denote $\check{\mathbf{m}}(\xi) = [\check{m}_1(\xi), \check{m}_2(\xi), \check{m}_3(\xi)]^\top$. Then clearly we have $\lim_{u \to +\infty} \check{\mathbf{m}}(\xi_u) = \mathbf{0}$. By the definition of $H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu})$, with simple calculations, we can see that $\lim_{u \to +\infty} H_1(\check{\mathbf{m}}(\xi_u); \mathbf{q}, \boldsymbol{\mu}) = q_4 \mu_{1,1}^2$. Thus by (V.1), we have

$$|\xi_u \cdot [\check{m}_1 - \mathsf{F}_1(\check{\mathbf{m}}(\xi_u); \xi_u, \mathbf{q}, \boldsymbol{\mu})]| = \psi_1 \cdot \left| \frac{q_2 \mu_{1,*}^2 + H_1(\check{\mathbf{m}}(\xi_u); \mathbf{q}, \boldsymbol{\mu})}{\xi_u - q_2 \mu_{1,*}^2 - H_1(\check{\mathbf{m}}(\xi_u); \mathbf{q}, \boldsymbol{\mu})} \right| = O_u\left(\frac{1}{u}\right).$$

Similarly, we can show that $|\xi_u \cdot [\check{m}_j - \mathsf{F}_1(\check{\mathbf{m}}(\xi_u); \xi_u, \mathbf{q}, \boldsymbol{\mu})]| = O_u(1/u)$, $j = 2, 3$. Therefore we have

$$\xi_u \cdot ||\check{\mathbf{m}}(\xi_u) - \mathbf{F}(\check{\mathbf{m}}(\xi_u); \xi_u, \mathbf{q}, \boldsymbol{\mu})||_2 = O_u(u^{-1}). \tag{V.2}$$

Moreover, by Lemma III.1, there exists a sufficiently large ξ_0 such that for any $K \geq \xi_0$, $\mathbf{F}(\cdot; \xi_u, \mathbf{q}, \boldsymbol{\mu})$ is $1/2$-Lipschitz on the domain $\mathbb{D}(2\psi_1/\xi_0) \times \mathbb{D}(2\psi_2/\xi_0) \times \mathbb{D}(2\psi_3/\xi_0)$. Therefore for sufficiently large K,

$$\begin{aligned}
&\|\check{\mathbf{m}}(\xi_u) - \mathbf{m}(\xi_u)\|_2 \\
&= \|\mathbf{F}(\check{\mathbf{m}}(\xi_u); \xi_u, \mathbf{q}, \boldsymbol{\mu}) - \mathbf{F}(\mathbf{m}(\xi_u); \xi_u, \mathbf{q}, \boldsymbol{\mu}) + \check{\mathbf{m}}(\xi_u) - \mathbf{F}(\check{\mathbf{m}}(\xi_u); \xi_u, \mathbf{q}, \boldsymbol{\mu})\|_2 \\
&\leq \|\mathbf{F}(\check{\mathbf{m}}(\xi_u); \xi_u, \mathbf{q}, \boldsymbol{\mu}) - \mathbf{F}(\mathbf{m}(\xi_u); \xi_u, \mathbf{q}, \boldsymbol{\mu})\|_2 + \|\check{\mathbf{m}}(\xi_u) - \mathbf{F}(\check{\mathbf{m}}(\xi_u); \xi_u, \mathbf{q}, \boldsymbol{\mu})\|_2 \\
&\leq \|\check{\mathbf{m}}(\xi_u) - \mathbf{m}(\xi_u)\|_2/2 + \|\check{\mathbf{m}}(\xi_u) - \mathbf{F}(\check{\mathbf{m}}(\xi_u); \xi_u, \mathbf{q}, \boldsymbol{\mu})\|_2,
\end{aligned}$$

where the first equality follows by the definition of $\mathbf{m}(\xi_u)$ as the fixed point of $\mathbf{F}(\cdot; \xi_u, \mathbf{q}, \boldsymbol{\mu})$, the first inequality follows by triangle inequality, and the second inequality follows by the $1/2$-Lipschitz continuity of $\mathbf{F}(\cdot; \xi_u, \mathbf{q}, \boldsymbol{\mu})$ on the domain $\mathbb{D}(2\psi_1/\xi_0) \times \mathbb{D}(2\psi_2/\xi_0) \times \mathbb{D}(2\psi_3/\xi_0)$ (note that $\mathbf{m}(\xi_u)$ is automatically in this domain according to Lemma III.1, and $\check{\mathbf{m}}(\xi_u)$ is also in this domain by its

definition). Rearranging terms then gives

$$\|\check{\mathbf{m}}(\xi_u) - \mathbf{m}(\xi_u)\|_2 \le 2\|\check{\mathbf{m}}(\xi_u) - \mathbf{F}(\check{\mathbf{m}}(\xi_u); \xi_u, \mathbf{q}, \boldsymbol{\mu})\|_2. \tag{V.3}$$

Thus for $j = 1, 2, 3$, we have

$$\begin{aligned} |m_j(\xi_r + \mathrm{i}u) \cdot (\xi_r + \mathrm{i}u) + \psi_j| &= \xi_u \cdot |m_j(\xi_u) - \check{m}_j(\xi_u)| \\ &\le 2\xi_u \cdot \|\check{\mathbf{m}}(\xi_u) - \mathbf{F}(\check{\mathbf{m}}(\xi_u); \xi_u, \mathbf{q}, \boldsymbol{\mu})\|_2 \\ &= O_u(u^{-1}), \end{aligned}$$

where the first inequality follows by (V.3), and the second equality follows by (V.2). This completes the proof. ■

The following lemma shows the asymptotics of the functions $G_d(\mathrm{i}u; \mathbf{q}, \boldsymbol{\mu})$ and $g(\mathrm{i}u; \mathbf{q}, \boldsymbol{\mu})$ (defined in Definition A.3 and (A.3) respectively) as u goes to infinity.

Lemma V.2. *Let $G_d(\xi; \mathbf{q}, \boldsymbol{\mu})$ be defined in Definition A.3 and $g(\xi; \mathbf{q}, \boldsymbol{\mu})$ defined in (A.3). The following limits hold:*

$$\begin{aligned} &\lim_{u \to +\infty} \sup_{d \ge 1} \mathbb{E}|G_d(\mathrm{i}u; \mathbf{q}, \boldsymbol{\mu}) - (\psi_1 + \psi_2 + \psi_3)\log(-\mathrm{i}u)| = 0, \\ &\lim_{u \to +\infty} |g(\mathrm{i}u; \mathbf{q}, \boldsymbol{\mu}) - (\psi_1 + \psi_2 + \psi_3)\log(-\mathrm{i}u)| = 0. \end{aligned}$$

Proof [Proof of Lemma V.2] The real and imaginary parts of $G_d(\mathrm{i}u; \mathbf{q}, \boldsymbol{\mu}) - (\psi_1 + \psi_2 + \psi_3)\log(-\mathrm{i}u)$ are

$$\begin{aligned} \Big|\Re\Big[\frac{1}{P}\sum\big(\log(\lambda_i(\mathbf{A}) - \mathrm{i}u) - \log(-\mathrm{i}u)\big)\Big]\Big| &= \frac{1}{2P}\sum_{i=1}^{P}\log(1 + \lambda_i(A)^2/u^2) \le \frac{\|\mathbf{A}\|_F^2}{2Pu^2}, \\ \Big|\Im\Big[\frac{1}{P}\sum\big(\log(\lambda_i(\mathbf{A}) - \mathrm{i}u) - \log(-\mathrm{i}u)\big)\Big]\Big| &= \frac{1}{P}\sum_{i=1}^{P}\arctan(\lambda_i(\mathbf{A})/u) \le \frac{\|A\|_F}{P^{1/2}u}. \end{aligned}$$

By the definition of the linear pencil matrix $\mathbf{A}$, it is easy to see that $\frac{1}{P}\mathbb{E}[\|\mathbf{A}\|_F^2] = O_d(1)$, thus

$$\lim_{u \to +\infty} \sup_{d \ge 1} \mathbb{E}|G_d(\mathrm{i}u; \mathbf{q}, \boldsymbol{\mu}) - (\psi_1 + \psi_2 + \psi_3)\log(-\mathrm{i}u)| = 0.$$

For the asymptotics of $g(\mathrm{i}u; \mathbf{q}, \boldsymbol{\mu})$, note that

$$L(\xi, z_1, z_2, z_3; \mathbf{q}, \boldsymbol{\mu}) = L_1(z_1, z_2, z_3; \mathbf{q}, \boldsymbol{\mu}) + L_2(\xi, z_1, z_2, z_3; \mathbf{q}, \boldsymbol{\mu}),$$

where

$$\begin{aligned} &L_1(z_1, z_2, z_3; \mathbf{q}, \boldsymbol{\mu}) \\ &\quad = \log\left[(1 + \mu_{1,1}^2 z_1 q_4 + \mu_{2,1}^2 z_2 q_4)(1 + z_3 q_5) - \mu_{1,1}^2(1 + q_1)^2 z_1 z_3 - \mu_{2,1}^2(1 + q_1)^2 z_2 z_3\right] \\ &\qquad - \mu_{1,*}^2 z_1 z_3 - \mu_{2,*}^2 z_2 z_3 + q_2\mu_{1,*}^2 z_1 + q_2\mu_{1,*}^2 z_2 + q_3 z_3, \\ &L_2(\xi, z_1, z_2, z_3; \mathbf{q}, \boldsymbol{\mu}) \\ &\quad = -\psi_1\log(z_1/\psi_1) - \psi_2\log(z_2/\psi_2) - \psi_3\log(z_3/\psi_3) - \xi(z_1 + z_2 + z_3) - \psi_1 - \psi_2 - \psi_3. \end{aligned}$$

