SUPPLEMENT TO "HARD PARAMETER SHARING IN HIGH-DIMENSIONAL LINEAR REGRESSION"

BY FAN YANG [*1] HONGYANG R. ZHANG[*2,†] SEN WU[3,‡] WEIJIE J. SU [1,*] AND CHRISTOPHER RÉ[3,§]

[1]*Department of Statistics, University of Pennsylvania, fyang75@wharton.upenn.edu;* [*] *suw@wharton.upenn.edu*

[2]*Khoury College of Computer Sciences, Northeastern University,* [†] *hrzhang@northeastern.edu*

[3]*Department of Computer Science, Stanford University,* [‡] *senwu@stanford.edu;* [§] *chrismre@cs.stanford.edu*

In this supplementary file, we provide the proofs of some technical results in the main manuscript, including Lemma 2.2, Theorem 5.1, Theorem 3.1, Proposition 3.3, Theorem 3.4 and Proposition 3.6.

A. Basic tools. In this section, we collect some basic tools that will be used in the proof. First, for our proof it is convenient to introduce the following notation.

DEFINITION A.1 (Overwhelming probability). We say Ξ holds *with overwhelming probability* (w.o.p.) if for any constant $D > 0$, $\mathbb{P}(\Xi) \geqslant 1 - p^{-D}$ for large enough p. Moreover, we say Ξ holds with overwhelming probability in an event Ω if for any constant $D > 0$, $\mathbb{P}(\Omega \setminus \Xi) \leqslant p^{-D}$ for large enough p.

The following notion of stochastic domination, which was first introduced in [10], is commonly used in the study of random matrices.

DEFINITION A.2 (Stochastic domination). Let $\xi \equiv \xi^{(n)}$ and $\zeta \equiv \zeta^{(n)}$ be two n-dependent random variables. We say that ξ is stochastically dominated by ζ, denoted by $\xi \prec \zeta$ or $\xi = \mathrm{O}_{\prec}(\zeta)$, if for any small constant $c > 0$ and any large constant $D > 0$, there exists a function $n_0(c, D)$ such that for all $n > n_0(c, D)$,

$$\mathbb{P}\left(|\xi| > n^c |\zeta|\right) \leqslant n^{-D}.$$

In other word, $\xi \prec \zeta$ if $|\xi| \leqslant n^c |\zeta|$ with overwhelming probability for any small constant $c > 0$. If $\xi(u)$ and $\zeta(u)$ are functions of u supported in $\mathcal{U}$, then we say $\xi(u)$ is stochastically dominated by $\zeta(u)$ uniformly in $\mathcal{U}$ if

$$\sup_{u \in \mathcal{U}} \mathbb{P}\left(|\xi(u)| > n^c |\zeta(u)|\right) \leqslant n^{-D}.$$

Given any event Ω, we say $\xi \prec \zeta$ on Ω if $\mathbf{1}_\Omega \xi \prec \zeta$.

REMARK A.3. We make several simple remarks. First, since we allow for an n^c factor in stochastic domination, we can ignore $\log$ factors without loss of generality since $(\log n)^C \prec 1$ for any constant $C > 0$. Second, given a random variable ξ with unit variance and finite moments up to any order as in (2.6), we have that $|\xi| \prec 1$. This is because by Markov's inequality, we have that

$$\mathbb{P}(|\xi| \geqslant n^c) \leqslant n^{-kc} \mathbb{E}|\xi|^k \leqslant n^{-D},$$

as long as k is taken to be larger than D/c.

[*]Fan Yang and Hongyang R. Zhang contributed equally.

MSC2020 subject classifications: Primary 62J05, 60B20; secondary 62E20, 62H10.

Keywords and phrases: Hard parameter sharing, high-dimensional linear regression, random matrix theory, sample covariance matrices.

The following lemma collects several basic properties of stochastic domination that will be used tacitly in the proof. Roughly speaking, it says that the stochastic domination "$\prec$" can be treated as the conventional less-than sign "$<$" in some sense.

LEMMA A.4 (Lemma 3.2 in [7]). *Let ξ and ζ be two families of nonnegative random variables depending on some parameters $u \in \mathcal{U}$ or $v \in \mathcal{V}$.*

(i) **Sum.** *Suppose that $\xi(u,v) \prec \zeta(u,v)$ uniformly in $u \in \mathcal{U}$ and $v \in \mathcal{V}$. If $|\mathcal{V}| \leqslant n^C$ for some constant $C > 0$, then $\sum_{v\in\mathcal{V}} \xi(u,v) \prec \sum_{v\in\mathcal{V}} \zeta(u,v)$ uniformly in u.*