We now calculate the limits of $L_1(m_1(\mathrm{i}u), m_2(\mathrm{i}u), m_3(\mathrm{i}u); \mathbf{q}, \boldsymbol{\mu})$ and $L_2(\mathrm{i}u, m_1(\mathrm{i}u), m_2(\mathrm{i}u), m_3(\mathrm{i}u); \mathbf{q}, \boldsymbol{\mu})$ separately. For L_1, by Lemma V.1, we have

$$\lim_{u\to+\infty} m_1(\mathrm{i}u) = 0, \quad \lim_{u\to+\infty} m_2(\mathrm{i}u) = 0, \quad \lim_{u\to+\infty} m_3(\mathrm{i}u) = 0,$$

which immediately implies that

$$\lim_{u\to+\infty} L_1(m_1(\mathrm{i}u), m_2(\mathrm{i}u), m_3(\mathrm{i}u); \mathbf{q}, \boldsymbol{\mu}) = 0.$$

For L_2, note that by Lemma V.1 we also have

$$\lim_{u\to+\infty} |m_1(\mathrm{i}u)\mathrm{i}u + \psi_1| = 0, \quad \lim_{u\to+\infty} |m_2(\mathrm{i}u)\mathrm{i}u + \psi_2| = 0, \quad \lim_{u\to+\infty} |m_3(\mathrm{i}u)\mathrm{i}u + \psi_3| = 0.$$

Therefore,

$$\begin{aligned}
&|L_2(\mathrm{i}u, m_1(\mathrm{i}u), m_2(\mathrm{i}u), m_3(\mathrm{i}u); \mathbf{q}, \boldsymbol{\mu}) - (\psi_1+\psi_2+\psi_3)\log(-\mathrm{i}u)| \\
\le&\psi_1|\log(-\mathrm{i}um_1(\mathrm{i}u)/\psi_1)| + \psi_2|\log(-\mathrm{i}um_2(\mathrm{i}u)/\psi_2)| + \psi_3|\log(-\mathrm{i}um_3(\mathrm{i}u)/\psi_3)| \\
&+ |\psi_1 + \mathrm{i}um_1(\mathrm{i}u)| + |\psi_2 + \mathrm{i}um_2(\mathrm{i}u)| + |\psi_3 + \mathrm{i}um_3(\mathrm{i}u)| \to 0,
\end{aligned}$$

which completes the proof. ■

The following lemma gives an important identity between $g(\xi; \mathbf{q}, \boldsymbol{\mu})$ and $m(\xi; \mathbf{q}, \boldsymbol{\mu})$.

Lemma V.3. *For all $\xi \in \mathbb{C}_+$, it holds that*

$$\frac{\partial g}{\partial \xi}(\xi; \mathbf{q}, \boldsymbol{\mu}) = -(m_1+m_2+m_3)(\xi; \mathbf{q}, \boldsymbol{\mu}) = -m(\xi; \mathbf{q}, \boldsymbol{\mu}).$$

Proof [Proof of Lemma V.3] By the definition of $L(\xi, z_1, z_2, z_3; \mathbf{q}, \boldsymbol{\mu})$, it is easy to see that

$$\begin{aligned}
&\partial_{z_1} L(\xi, z_1, z_2, z_3; \mathbf{q}, \boldsymbol{\mu}) \\
&\quad= -\mu_{1,*}^2 z_3 + q_2\mu_{1,*}^2 - \psi_1/z_1 - \xi \\
&\qquad+ \frac{\mu_{1,1}^2 q_4(1+z_3q_5) - \mu_{1,1}^2(1+q_1)^2 z_3}{(1+\mu_{1,1}^2 z_1 q_4 + \mu_{2,1}^2 z_2 q_4)(1+z_3q_5) - \mu_{2,1}^2(1+q_1)^2 z_2 z_3 - \mu_{1,1}^2(1+q_1)^2 z_1 z_3} \\
&\quad= \psi_1\left(\frac{1}{\mathsf{F}_1(\mathbf{z})} - \frac{1}{z_1}\right),
\end{aligned}$$

$$\begin{aligned}
&\partial_{z_2} L(\xi, z_1, z_2, z_3; \mathbf{q}, \boldsymbol{\mu}) \\
&\quad= -\mu_{2,*}^2 z_3 + q_2\mu_{2,*}^2 - \psi_2/z_2 - \xi \\
&\qquad+ \frac{\mu_{2,1}^2 q_4(1+z_3q_5) - \mu_{2,1}^2(1+q_1)^2 z_3}{(1+\mu_{2,1}^2 z_2 q_4 + \mu_{1,1}^2 z_1 q_4)(1+z_3q_5) - \mu_{1,1}^2(1+q_1)^2 z_1 z_3 - \mu_{2,1}^2(1+q_1)^2 z_2 z_3} \\
&\quad= \psi_2\left(\frac{1}{\mathsf{F}_2(\mathbf{z})} - \frac{1}{z_2}\right),
\end{aligned}$$

$$
\begin{aligned}
&\partial_{z_3} L(\xi, z_1, z_2, z_3; \mathbf{q}, \boldsymbol{\mu}) \\
&\quad = -\mu_{1,*}^2 z_1 - \mu_{2,*}^2 z_2 + q_3 - \psi_3/z_3 - \xi \\
&\qquad + \frac{q_5(1+\mu_{1,1}^2 z_1 q_4 + \mu_{2,1}^2 z_2 q_4) - \mu_{2,1}^2(1+q_1)^2 z_2 - \mu_{1,1}^2(1+q_1)^2 z_1}{(1+\mu_{2,1}^2 z_2 q_4 + \mu_{1,1}^2 z_1 q_4)(1+z_3 q_5) - \mu_{1,1}^2(1+q_1)^2 z_1 z_3 - \mu_{2,1}^2(1+q_1)^2 z_2 z_3} \\
&\quad = \psi_3 \bigg(\frac{1}{\mathsf{F}_3(\mathbf{z})} - \frac{1}{z_3} \bigg),
\end{aligned}
$$

where we utilize the definition of $\mathbf{F}$ in Definition A.5 and write $\mathbf{z} = [z_1, z_2, z_3]$. Then by Proposition A.6, we have

$$
\nabla_{\mathbf{z}} L(\xi, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})|_{\mathbf{z}=\mathbf{m}} \equiv \mathbf{0}
$$

for all $\xi \in \mathbb{C}_+$. By the formula of implicit differentiation, we have

$$
\begin{aligned}
\frac{\partial g(\xi; \mathbf{q}, \boldsymbol{\mu})}{\partial \xi} &= \big[\big\langle \nabla_{\mathbf{z}} L(\xi, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})|_{\mathbf{z}=\mathbf{m}}, \partial_\xi \mathbf{m} \big\rangle + \partial_\xi L(\xi, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})|_{\mathbf{z}=\mathbf{m}} \big] \\
&= 0 + \frac{\mathrm{d}\, L(\xi, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})}{\mathrm{d}\, \xi}\Big|_{\mathbf{z}=\mathbf{m}} = -m(\xi; \mathbf{q}, \boldsymbol{\mu}).
\end{aligned}
$$

This completes the proof of Lemma V.3. ■

The following lemma further shows that the derivatives of G_d and g_d are asymptotically bounded.

Lemma V.4. *For fixed $\xi \in \mathbb{C}_+$, the following limits hold:*

$$
\begin{aligned}
&\limsup_{d \to +\infty} \big\{ \mathbb{E} \sup_{\mathbf{q} \in \mathcal{Q}} \|\nabla_{\mathbf{q}} G_d(\xi; \mathbf{q}, \boldsymbol{\mu})\|_2 \big\} + \sup_{\mathbf{q} \in \mathcal{Q}} \|\nabla_{\mathbf{q}} g(\xi; \mathbf{q}, \boldsymbol{\mu})\|_2 < +\infty, \\
&\limsup_{d \to +\infty} \big\{ \mathbb{E} \sup_{\mathbf{q} \in \mathcal{Q}} \|\nabla_{\mathbf{q}}^2 G_d(\xi; \mathbf{q}, \boldsymbol{\mu})\|_{\mathrm{op}} \big\} + \sup_{\mathbf{q} \in \mathcal{Q}} \|\nabla_{\mathbf{q}}^2 g(\xi; \mathbf{q}, \boldsymbol{\mu})\|_{\mathrm{op}} < +\infty, \\
&\limsup_{d \to +\infty} \big\{ \mathbb{E} \sup_{\mathbf{q} \in \mathcal{Q}} \|\nabla_{\mathbf{q}}^3 G_d(\xi; \mathbf{q}, \boldsymbol{\mu})\|_{\mathrm{op}} \big\} + \sup_{\mathbf{q} \in \mathcal{Q}} \|\nabla_{\mathbf{q}}^3 g(\xi; \mathbf{q}, \boldsymbol{\mu})\|_{\mathrm{op}} < +\infty.
\end{aligned}
$$

Proof [Proof of Lemma V.4] Let $\xi = \xi_r + \mathrm{i}u$, where $\xi_r \in \mathbb{R}$ and $u \in \mathbb{R}_+$ are both fixed. We also denote

$$
\begin{aligned}
&\mathbf{S}_1 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \widetilde{\mathbf{Z}}_1^\top \\ \mathbf{0} & \mathbf{0} & \widetilde{\mathbf{Z}}_2^\top \\ \widetilde{\mathbf{Z}}_1 & \widetilde{\mathbf{Z}}_2 & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{Z}} \\ \widetilde{\mathbf{Z}} & \mathbf{0} \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} \mathbf{M}_* \mathbf{M}_* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \\
&\mathbf{S}_3 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix}, \quad \mathbf{S}_4 = \begin{bmatrix} \mathbf{M}_1 \frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\top}{d} \mathbf{M}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{S}_5 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\mathbf{X}\mathbf{X}^\top}{d} \end{bmatrix}.
\end{aligned}
$$

Then $\mathbf{S}_1, \ldots, \mathbf{S}_5$ are not related to $\mathbf{q}$, and it is easy to see that

$$
\limsup_{d \to +\infty} \mathbb{E} \sup_{\mathbf{q} \in \mathcal{Q}} \|\mathbf{S}_i\|_{\mathrm{op}}^{2k} < +\infty
$$

for any fixed $k \in \mathbb{N}$. Moreover, define $\mathbf{R} = \mathbf{R}(\mathbf{q}) = (\mathbf{A}(\mathbf{q}) - \xi_r \mathbf{I}_P - \mathrm{i}u\mathbf{I}_P)^{-1}$. Since $\mathbf{A}(\mathbf{q})$ is a real symmetric matrix, the imaginary parts in the eigenvalues of $\mathbf{A}(\mathbf{q}) - \xi_r \mathbf{I}_P - \mathrm{i}u\mathbf{I}_P$ are all $-\mathrm{i}u$, and

hence we deterministically have

$$\sup_{\mathbf{q}} \|\mathbf{R}\|_{\text{op}} \leq 1/u. \tag{V.4}$$

Therefore, by (II.2), (II.3) and the definition of the linear pencil matrix $\mathbf{A}$, we have

$$\mathbb{E} \sup_{\mathbf{q}\in\mathcal{Q}} |\partial_{q_i} G_d(\xi;\mathbf{q})| = \mathbb{E} \sup_{\mathbf{q}\in\mathcal{Q}} \frac{1}{d}|\text{tr}(\mathbf{R}\mathbf{S}_i)| \leq \mathbb{E} \sup_{\mathbf{q}\in\mathcal{Q}} \frac{1}{u}[\|\mathbf{S}_i\|_{\text{op}}] = O_d(1),$$

$$\mathbb{E} \sup_{\mathbf{q}\in\mathcal{Q}} |\partial^2_{q_i,q_j} G_d(\xi;\mathbf{q})| = \mathbb{E} \sup_{\mathbf{q}\in\mathcal{Q}} \frac{1}{d}|\text{tr}(\mathbf{R}\mathbf{S}_i\mathbf{R}\mathbf{S}_j)| \leq \left(\mathbb{E} \sup_{\mathbf{q}\in\mathcal{Q}} \frac{1}{u^2}[\|\mathbf{S}_i\|^2_{\text{op}}\|\mathbf{S}_j\|^2_{\text{op}}]\right)^{\frac{1}{2}} = O_d(1).$$

Similarly, for the third order derivatives, we also have

$$\begin{aligned}\mathbb{E} \sup_{\mathbf{q}\in\mathcal{Q}} |\partial^3_{q_i,q_j,q_l} G_d(\xi;\mathbf{q})| =& \mathbb{E}\Big\{ \sup_{\mathbf{q}\in\mathcal{Q}} \frac{1}{d}|\text{tr}(\mathbf{R}\mathbf{S}_i\mathbf{R}\mathbf{S}_j\mathbf{R}\mathbf{S}_l) + |\text{tr}(\mathbf{R}\mathbf{S}_i\mathbf{R}\mathbf{S}_l\mathbf{R}\mathbf{S}_j)|\Big\} \\ \leq& \frac{2}{u^3}\Big(\mathbb{E} \sup_{\mathbf{q}\in\mathcal{Q}} [\|\mathbf{S}_i\|^4_{\text{op}}\|\mathbf{S}_j\|^4_{\text{op}}\|\mathbf{S}_j\|^4_{\text{op}}]\Big)^{\frac{1}{4}} = O_d(1).\end{aligned}$$

This completes the Proof for $G_d(\xi;\mathbf{q},\boldsymbol{\mu})$. As for $g(\xi;\mathbf{q},\boldsymbol{\mu})$, we first show that if $\mathbf{q}_1 \neq \mathbf{q}_2$, the following property holds:

$$\begin{aligned}\frac{|m(\xi;\mathbf{q}_1,\boldsymbol{\mu}) - m(\xi;\mathbf{q}_2,\boldsymbol{\mu})|}{\|\mathbf{q}_1-\mathbf{q}_2\|_2} &= \frac{\Big|\lim\limits_{d\to\infty} \mathbb{E}\big(M_d(\xi;\mathbf{q}_1,\boldsymbol{\mu}) - M_d(\xi;\mathbf{q}_2,\boldsymbol{\mu})\big)\Big|}{\|\mathbf{q}_1-\mathbf{q}_2\|_2} \\ &= \lim_{d\to\infty} \frac{\Big|\mathbb{E}\big[\text{tr}\big(\mathbf{R}(\mathbf{q}_1) - \mathbf{R}(\mathbf{q}_2)\big)\big]\Big|}{d\|\mathbf{q}_1-\mathbf{q}_2\|_2} \\ &= \lim_{d\to\infty} \frac{\Big|\mathbb{E}\big[\text{tr}\big(\mathbf{R}(\mathbf{q}_1)\big(\mathbf{A}(\mathbf{q}_1) - \mathbf{A}(\mathbf{q}_2)\big)\mathbf{R}(\mathbf{q}_2)\big)\big]\Big|}{d\|\mathbf{q}_1-\mathbf{q}_2\|_2} \\ &\leq \lim_{d\to\infty} \frac{P}{d}\cdot \mathbb{E}\frac{\|\mathbf{A}(\mathbf{q}_1) - \mathbf{A}(\mathbf{q}_2)\|_{\text{op}}}{u^2\cdot\|\mathbf{q}_1-\mathbf{q}_2\|_2} < +\infty,\end{aligned}$$

where the first equality follows by Proposition A.6, the third equality follows by the identity $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})\mathbf{B}^{-1}$ for any invertible matrices $\mathbf{A},\mathbf{B}$, the first inequality follows by $|\text{tr}(\mathbf{A}\mathbf{B})| \leq P\cdot\|\mathbf{A}\|_{\text{op}}\|\mathbf{B}\|_{\text{op}}$ for all $\mathbf{A},\mathbf{B}\in\mathbb{C}^{P\times P}$ and (V.4), and the last inequality follows by the linearity of $\mathbf{A}(\mathbf{q})$ in $\mathbf{q}$. Therefore we have $\sup_{\mathbf{q}\in\mathcal{Q}}\|\nabla_{\mathbf{q}} m(\xi;\mathbf{q},\boldsymbol{\mu})\|_2 < +\infty$. Similarly, we can also show that $\sup_{\mathbf{q}\in\mathcal{Q}}\|\nabla^j_{\mathbf{q}} m(\xi;\mathbf{q},\boldsymbol{\mu})\|_{\text{op}} < +\infty$ for any fixed $\xi\in\mathbb{C}_+$ and $j=2,3$. Moreover, by Lemma V.3, we have

$$\frac{d}{d\xi} g(\xi;\mathbf{q},\boldsymbol{\mu}) = -m(\xi;\mathbf{q},\boldsymbol{\mu}).$$

Then $\sup_{\mathbf{q}\in\mathcal{Q}}\|\nabla^j_{\mathbf{q}} m(\xi;\mathbf{q},\boldsymbol{\mu})\|_{\text{op}} < +\infty$ indicates that $\sup_{\mathbf{q}\in\mathcal{Q}}\|\nabla^j_{\mathbf{q}} g(\xi;\mathbf{q},\boldsymbol{\mu})\|_{\text{op}} < +\infty$. This completes the proof of Lemma V.4. ■

Finally, we present a classic result which shows that the derivatives of a function in a compact region can be upper bounded by the function value and the second derivatives of the function in the region.

Lemma V.5 (lemma 11.4 in Mei and Montanari (2022)). *Let $f \in C^2([a,b])$. Then we have*

$$\sup_{x\in[a,b]} |f'(x)| \le \Big|\frac{f(a)-f(b)}{a-b}\Big| + \frac{1}{2}\sup_{x\in[a,b]} |f''(x)|\cdot|a-b|.$$

Moreover, letting, $f \in C^2(\boldsymbol{B}(\mathbf{x}_0, 2r))$ where $\boldsymbol{B}(\mathbf{x}_0, 2r) = \{\mathbf{x}\in\mathbb{R}^d : \|\mathbf{x}-\mathbf{x}_0\|_2 \le r\}$ with a point $\mathbf{x}_0$, we have

$$\sup_{\mathbf{x}\in\boldsymbol{B}(\mathbf{x}_0,2r)} \|\nabla f(\mathbf{x})\|_2 \le r^{-1}\sup_{\mathbf{x}\in\boldsymbol{B}(\mathbf{x}_0,2r)} |f(\mathbf{x})| + 2r\sup_{\mathbf{x}\in\boldsymbol{B}(\mathbf{x}_0,2r)} \|\nabla^2 f(\mathbf{x})\|_{\mathrm{op}}.$$

V.2 Completion of the proof

By Lemma V.3, we have $\frac{\partial g}{\partial \xi}(\xi;\mathbf{q},\boldsymbol{\mu}) = -m(\xi;\mathbf{q},\boldsymbol{\mu})$. Hence, for $\xi\in\mathbb{C}_+$, $u\in\mathbb{R}_+$, and any compact continuous path $c(\xi,\mathrm{i}u)$ connecting ξ and $\mathrm{i}u$, we have

$$g(\xi;\mathbf{q},\boldsymbol{\mu}) - g(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu}) = \int_{c(\xi,\mathrm{i}u)} m(x;\mathbf{q},\boldsymbol{\mu})dx.$$

Moreover, from Definition A.3, we also have $\frac{\mathrm{d}\,G_d(\xi;\mathbf{q},\boldsymbol{\mu})}{\mathrm{d}\,\xi} = -M_d(\xi;\mathbf{q},\boldsymbol{\mu})$, and

$$G_d(\xi;\mathbf{q},\boldsymbol{\mu}) - G_d(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu}) = \int_{c(\xi,\mathrm{i}u)} M_d(x;\mathbf{q},\boldsymbol{\mu})dx.$$

The two equations imply that

$$\begin{aligned}&\mathbb{E}[|G_d(\xi;\mathbf{q},\boldsymbol{\mu}) - g(\xi;\mathbf{q},\boldsymbol{\mu})|]\\ &\quad\le \mathbb{E}|G_d(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu}) - g(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})| + \int_{c(\xi,\mathrm{i}u)} \mathbb{E}|M_d(x;\mathbf{q},\boldsymbol{\mu}) - m(x;\mathbf{q},\boldsymbol{\mu})|dx. \qquad \text{(V.5)}\end{aligned}$$

Therefore, by taking supremum limit on both sides above and using Proposition A.6, we have

$$\begin{aligned}&\limsup_{d\to+\infty}\ \mathbb{E}[|G_d(\xi;\mathbf{q},\boldsymbol{\mu}) - g(\xi;\mathbf{q},\boldsymbol{\mu})|]\\ &\quad\le \limsup_{d\to+\infty}\mathbb{E}|G_d(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu}) - g(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})| + \limsup_{d\to+\infty}\int_{c(\xi,\mathrm{i}u)}\mathbb{E}|M_d(x;\mathbf{q},\boldsymbol{\mu}) - m(x;\mathbf{q},\boldsymbol{\mu})|dx\\ &\quad= \limsup_{d\to+\infty}\mathbb{E}|G_d(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu}) - g(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})|.\end{aligned}$$

Now the left hand side above does not depend on u. Moreover, by Lemma V.2, we have

$$\lim_{u\to+\infty}\limsup_{d\to+\infty}\mathbb{E}|G_d(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu}) - g(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})| = 0.$$

Therefore

$$\lim_{d\to+\infty}\mathbb{E}[|G_d(\xi;\mathbf{q},\boldsymbol{\mu}) - g(\xi;\mathbf{q},\boldsymbol{\mu})|] = 0, \qquad \text{(V.6)}$$

which proves the first equality in Proposition A.7.