(ii) **Product.** *If $\xi_1(u) \prec \zeta_1(u)$ and $\xi_2(u) \prec \zeta_2(u)$ uniformly in $u \in \mathcal{U}$, then $\xi_1(u)\xi_2(u) \prec \zeta_1(u)\zeta_2(u)$ uniformly in $u \in \mathcal{U}$.*

(iii) **Expectation.** *Suppose that $\Psi(u) \geqslant n^{-C}$ is a family of deterministic parameters, and $\xi(u)$ satisfies $\mathbb{E}\xi(u)^2 \leqslant n^C$. If $\xi(u) \prec \Psi(u)$ uniformly in u, then we also have $\mathbb{E}\xi(u) \prec \Psi(u)$ uniformly in u.*

We say that a random matrix $Z \in \mathbb{R}^{n\times p}$ satisfies the *bounded support condition* with Q or Z has support Q if

$$\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p} |Z_{ij}| \prec Q. \tag{A.1}$$

As shown in the example above, if the entries of Z have finite moments up to any order, then Z has bounded support 1. More generally, if the entries of Z have finite φ-th moment, then using Markov's inequality and a simple union bound we get that

$$\begin{aligned}\mathbb{P}\left(\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p} |Z_{ij}| \geqslant (\log n) n^{\frac{2}{\varphi}}\right) &\leqslant \sum_{i=1}^n \sum_{j=1}^p \mathbb{P}\left(|Z_{ij}| \geqslant (\log n) n^{\frac{2}{\varphi}}\right) \\ &\lesssim \sum_{i=1}^n \sum_{j=1}^p \left[(\log n) n^{\frac{2}{\varphi}}\right]^{-\varphi} = \mathrm{O}((\log n)^{-\varphi}).\end{aligned} \tag{A.2}$$

In other words, Z has bounded support $Q = n^{\frac{2}{\varphi}}$ with high probability.

The following lemma gives sharp concentration bounds for linear and quadratic forms of random variables with bounded support.

LEMMA A.5 (Lemma 3.8 of [11] and Theorem B.1 of [12]). *Let (x_i), (y_j) be independent families of centered and independent random variables, and (A_i), (B_{ij}) be families of deterministic complex numbers. Suppose the entries x_i and y_j have variance at most 1, and satisfy the bounded support condition (A.1) for a deterministic parameter $Q \geqslant 1$. Then we have the following results:*

$$\left|\sum_{i=1}^n A_i x_i\right| \prec Q \max_i |A_i| + \Big(\sum_i |A_i|^2\Big)^{1/2}, \tag{A.3}$$

$$\left|\sum_{i,j=1}^n x_i B_{ij} y_j\right| \prec Q^2 B_d + Q n^{1/2} B_o + \Big(\sum_{i\neq j} |B_{ij}|^2\Big)^{1/2}, \tag{A.4}$$

$$\left|\sum_{i=1}^n (|x_i|^2 - \mathbb{E}|x_i|^2) B_{ii}\right| \prec Q n^{1/2} B_d, \tag{A.5}$$

$$\left|\sum_{1\leqslant i\neq j\leqslant n} \bar{x}_i B_{ij} x_j\right| \prec Q n^{1/2} B_o + \Big(\sum_{i\neq j} |B_{ij}|^2\Big)^{1/2}, \tag{A.6}$$

where we denote $B_d := \max_i |B_{ii}|$ *and* $B_o := \max_{i\neq j} |B_{ij}|$. *Moreover, if the moments of* x_i *and* y_j *exist up to any order, then we have the following stronger results:*

$$\Big|\sum_i A_i x_i\Big| \prec \Big(\sum_i |A_i|^2\Big)^{1/2}, \tag{A.7}$$

$$\Big|\sum_{i,j} x_i B_{ij} y_j\Big| \prec \Big(\sum_{i,j} |B_{ij}|^2\Big)^{1/2}, \tag{A.8}$$

$$\Big|\sum_i (|x_i|^2 - \mathbb{E}|x_i|^2) B_{ii}\Big| \prec \Big(\sum_i |B_{ii}|^2\Big)^{1/2}, \tag{A.9}$$

$$\Big|\sum_{i\neq j} \bar{x}_i B_{ij} x_j\Big| \prec \Big(\sum_{i\neq j} |B_{ij}|^2\Big)^{1/2}. \tag{A.10}$$

It is well-known that the limiting eigenvalue distributions of $(Z^{(1)})^\top Z^{(1)}$ and $(Z^{(2)})^\top Z^{(2)}$ satisfy the Marchenko-Pastur (MP) law [16]. Moreover, their eigenvalues are all inside the support of the MP with high probability [3]. In the proof, we shall need a slightly stronger result that holds with overwhelming probability as given by the following lemma.