Next, we will prove the second and third equalities in Proposition A.7. Define $V_d(\mathbf{q}) = G_d(\xi;\mathbf{q},\boldsymbol{\mu}) - g(\xi;\mathbf{q},\boldsymbol{\mu})$. Then by Lemma V.5, we have

$$\sup_{\widetilde{\mathbf{q}}\in\mathbf{B}(\mathbf{0},\varepsilon)} \|\nabla V_d(\widetilde{\mathbf{q}})\|_2 \le \frac{\sup_{\widetilde{\mathbf{q}}\in\mathbf{B}(\mathbf{0},\varepsilon)} |V_d(\widetilde{\mathbf{q}})|}{\varepsilon} + 2\varepsilon \sup_{\widetilde{\mathbf{q}}\in\mathbf{B}(\mathbf{0},\varepsilon)} \|\nabla^2 V_d(\widetilde{\mathbf{q}})\|_{\mathrm{op}}. \tag{V.7}$$

By equation (V.6), Lemma V.4 and the covering number argument (similar to Section IV.6 and the proof in Section 11.2 in Mei and Montanari (2022)), we get that $\lim_{d\to+\infty} \mathbb{E} \sup_{\widetilde{\mathbf{q}}\in\mathcal{Q}_\star} |V_d(\widetilde{\mathbf{q}})| = 0$. Again from Lemma V.4 and its proof, we already have

$$\lim_{d\to+\infty} \mathbb{E} \sup_{\widetilde{\mathbf{q}}\in\mathbf{B}(\mathbf{0},\varepsilon)} |\nabla^2 V_d(\widetilde{\mathbf{q}})| < C,$$

for some absolute value C. Therefore, by (V.7), we have

$$\lim_{d\to+\infty} \mathbb{E}[\|\partial_{\mathbf{q}} G_d(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}} - \partial_{\mathbf{q}} g(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}\|_2] \le C\varepsilon.$$

Taking $\varepsilon \to 0^+$, we have

$$\lim_{d\to+\infty} \mathbb{E}[\|\partial_{\mathbf{q}} G_d(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}} - \partial_{\mathbf{q}} g(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}\|_2] = 0.$$

This completes the proof of the second equality in Proposition A.7. The proof of the third equation in Proposition A.7 follows by a similar argument.

VI. Proof of Proposition A.8

The existence result is obtained by directly checking that $\mathbf{m}(\xi;\mathbf{0},\boldsymbol{\mu})$ satisfies the two properties stated in Proposition A.8 as follows. The second property follows by the original definition of $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$ on $\{\xi : \Im(\xi) \ge \xi_0\}$. For the first property, by Proposition A.6, the analytic continuation of $\mathbf{m}$ satisfies $\mathbf{m}(\xi,\mathbf{0},\boldsymbol{\mu}) \equiv \mathbf{F}[\mathbf{m}(\xi,\mathbf{0},\boldsymbol{\mu});\xi,\mathbf{0},\boldsymbol{\mu}]$ for all $\xi \in \mathbb{C}_+$. This directly implies that $\mathbf{m}(\xi,\mathbf{0},\boldsymbol{\mu})$ solves the system (3.1) for all $\xi \in \mathbb{C}_+$, which verifies the first property in Proposition A.8. Moreover, by Proposition A.6, $\mathbf{m}(\xi;\mathbf{0},\boldsymbol{\mu})$ is analytic, and $\mathbf{m}(\xi;\mathbf{0},\boldsymbol{\mu}) \in \mathbb{C}_+^3$ for all $\xi \in \mathbb{C}_+$. Therefore $\mathbf{m}(\xi;\mathbf{0},\boldsymbol{\mu})$ is indeed an analytic function from $\mathbb{C}_+$ to $\mathbb{C}_+^3$. This completes the proof of existence.

To show the uniqueness, suppose that an analytic function $\boldsymbol{\nu} : \mathbb{C}_+ \to \mathbb{C}_+^3$ satisfies the properties satisfied in Proposition A.6. It then suffices to show that $\boldsymbol{\nu}(\xi;\boldsymbol{\mu}) \equiv \mathbf{m}(\xi;\mathbf{0},\boldsymbol{\mu})$. By Lemma III.1, we clearly have $\boldsymbol{\nu}(\xi;\boldsymbol{\mu}) = \mathbf{m}(\xi;\mathbf{0},\boldsymbol{\mu})$ for all ξ with $\Im(\xi) > \xi_0$. Now since both $\boldsymbol{\nu}(\xi;\boldsymbol{\mu})$ and $\mathbf{m}(\xi;\mathbf{0},\boldsymbol{\mu})$ are analytic on $\mathbb{C}_+$, the result follows by the uniqueness of analytic continuation.

We denote by $\boldsymbol{\nu}^* = \boldsymbol{\nu}(\sqrt{\lambda}\cdot\mathrm{i};\boldsymbol{\mu})$. From the definition of $\overline{\mathbf{m}}_d(\xi)$ in Lemma IV.5, we easily get that the elements in $\overline{\mathbf{m}}_d(\xi)$ are purely imaginary in the upper half-plane of $\mathbb{C}$ when $\mathbf{q} = 0$ and $\xi = \sqrt{\lambda}\cdot\mathrm{i}$. (IV.23) further indicates that elements in $\boldsymbol{\nu}^*$ are purely imaginary. Based on the proof above, we have $\nu_j^*/\mathrm{i} \in \mathbb{R}_+$.

VII. Proofs of Lemmas and Propositions in Appendix C

VII.1 Proof of Lemma C.6

When $\Im(\xi) \geq \xi_0$ for some sufficiently large ξ_0, we prove the existence and uniqueness of the solution by the Banach fixed point theorem. To do so, we want to show that

1. $\mathsf{F}(\cdot; \mathbf{q}, \boldsymbol{\mu})$ maps domain $\mathbb{D}(2\psi_1/\xi_0) \times \cdots \times \mathbb{D}(2\psi_K/\xi_0) \times \mathbb{D}(2\psi_{K+1}/\xi_0)$ into itself.
2. $\mathsf{F}(\cdot; \mathbf{q}, \boldsymbol{\mu})$ is Lipschitz continuous with a Liptichz constant smaller than 1.

For $\mathsf{F}_1(\cdot; \mathbf{q}, \boldsymbol{\mu})$, by Definition C.5, we have

$$\mathsf{F}_1(\mathbf{m}; \xi, \mathbf{q}, \boldsymbol{\mu}) = \frac{\psi_1}{-\xi + q_2\mu_{1,*}^2 + H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu})},$$

where

$$H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu}) = -\mu_{1,*}^2 m_3 + \frac{1}{m_1 + \frac{-\sum_{j=2}^K \mu_{j,1}^2 (1+q_1)^2 m_j m_{K+1} + (1+\sum_{j=2}^K \mu_{j,1}^2 m_j q_4)(1+m_{K+1}q_5)}{\mu_{1,1}^2 q_4 (1+m_3 q_5) - \mu_{1,1}^2 (1+q_1)^2 m_3}} \tag{VII.1}$$

Note that $q_4, q_5 \leq (1+q_1)/2$, it is easy to see that for r_0 small enough and $\mathbf{m} \in \mathbb{D}(r_0) \times \mathbb{D}(r_0) \times \mathbb{D}(r_0)$, we have

$$|H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu})| \leq 2 + 2|q_4|\mu_{1,1}^2. \tag{VII.2}$$

Now as long as $\xi_0 \geq 4 + 4|q_4|\mu_{1,1}^2$, it is clear that for ξ with $\Im(\xi) \geq \xi_0$ we have

$$\Im(\xi) \geq \xi_0/2 + \xi_0/2 \geq \xi_0/2 + 2 + 2|q_4|\mu_{1,1}^2 \geq \xi_0/2 + |H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu})|, \tag{VII.3}$$

where the last inequality follows by (VII.2). Therefore we have

$$\begin{aligned} |\mathsf{F}_1(\mathbf{m}; \xi, \mathbf{q}, \boldsymbol{\mu})| \leq & \frac{\psi_1}{|\Im\big(\xi - q_2\mu_{1,*}^2 - H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu})\big)|} \\ \leq & \frac{\psi_1}{\Im(\xi) - |H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu})|} \leq \frac{2\psi_1}{\xi_0}, \end{aligned}$$

where the inequalities follow from (VII.3).

Similarly, for F_j, $j = 2, \cdots, K+1$, we also have $|\mathsf{F}_j(\mathbf{m}; \xi, \mathbf{q}, \boldsymbol{\mu})| \leq 2\psi_j/\xi_0$ provided $\xi_0 \geq 4 + 4\max_j\{|q_4|\mu_{j,1}^2, |q_5|\}$. Therefore if ξ_0 satisfies $2\max\{\psi_1, \cdots, \psi_{K+1}\}/\xi_0 \leq r_0$ and $\xi_0 \geq 4 + 4\max_j\{|q_4|\mu_{j,1}^2, |q_5|\}$, it is clear that F maps domain $\mathbb{D}(2\psi_1/\xi_0) \times \cdots \times \mathbb{D}(2\psi_{K+1}/\xi_0)$ into itself.

As for the Lipschitz continuity of $\mathsf{F}(\cdot; \mathbf{q}, \boldsymbol{\mu})$, note that

$$\nabla_{\mathbf{m}} \mathsf{F}_1(\mathbf{m}; \xi, \mathbf{q}, \boldsymbol{\mu}) = -\frac{\psi_1}{(-\xi + q_2\mu_{1,*}^2 + H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu}))^2} \cdot \nabla_{\mathbf{m}} H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu}).$$

It is easy to see that when ξ_0 is sufficiently large, $\|\nabla_{\mathbf{m}} H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu})\|_2 \leq C(\mathbf{q}, \boldsymbol{\mu})$ for all $\mathbf{m} \in \mathbb{D}(2\psi_1/\xi_0) \times \cdots \times \mathbb{D}(2\psi_{K+1}/\xi_0)$, where $C(\mathbf{q}, \boldsymbol{\mu})$ is a constant that only depends on $\mathbf{q}$ and $\boldsymbol{\mu}$. Thus when ξ_0 is sufficiently large, for ξ with $\Im(\xi) \geq \xi_0$,

$$\big\|\nabla_{\mathbf{m}} \mathsf{F}_1(\mathbf{m}; \xi, \mathbf{q}, \boldsymbol{\mu})\big\|_2 \leq \frac{C(\mathbf{q}, \boldsymbol{\mu}) \cdot \psi_1}{\Im(\xi) - |H_1(\mathbf{m}; \mathbf{q}, \boldsymbol{\mu})|} \leq \frac{4C(\mathbf{q}, \boldsymbol{\mu}) \cdot \psi_1}{\xi_0} \leq \frac{1}{4K},$$

where we again utilize (VII.3). We can apply the same argument for $\mathsf{F}_2, \ldots, \mathsf{F}_{K+1}$, and conclude that $\mathbf{F}$ is $\frac{1}{2}$-Lipschitz on $\mathbf{m} \in \mathbb{D}(2\psi_1/\xi_0) \times \cdots \times \mathbb{D}(2\psi_{K+1}/\xi_0)$. Therefore by Banach fixed point theorem, there exists a unique fixed point of $\mathbf{F}$. Thus the fixed point of the functions defined in Definition C.5 exists and is unique.

VII.2 Proof of Proposition C.7

Following the same argument as in Lemma IV.3, we may assume that all the elements in $\overline{\mathbf{X}}$ and $\overline{\mathbf{\Theta}}_j$ are independently generated from standard normal $\mathrm{N}(0,1)$, and the activation functions are polynomials and centralized as $\phi_j(x) = \sigma_j(x) - \mu_{j,0}$. The linear pencil matrix of this Gaussian version is defined as

$$\overline{\mathbf{A}}(\mathbf{q}, \boldsymbol{\mu}) = \begin{bmatrix} q_2\mu_{1,*}^2\mathbf{I}_{N_1} + q_4\mu_{1,1}^2\frac{\overline{\mathbf{\Theta}}_1\overline{\mathbf{\Theta}}_1^\top}{d} & \cdots & q_4\mu_{1,1}\mu_{K,1}\frac{\overline{\mathbf{\Theta}}_1\overline{\mathbf{\Theta}}_K^\top}{d} & \overline{\mathbf{Z}}_1^\top \\ \vdots & \ddots & \vdots & \vdots \\ q_4\mu_{K,1}\mu_{1,1}\frac{\overline{\mathbf{\Theta}}_K\overline{\mathbf{\Theta}}_1^\top}{d} & \cdots & q_2\mu_{K,*}^2\mathbf{I}_{N_K} + q_4\mu_{K,1}^2\frac{\overline{\mathbf{\Theta}}_K\overline{\mathbf{\Theta}}_K^\top}{d} & \overline{\mathbf{Z}}_K^\top \\ \overline{\mathbf{Z}}_1 & \cdots & \overline{\mathbf{Z}}_K & q_3\mathbf{I}_n + q_5\frac{\overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top}{d} \end{bmatrix}.$$

Here $\overline{\mathbf{Z}}_j = \Phi_j\left(\overline{\mathbf{X}}\,\overline{\mathbf{\Theta}}_j^\top/\sqrt{d}\right)/\sqrt{d} \in \mathbb{R}^{n\times N_j}$, and $\Phi_j(x)$ is defined as $\Phi_j(x) = \phi_j(x) + q_1\mu_{j,1}x$. Moreover, for $j = 1, \ldots, K$ and with $G \sim \mathrm{N}(0,1)$, we denote $\phi_{j,0} \triangleq \mathbb{E}\{\Phi_j(G)\}$, $\phi_{j,1} \triangleq \mathbb{E}\{G\Phi_j(G)\}$, $\phi_{j,*} \triangleq \mathbb{E}\{\Phi_j(G)^2\} - \phi_{j,0}^2 - \phi_{j,1}^2$. It is easy to see $\phi_{j,0} = 0$, $\phi_{j,1}^2 = \mu_{j,1}^2(1+q_1)^2$, $\phi_{j,*}^2 = \mu_{j,*}^2$.

We remind readers that $\mathcal{N}_j$ is the index set of units that use the j-th activation function σ_j. Define the following terms:

$$\overline{m}_{j,d}(\xi;\mathbf{q},\boldsymbol{\mu}) = \mathbb{E}\big[\overline{M}_{j,d}(\xi;\mathbf{q},\boldsymbol{\mu})\big], \quad \overline{M}_{j,d}(\xi;\mathbf{q},\boldsymbol{\mu}) = \frac{1}{d}\mathrm{tr}_{\mathcal{N}_j}\big[\overline{\mathbf{A}}(\mathbf{q},\boldsymbol{\mu}) - \xi\mathbf{I}_P)^{-1}\big], \quad j = 1, \ldots, K$$
$$\overline{m}_{K+1,d}(\xi;\mathbf{q},\boldsymbol{\mu}) = \mathbb{E}\big[\overline{M}_{K+1,d}(\xi;\mathbf{q},\boldsymbol{\mu})\big], \quad \overline{M}_{K+1,d}(\xi;\mathbf{q},\boldsymbol{\mu}) = \frac{1}{d}\mathrm{tr}_{[N+1:P]}\big[\overline{\mathbf{A}}(\mathbf{q},\boldsymbol{\mu}) - \xi\mathbf{I}_P)^{-1}\big].$$

With the same argument as in Lemma IV.3, we obtain

$$\mathbb{E}\Big|\sum_{j=1}^{K+1}\overline{M}_{j,d}(\xi;\mathbf{q},\boldsymbol{\mu}) - M_d(\xi;\mathbf{q},\boldsymbol{\mu})\Big| = o_d(1), \quad \text{for any fixed } \xi \in \mathbb{C}_+.$$

Next, by contraction properties we have

$$\mathbb{E}\Big|\overline{M}_{j,d}(\xi;\mathbf{q},\boldsymbol{\mu}) - \overline{m}_{j,d}(\xi;\mathbf{q},\boldsymbol{\mu})\Big| = o_d(1), \quad \text{for any fixed } \xi \in \mathbb{C}_+.$$

To study $\overline{M}_d(\xi;\mathbf{q},\boldsymbol{\mu})$, which is the Stieltjes transform of the empirical eigenvalue distribution of $\overline{A}(\mathbf{q},\boldsymbol{\mu})$, it suffices to derive the resolvent equations for $\overline{m}_d(\xi;\mathbf{q},\boldsymbol{\mu})$ here. This is done by the following lemma.

Lemma VII.1. *Let $\overline{\mathbf{m}}_d(\xi) = [\overline{m}_{1,d}(\xi), \ldots, \overline{m}_{K+1,d}(\xi)]^\top$. Then for any fixed $\xi \in \mathbb{C}_+$, the following property holds:*

$$\|\overline{\mathbf{m}}_d(\xi) - \mathbf{F}(\overline{\mathbf{m}}_d(\xi))\|_2 = o_d(1).$$

Proof [Proof of Lemma VII.1] Since $\overline{\mathbf{m}}_d(\xi), \mathbf{F}(\overline{\mathbf{m}}_d(\xi)) \in \mathbb{C}^{K+1}$, Lemma VII.1 essentially contains results showing that each element of $\overline{\mathbf{m}}_d(\xi) - \mathbf{F}(\overline{\mathbf{m}}_d(\xi))$ is asymptotically zero. Since the proofs of the results are almost the same, we mainly focus on the proof of the first element $\overline{m}_{1,d}$. The proof still consists of three main steps similar to the proof of Lemma IV.5.

Step 1. We first use a leave-one-out argument to calculate $\overline{m}_{1,d}$. Let $\overline{\mathbf{A}}_{\cdot,N_1}$ be the N_1^{th} column of $\overline{\mathbf{A}}$, with the N_1^{th} entry removed. We further denote $\overline{\mathbf{B}} \in \mathbb{R}^{(P-1)\times(P-1)}$ the matrix from $\overline{\mathbf{A}}$ by removing the N_1^{th} row and N_1^{th} column. From the Schur complement formula, we get

$$\overline{m}_{1,d} = \psi_1 \mathbb{E}\left(-\xi + q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2\|\overline{\boldsymbol{\theta}}_{N_1}\|_2^2/d - \overline{\mathbf{A}}_{\cdot,N_1}^\top(\overline{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\overline{\mathbf{A}}_{\cdot,N_1}\right)^{-1}. \tag{VII.4}$$

We decompose the vectors $\overline{\boldsymbol{\theta}}_a$, $a \in [N]$ and $\overline{\mathbf{x}}_i$, $i \in [n]$ into components along the direction of $\overline{\boldsymbol{\theta}}_{N_1}$ and the other orthogonal directions:

$$\begin{aligned}
\overline{\boldsymbol{\theta}}_a &= \eta_a \frac{\overline{\boldsymbol{\theta}}_{N_1}}{\|\overline{\boldsymbol{\theta}}_{N_1}\|} + \widetilde{\boldsymbol{\theta}}_a,\ \langle \overline{\boldsymbol{\theta}}_{N_1}, \widetilde{\boldsymbol{\theta}}_a\rangle = 0, a \in [N]\backslash\{N_1\},\\
\overline{\mathbf{x}}_i &= u_i \frac{\overline{\boldsymbol{\theta}}_{N_1}}{\|\overline{\boldsymbol{\theta}}_{N_1}\|} + \widetilde{\mathbf{x}}_i,\ \langle \overline{\boldsymbol{\theta}}_{N_1}, \widetilde{\mathbf{x}}_i\rangle = 0, i \in [n].
\end{aligned} \tag{VII.5}$$

Note that for any $a \in [N]\backslash\{N_1\}$ and $i \in [n]$, η_a, u_i are standard Gaussian and are independent of $\widetilde{\boldsymbol{\theta}}_a$ and $\widetilde{\mathbf{x}}_i$. Moreover, $\widetilde{\boldsymbol{\theta}}_a$ and $\widetilde{\mathbf{x}}_i$ are conditionally independent on each other given $\overline{\boldsymbol{\theta}}_{N_1}$, with $\widetilde{\boldsymbol{\theta}}_a, \widetilde{\mathbf{x}}_i \sim N(0, P_\perp)$, where $P_\perp$ is the projector orthogonal to $\overline{\boldsymbol{\theta}}_{N_1}$. We then have $\overline{\mathbf{A}}_{\cdot,N_1} = (\overline{\mathbf{A}}_{1,N_1}, ..., \overline{\mathbf{A}}_{P-1,N_1})^\top \in \mathbb{R}^{P-1}$ with

$$\overline{\mathbf{A}}_{i,N_1} = \begin{cases} \dfrac{q_4\mu_{1,1}^2\eta_i}{d}\|\overline{\boldsymbol{\theta}}_{N_1}\|_2, & \text{if } i \in [1, N_1 - 1],\\ \dfrac{q_4\mu_{1,1}\mu_{j,1}\eta_{i+1}}{d}\|\overline{\boldsymbol{\theta}}_{N_1}\|_2, & \text{if } i+1 \in \mathcal{N}_j,\ j \geq 2,\\ \dfrac{1}{\sqrt{d}}\Phi_1\Big(\dfrac{1}{\sqrt{d}}u_{i-N+1}\|\overline{\boldsymbol{\theta}}_{N_1}\|_2\Big), & \text{if } i \geq N. \end{cases}$$

To calculate the resolvent equations, we need to further represent the matrix $\overline{\mathbf{B}}$ in (VII.4) with η_a, $\widetilde{\boldsymbol{\theta}}_a$, u_i, and $\widetilde{\mathbf{x}}_i$ for $a \in [N]\backslash\{N_1\}$ and $i \in [n]$. Below we first list some additional notations for easier reference. Write $\boldsymbol{\eta}_1 = [\eta_1, ..., \eta_{N_1-1}] \in \mathbb{R}^{N_1-1}$, $\boldsymbol{\eta}_j = (\eta_{\mathcal{N}_j}) \in \mathbb{R}^{N_j}$, $j = 2, \ldots, K$ $\boldsymbol{\eta} = [\boldsymbol{\eta}_1^\top, \ldots, \boldsymbol{\eta}_K^\top]^\top \in \mathbb{R}^{N-1}$, $\mathbf{u} = (u_1, ..., u_n)^\top \in \mathbb{R}^n$, $\widetilde{\boldsymbol{\Theta}}_1 = [\widetilde{\boldsymbol{\theta}}_1, ..., \widetilde{\boldsymbol{\theta}}_{N_1-1}]^\top$, $\widetilde{\boldsymbol{\Theta}}_j = [\widetilde{\boldsymbol{\theta}}_{\mathcal{N}_j}]^\top$,

$$\widetilde{\boldsymbol{\Theta}} = \begin{bmatrix} \widetilde{\boldsymbol{\Theta}}_1 \\ \vdots \\ \widetilde{\boldsymbol{\Theta}}_K \end{bmatrix} \in \mathbb{R}^{(N-1)\times d}, \quad \widetilde{\mathbf{M}}_1 = \begin{bmatrix} \mu_{1,1}\mathbf{I}_{N_1-1} & & \\ & \ddots & \\ & & \mu_{K,1}\mathbf{I}_{N_K} \end{bmatrix}, \quad \widetilde{\mathbf{M}}_* = \begin{bmatrix} \mu_{1,*}\mathbf{I}_{N_1-1} & & \\ & \ddots & \\ & & \mu_{K,*}\mathbf{I}_{N_K} \end{bmatrix}.$$

With $\overline{\mathbf{B}}$ defined previously and (VII.5), $\overline{\mathbf{B}}_{[1:N-1],[1:N-1]}$ is decomposed into

$$\overline{\mathbf{B}}_{[1:N-1],[1:N-1]} = q_2\widetilde{\mathbf{M}}_*\widetilde{\mathbf{M}}_* + \frac{q_4}{d}\widetilde{\mathbf{M}}_1\widetilde{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Theta}}^\top\widetilde{\mathbf{M}}_1 + \frac{q_4}{d}\widetilde{\mathbf{M}}_1\boldsymbol{\eta}\boldsymbol{\eta}^\top\widetilde{\mathbf{M}}_1. \tag{VII.6}$$

Moreover, for $i, j \in [n]$ and $a \in [N]\backslash\{N_1\}$, we define

$$\big(\widetilde{\mathbf{H}}\big)_{ij} = \frac{1}{d}\langle \widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}_j\rangle.$$

Then we could decompose $\overline{\mathbf{B}}_{[N:P-1],[N:P-1]}$ into

$$\overline{\mathbf{B}}_{[N:P-1],[N:P-1]} = q_3\mathbf{I}_n + q_5\widetilde{\mathbf{H}} + \frac{q_5}{d}\mathbf{u}\mathbf{u}^\top. \tag{VII.7}$$

$\overline{\mathbf{B}}_{[N:P-1],[1:N-1]} = \overline{\mathbf{B}}^\top_{[1:N-1],[N:P-1]}$ holds due to the symmetry of $\overline{\mathbf{B}}$. For $i, j \in [n]$ and $a \in \mathcal{N}_j \backslash \{N_1\}$, elementally we have

$$\begin{aligned}(\overline{\mathbf{Z}})_{i,a} &= \frac{1}{\sqrt{d}}\Phi_j\Big(\frac{1}{\sqrt{d}}\langle \overline{\mathbf{x}}_i, \overline{\boldsymbol{\theta}}_a\rangle\Big) = \frac{1}{\sqrt{d}}\Phi_j\Big(\frac{1}{\sqrt{d}}\langle \widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a\rangle + \frac{1}{d}u_i\eta_a\Big)\\ &= \frac{1}{\sqrt{d}}\Phi_j\Big(\frac{1}{\sqrt{d}}\langle \widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a\rangle\Big) + \frac{\phi_{j,1}}{d}u_i\eta_a + \frac{1}{\sqrt{d}}\Big[\Phi_{j,\perp}\Big(\frac{1}{\sqrt{d}}\langle \widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a\rangle + \frac{1}{\sqrt{d}}u_i\eta_a\Big) - \Phi_{j,\perp}\Big(\frac{1}{\sqrt{d}}\langle \widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a\rangle\Big)\Big],\end{aligned}$$

where $\Phi_{j,\perp}(x) = \Phi_j(x) - \phi_{j,1}x$. By the symmetry of $\overline{\mathbf{B}}$, we can then decompose $\overline{\mathbf{B}}_{[N:P-1],[1:N-1]}$ into

$$\overline{\mathbf{B}}_{[N:P-1],[1:N-1]} = \widetilde{\mathbf{Z}} + \frac{1}{d}\mathbf{u}\boldsymbol{\eta}\mathbf{M}_\phi + [\mathbf{E}_1, \mathbf{E}_2]. \tag{VII.8}$$

Here, we define

$$\widetilde{\mathbf{Z}} = [\widetilde{\mathbf{Z}}_1, \ldots, \widetilde{\mathbf{Z}}_K], \quad (\widetilde{\mathbf{Z}}_j)_{i,a} = \frac{1}{\sqrt{d}}\Phi_j\Big(\frac{1}{\sqrt{d}}\langle \widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a\rangle\Big), \quad \mathbf{M}_\phi = \begin{bmatrix}\phi_{1,1}\mathbf{I}_{N_1-1} & & \\ & \ddots & \\ & & \phi_{K,1}\mathbf{I}_{N_K}\end{bmatrix}$$

$$(\mathbf{E}_j)_{i,a} = \frac{1}{\sqrt{d}}\Big[\Phi_{j,\perp}\Big(\frac{1}{\sqrt{d}}\langle \widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a\rangle + \frac{1}{\sqrt{d}}u_i\eta_a\Big) - \Phi_{j,\perp}\Big(\frac{1}{\sqrt{d}}\langle \widetilde{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}}_a\rangle\Big)\Big].$$

Combined (VII.6), (VII.7) and (VII.8), we decompose $\overline{\mathbf{B}}$ into

$$\overline{\mathbf{B}} = \widetilde{\mathbf{B}} + \boldsymbol{\Delta} + \mathbf{E} \in \mathbb{R}^{(P-1)\times(P-1)},$$

where

$$\begin{aligned}\widetilde{\mathbf{B}} &= \begin{bmatrix} q_2\widetilde{\mathbf{M}}_*\widetilde{\mathbf{M}}_* + \frac{q_4}{d}\widetilde{\mathbf{M}}_1\widetilde{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Theta}}^\top\widetilde{\mathbf{M}}_1 & \widetilde{\mathbf{Z}}^\top \\ \widetilde{\mathbf{Z}} & q_3\mathbf{I}_n + q_5\widetilde{\mathbf{H}}\end{bmatrix}\\ &= \begin{bmatrix} q_2\mu_{1,*}^2\mathbf{I}_{N_1} + q_4\mu_{1,1}^2\frac{\widetilde{\boldsymbol{\Theta}}_1\widetilde{\boldsymbol{\Theta}}_1^\top}{d} & \cdots & q_4\mu_{1,1}\mu_{K,1}\frac{\widetilde{\boldsymbol{\Theta}}_1\widetilde{\boldsymbol{\Theta}}_K^\top}{d} & \widetilde{\mathbf{Z}}_1^\top \\ \vdots & \ddots & \vdots & \vdots \\ q_4\mu_{K,1}\mu_{1,1}\frac{\widetilde{\boldsymbol{\Theta}}_K\widetilde{\boldsymbol{\Theta}}_1^\top}{d} & \cdots & q_2\mu_{K,*}^2\mathbf{I}_{N_K} + q_4\mu_{K,1}^2\frac{\widetilde{\boldsymbol{\Theta}}_K\widetilde{\boldsymbol{\Theta}}_K^\top}{d} & \widetilde{\mathbf{Z}}_K^\top \\ \widetilde{\mathbf{Z}}_1 & \cdots & \widetilde{\mathbf{Z}}_K & q_3\mathbf{I}_n + q_5\widetilde{\mathbf{H}}\end{bmatrix},\\ \boldsymbol{\Delta} &= \begin{bmatrix}\frac{q_4}{d}\widetilde{\mathbf{M}}_1\boldsymbol{\eta}\boldsymbol{\eta}^\top\widetilde{\mathbf{M}}_1 & \frac{1}{d}\mathbf{M}_\phi\boldsymbol{\eta}\mathbf{u}^\top \\ \frac{1}{d}\mathbf{u}\boldsymbol{\eta}\mathbf{M}_\phi & \frac{q_5}{d}\mathbf{u}\mathbf{u}^\top\end{bmatrix}\\ &= \begin{bmatrix}\frac{q_4\mu_{1,1}^2}{d}\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top & \cdots & \frac{q_4\mu_{1,1}\mu_{K,1}}{d}\boldsymbol{\eta}_1\boldsymbol{\eta}_K^\top & \frac{\phi_{1,1}}{d}\boldsymbol{\eta}_1\mathbf{u}^\top \\ \vdots & \ddots & \vdots & \vdots \\ \frac{q_4\mu_{K,1}\mu_{1,1}}{d}\boldsymbol{\eta}_K\boldsymbol{\eta}_1^\top & \cdots & \frac{q_4\mu_{K,1}^2}{d}\boldsymbol{\eta}_K\boldsymbol{\eta}_K^\top & \frac{\phi_{K,1}}{d}\boldsymbol{\eta}_K\mathbf{u}^\top \\ \frac{\phi_{1,1}}{d}\mathbf{u}\boldsymbol{\eta}_1^\top & \cdots & \frac{\phi_{K,1}}{d}\mathbf{u}\boldsymbol{\eta}_K^\top & \frac{q_5}{d}\mathbf{u}\mathbf{u}^\top\end{bmatrix}, \quad \mathbf{E} = \begin{bmatrix}\mathbf{0} & \cdots & \mathbf{0} & \mathbf{E}_1^\top \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{E}_K^\top \\ \mathbf{E}_1 & \cdots & \mathbf{E}_K & \mathbf{0}\end{bmatrix}.\end{aligned}$$

Clearly, by the definition of $\widetilde{\mathbf{B}}$, the Stieltjes transform corresponding to $\widetilde{\mathbf{B}}$ shares the same asymptotics as the Stieltjes transform corresponding to $\overline{\mathbf{A}}$.

Step 2. Define $w_2 = \Big(-\xi + q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2 - \overline{\mathbf{A}}_{\cdot,N_1}^\top(\widetilde{\mathbf{B}} + \mathbf{\Delta} - \xi\mathbf{I}_{P-1})^{-1}\overline{\mathbf{A}}_{\cdot,N_1}\Big)^{-1}$. Similar to the argument in Section IV.2, we have $\overline{m}_{1,d} = \psi_1\mathbb{E}w_2 + o_d(1)$.

Step 3. We calculate $\mathbb{E}w_2$ by mathematical induction. Similar to Section IV.2, we give some notations which will be used in the following calculation on $\mathbb{E}w_2$. Let

$$\mathbf{v} = \overline{\mathbf{A}}_{\cdot,N_1}, \quad \mathbf{v}_i = \overline{\mathbf{A}}_{i,N_1} = \begin{cases} \dfrac{q_4\mu_{1,1}^2\eta_i}{d}\|\overline{\boldsymbol{\theta}}_{N_1}\|_2, & \text{if } i \in [1, N_1 - 1], \\ \dfrac{q_4\mu_{1,1}\mu_{j,1}\eta_{i+1}}{d}\|\overline{\boldsymbol{\theta}}_{N_1}\|_2, & \text{if } i \in \mathcal{N}_j - 1,\ j \geq 2, \\ \dfrac{1}{\sqrt{d}}\Phi_1\Big(\dfrac{1}{\sqrt{d}}u_{i-N+1}\|\overline{\boldsymbol{\theta}}_{N_1}\|_2\Big), & \text{if } i \geq N, \end{cases}$$

and

$$\mathbf{U} = \frac{1}{\sqrt{d}}\begin{bmatrix} \boldsymbol{\eta}_1 & & & \\ & \boldsymbol{\eta}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{\eta}_K \\ & & & \mathbf{u} \end{bmatrix} \in \mathbb{R}^{(P-1)\times(K+1)}, \quad \mathbf{M} = \begin{bmatrix} q_4\mu_{1,1}^2 & \cdots & q_4\mu_{1,1}\mu_{K,1} & \phi_{1,1} \\ \vdots & \ddots & \vdots & \vdots \\ q_4\mu_{1,1}\mu_{K,1} & \cdots & q_4\mu_{K,1}^2 & \phi_{K,1} \\ \phi_{1,1} & \cdots & \phi_{K,1} & q_5 \end{bmatrix},$$

respectively. Then after direct calculation, we have the decomposition of $\mathbf{\Delta}$ as

$$\mathbf{\Delta} = \mathbf{U}\mathbf{M}\mathbf{U}^\top.$$

Similar to (IV.16), we again get that

$$\begin{aligned} w_2 = \Big(& -\xi + q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2 - \mathbf{v}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{v} \\ & + \mathbf{v}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{U}(\mathbf{M}^{-1} + \mathbf{U}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{U})^{-1}\mathbf{U}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{v}\Big)^{-1}. \end{aligned} \tag{VII.9}$$

To continue the calculation, we still require to study the terms $\mathbf{v}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{v}$, $\mathbf{v}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{U}$ and $\mathbf{U}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{U}$ in the denominator of (VII.9). To do so, we note that $\widetilde{\mathbf{B}}$ is independent on $\mathbf{v}$ and $\mathbf{U}$. Moreover, by the leave-one-out argument, the Stieltjes transform corresponding to $\widetilde{\mathbf{B}}$ shares the same asymptotics as the Stieltjes transform corresponding to $\overline{\mathbf{A}}$. Notice that η_i is independent on $\widetilde{\mathbf{B}}$ conditioned on $\overline{\boldsymbol{\theta}}_{N_1}$, and $\widetilde{\mathbf{B}}$ is independent on $\overline{\boldsymbol{\theta}}_{N_1}$. Similar to (IV.17)-(IV.19), we have

$$\mathbf{v}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{v} = q_4^2\mu_{1,1}^2\Big(\sum_{j=1}^K \mu_{j,1}^2\overline{m}_{j,d}\Big) + (\phi_{1,1}^2 + \phi_{1,*}^2)\overline{m}_{K+1,d} + o_{\mathbb{P}}(1), \tag{VII.10}$$

$$\mathbf{v}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{U} = \begin{bmatrix} q_4\mu_{1,1}^2\overline{m}_{1,d} & \cdots & q_4\mu_{1,1}\mu_{K,1}\overline{m}_{K,d} & \phi_{1,1}\overline{m}_{K+1,d} \end{bmatrix} + o_{\mathbb{P}}(1), \tag{VII.11}$$

$$\mathbf{U}^\top(\widetilde{\mathbf{B}} - \xi\mathbf{I}_{P-1})^{-1}\mathbf{U} = \begin{bmatrix} \overline{m}_{1,d} & & \\ & \ddots & \\ & & \overline{m}_{K+1,d} \end{bmatrix} + o_{\mathbb{P}}(1). \tag{VII.12}$$

Since $|w_2| \leq \Im(\xi)$ is deterministically bounded, by dominated convergence theorem, we have the L_1 convergence of w_2 by plugging (VII.10)-(VII.12) into (VII.9). We have

$$\overline{m}_{1,d} = \psi_1 \big\{ -\xi + q_2\mu_{1,*}^2 + q_4\mu_{1,1}^2 - \phi_{1,*}^2 \overline{m}_{K+1,d} - \boldsymbol{l}_K^\top \mathbf{M}_K^{-1} \boldsymbol{l}_K \big\}^{-1} + o_d(1). \tag{VII.13}$$

Here we define $\boldsymbol{l}_K = \begin{bmatrix} q_4\mu_{1,1}^2 & \cdots & q_4\mu_{1,1}\mu_{K,1} & \phi_{1,1} \end{bmatrix}^\top \in \mathbb{R}^{(K+1)\times 1}$, and

$$\mathbf{M}_K = \begin{bmatrix} q_4\mu_{1,1}^2 + \frac{1}{\overline{m}_{1,d}} & \cdots & q_4\mu_{1,1}\mu_{K,1} & \phi_{1,1} \\ \vdots & \ddots & \vdots & \vdots \\ q_4\mu_{1,1}\mu_{K,d} & \cdots & q_4\mu_{K,1}^2 + \frac{1}{\overline{m}_{K,d}} & \phi_{K,1} \\ \phi_{1,1} & \cdots & \phi_{K,1} & q_5 + \frac{1}{\overline{m}_{K+1,d}} \end{bmatrix}.$$

Note that $\phi_{j,1} = \mu_{j,1}(1+q_1)$, $\phi_{j,*} = \mu_{j,*}$, we aim to prove the following equality:

$$q_4\mu_{1,1}^2 - \boldsymbol{l}_K^\top \mathbf{M}_K^{-1} \boldsymbol{l}_K = \frac{\mu_{1,1}^2 q_4(1+q_5\overline{m}_{K+1,d}) - \mu_{1,1}^2(1+q_1)^2\overline{m}_{K+1,d}}{\Big(1 + q_4 \sum_{j=1}^K \mu_{j,1}^2 \overline{m}_{j,d}\Big)(1+q_5\overline{m}_{K+1,d}) - (1+q_1)^2 \sum_{j=1}^K \mu_{j,1}^2 \overline{m}_{j,d}\overline{m}_{K+1,d}}. \tag{VII.14}$$

We prove (VII.14) by mathematical induction. For $K = 2$, (VII.14) holds from Section IV.2. We assume that

$$q_4\mu_{1,1}^2 - \boldsymbol{l}_{K-1}^\top \mathbf{M}_{K-1}^{-1} \boldsymbol{l}_{K-1} = \frac{\mu_{1,1}^2 q_4(1+q_5\overline{m}_{K,d}) - \mu_{1,1}^2(1+q_1)^2\overline{m}_{K,d}}{\Big(1 + q_4 \sum_{j=1}^{K-1} \mu_{j,1}^2 \overline{m}_{j,d}\Big)(1+q_5\overline{m}_{K,d}) - (1+q_1)^2 \sum_{j=1}^{K-1} \mu_{j,1}^2 \overline{m}_{j,d}\overline{m}_{K,d}} \tag{VII.15}$$

holds under the case $K-1$. We aim to prove (VII.14) for general K under the assumption that (VII.15) holds. To prove so, define $\boldsymbol{\mu}_K = [\mu_{1,1}, \ldots, \mu_{K,1}]^\top$. The vector $\boldsymbol{l}_K$ could be separated into $\boldsymbol{l}_K = [q_4\mu_{1,1} \cdot \boldsymbol{\mu}_K^\top \quad (1+q_1)\mu_{1,1}]^\top$. If we further define

$$\mathbf{V}_0 = \begin{bmatrix} q_4\mu_{1,1}^2 + \frac{1}{\overline{m}_{1,d}} & \cdots & q_4\mu_{1,1}\mu_{K,1} \\ \vdots & \ddots & \vdots \\ q_4\mu_{1,1}\mu_{K,1} & \cdots & q_4\mu_{K,1}^2 + \frac{1}{\overline{m}_{K,d}} \end{bmatrix},$$

the target equation (VII.14) could be rewritten as

$$q_4\mu_{1,1}^2 - \boldsymbol{l}_K^\top \mathbf{M}_K^{-1} \boldsymbol{l}_K = q_4\mu_{1,1}^2 - \begin{bmatrix} q_4\mu_{1,1} \cdot \boldsymbol{\mu}_K \\ (1+q_1)\mu_{1,1} \end{bmatrix}^\top \begin{bmatrix} \mathbf{V}_0 & (1+q_1)\boldsymbol{\mu}_K^\top \\ (1+q_1)\boldsymbol{\mu}_K & q_5 + \frac{1}{\overline{m}_{K+1,d}} \end{bmatrix}^{-1} \begin{bmatrix} q_4\mu_{1,1} \cdot \boldsymbol{\mu}_K \\ (1+q_1)\mu_{1,1} \end{bmatrix}. \tag{VII.16}$$

Clearly, the formula in (VII.16) requires us to investigate $\boldsymbol{\mu}_K^\top \mathbf{V}_0^{-1} \boldsymbol{\mu}_K$ first. Under the case $K-1$, (VII.15) holds from the induction hypothesis. Thus if we set $(1+q_1) = q_4\mu_{K,1}$, $q_5 = q_4\mu_{K,1}^2$, we have $\boldsymbol{l}_{K-1} = q_4\mu_{1,1} \cdot \boldsymbol{\mu}_K$. Plugging $\boldsymbol{l}_{K-1} = q_4\mu_{1,1} \cdot \boldsymbol{\mu}_K$ into (VII.15) we obtain that

$$\boldsymbol{\mu}_K^\top \mathbf{V}_0^{-1} \boldsymbol{\mu}_K \tag{VII.17}$$

$$= \frac{1}{q_4^2\mu_{1,1}^2}\left(q_4\mu_{1,1}^2 - \frac{\mu_{1,1}^2 q_4(1+\mu_{K,1}^2 q_4 \overline{m}_{K,d}) - \mu_{1,1}^2 q_4^2 \mu_{K,1}^2 \overline{m}_{K,d}}{\left(1+q_4\sum_{j=1}^{K-1}\mu_{j,1}^2\overline{m}_{j,d}\right)(1+\mu_{K,1}^2 q_4\overline{m}_{K,d}) - q_4^2\mu_{K,1}^2\sum_{j=1}^{K-1}\mu_{j,1}^2\overline{m}_{j,d}\overline{m}_{K,d}}\right)$$

$$= \frac{1}{q_4\mu_{1,1}^2}\left(\mu_{1,1}^2 - \frac{\mu_{1,1}^2}{1+q_4\sum_{j=1}^{K}\mu_{j,1}^2\overline{m}_{j,d}}\right) = \frac{\sum_{j=1}^{K}\mu_{j,1}^2\overline{m}_{j,d}}{1+q_4\sum_{j=1}^{K}\mu_{j,1}^2\overline{m}_{j,d}}.$$

Therefore, for the case K, we have

$$q_4\mu_{1,1}^2 - \boldsymbol{l}_K^\top \mathbf{M}_K^{-1}\boldsymbol{l}_K = q_4\mu_{1,1}^2 - \begin{bmatrix} q_4\mu_{1,1}\cdot\boldsymbol{\mu}_K \\ (1+q_1)\mu_{1,1}\end{bmatrix}^\top \begin{bmatrix}\mathbf{V}_0 & (1+q_1)\boldsymbol{\mu}_K^\top \\ (1+q_1)\boldsymbol{\mu}_K & q_5 + \frac{1}{\overline{m}_{K+1,d}}\end{bmatrix}^{-1}\begin{bmatrix} q_4\mu_{1,1}\cdot\boldsymbol{\mu}_K \\ (1+q_1)\mu_{1,1}\end{bmatrix}$$

$$= \frac{\mu_{1,1}^2 q_4(1+q_5\overline{m}_{K,d}) - \mu_{1,1}^2(1+q_1)^2\overline{m}_{K,d}}{\left(1+q_4\sum_{j=1}^{K-1}\mu_{j,1}^2\overline{m}_{j,d}\right)(1+q_5\overline{m}_{K,d}) - (1+q_1)^2\sum_{j=1}^{K-1}\mu_{j,1}^2\overline{m}_{j,d}\overline{m}_{K,d}}.$$

Here, the first equality directly comes from (VII.16), and the second equality comes from Schur complement and (VII.17) after direct calculation. We completed the mathematical induction for the general case K and equation (VII.14) is proved. Then we have

$$\overline{m}_{1,d} = \psi_1\left\{-\xi + q_2\mu_{1,*}^2 - \mu_{1,*}^2\overline{m}_{K+1,d} + \frac{H_{1,d}}{H_{D,d}}\right\}^{-1} + o_d(1),$$

where

$$H_{1,d} = \mu_{j,1}^2 q_4(1+q_5\overline{m}_{K+1,d}) - \mu_{j,1}^2(1+q_1)^2\overline{m}_{K+1,d}, \quad j=1,\dots,K,$$

$$H_{D,d} = \left(1+q_4\sum_{j=1}^{K}\mu_{j,1}^2\overline{m}_{j,d}\right)(1+q_5\overline{m}_{K+1,d}) - (1+q_1)^2\sum_{j=1}^{K}\mu_{j,1}^2\overline{m}_{j,d}\overline{m}_{K+1,d}.$$

After similar argument, we conclude that

$$\overline{m}_{j,d} = \psi_j\left\{-\xi + s_j\mu_{j,*}^2 - \mu_{j,*}^2\overline{m}_{K+1,d} + \frac{H_{j,d}}{H_{D,d}}\right\}^{-1} + o_d(1), \quad j=1,\dots,K+1,$$

$$\overline{m}_{K+1,d} = \psi_{K+1}\left\{-\xi + q_3 - \sum_{j=1}^{K}\mu_{j,*}^2\overline{m}_{j,d} + \frac{H_{K+1,d}}{H_{D,d}}\right\}^{-1} + o_d(1),$$

where

$$H_{j,d} = \mu_{j,1}^2 q_4(1+q_5\overline{m}_{K+1,d}) - \mu_{j,1}^2(1+q_1)^2\overline{m}_{K+1,d}, \quad j=1,\dots,K,$$

$$H_{K+1,d} = q_5\left(1+q_4\sum_{j=1}^{K}\mu_{j,1}^2\overline{m}_{j,d}\right) - (1+q_1)^2\sum_{j=1}^{K}\mu_{j,1}^2\overline{m}_{j,d}.$$

We get that each element of $\overline{\mathbf{m}}_d(\xi) - \mathbf{F}(\overline{\mathbf{m}}_d(\xi))$ is asymptotically zero. Therefore $\|\overline{\mathbf{m}}_d(\xi) - \mathbf{F}(\overline{\mathbf{m}}_d(\xi))\|_2 = o_d(1)$. The remaining arguments are similar to those in Section IV and the details are skipped. Wrapping all together, we complete the proof of Proposition C.7. ■

References

Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04), 2013.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Luogeng Hua. *Harmonic Analysis of Functions of Several Complex Variables in the Classical Domains*. American Mathematical Soc., 1963.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Terence Tao. *Topics in Random Matrix Theory*. American Mathematical Soc., 2012.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.