LEMMA A.6. *Suppose* $Z \in \mathbb{R}^{n\times p}$ *is an* $n \times p$ *random matrix satisfying the same assumptions as* $Z^{(1)}$ *and* $Z^{(2)}$ *in Assumption 2.1, and* $\rho := n/p$ *satisfies* (2.7). *Moreover, suppose* Z *satisfies the bounded support condition* (A.1) *for a deterministic parameter* Q *satisfying* $1 \leqslant Q \leqslant n^{1/2-c_Q}$ *for a small constant* $c_Q > 0$. *Then for any constant* $c > 0$, *we have that*

$$(\sqrt{n} - \sqrt{p})^2 - \mathrm{O}_\prec(n \cdot Q) \leqslant \lambda_p(Z^\top Z) \leqslant \lambda_1(Z^\top Z) \leqslant (\sqrt{n} - \sqrt{p})^2 + \mathrm{O}_\prec(n \cdot Q). \tag{A.11}$$

PROOF. When Q is of order 1, this lemma follows from [7, Theorem 2.10]. The estimate for the case with larger Q satisfying $1 \leqslant Q \leqslant n^{1/2-c_Q}$ follows from [9, Lemma 3.11]. □

Using a standard cut-off argument, we can extend Lemma A.5 and Lemma A.6 to the random matrices whose entries only satisfy certain moment assumptions but not necessarily the bounded support condition.

COROLLARY A.7. *Suppose* $Z \in \mathbb{R}^{n\times p}$ *is an* $n \times p$ *random matrix satisfying the same assumptions as* $Z^{(1)}$ *and* $Z^{(2)}$ *in Assumption 2.1, and* $\rho := n/p$ *satisfies* (2.7). *Then* (A.11) *holds on a high probability event with* $Q = n^{2/\varphi}$, *where* φ *is the constant in* (2.3).

PROOF. For $Q = n^{2/\varphi}$, we introduce a truncated matrix $\widetilde{Z}$ with entries

$$\widetilde{Z}_{ij} := \mathbf{1}\left(|Z_{ij}| \leqslant Q \log n\right) \cdot Z_{ij}. \tag{A.12}$$

From equation (A.2), we get

$$\mathbb{P}(\widetilde{Z} = Z) = 1 - \mathbb{P}\left(\max_{i,j} |Z_{ij}| > (\log n) n^{\frac{2}{\varphi}}\right) = 1 - \mathrm{O}((\log n)^{-\varphi}). \tag{A.13}$$

By definition, we have

$$\begin{aligned} \mathbb{E}\widetilde{Z}_{ij} &= -\mathbb{E}\left[\mathbf{1}\left(|Z_{ij}| > Q \log n\right) Z_{ij}\right], \\ \mathbb{E}|\widetilde{Z}_{ij}|^2 &= 1 - \mathbb{E}\left[\mathbf{1}\left(|Z_{ij}| > Q \log n\right) |Z_{ij}|^2\right]. \end{aligned} \tag{A.14}$$

Using the formula for expectation in terms of the tail probabilities, we can check that

$$\begin{aligned}\mathbb{E}\left|\mathbf{1}\left(|Z_{ij}|>Q\log n\right)Z_{ij}\right| &= \int_0^\infty \mathbb{P}\left(\left|\mathbf{1}\left(|Z_{ij}|>Q\log n\right)Z_{ij}\right|>s\right)\mathrm{d}s\\ &= \int_0^{Q\log n}\mathbb{P}\left(|Z_{ij}|>Q\log n\right)\mathrm{d}s+\int_{Q\log n}^\infty \mathbb{P}\left(|Z_{ij}|>s\right)\mathrm{d}s\\ &\lesssim \int_0^{Q\log n}(Q\log n)^{-\varphi}\,\mathrm{d}s+\int_{Q\log n}^\infty s^{-\varphi}\mathrm{d}s \leqslant n^{-2(\varphi-1)/\varphi},\end{aligned}$$

where in the third step we used the finite φ-th moment condition (2.3) for Z_{ij} and Markov's inequality. Similarly, we can obtain that

$$\begin{aligned}\mathbb{E}\left|\mathbf{1}\left(|Z_{ij}|>Q\log n\right)Z_{ij}\right|^2 &= 2\int_0^\infty s\mathbb{P}\left(\left|\mathbf{1}\left(|Z_{ij}|>Q\log n\right)Z_{ij}\right|>s\right)\mathrm{d}s\\ &= 2\int_0^{Q\log n}s\mathbb{P}\left(|Z_{ij}|>Q\log n\right)\mathrm{d}s+2\int_{Q\log n}^\infty s\mathbb{P}\left(|Z_{ij}|>s\right)\mathrm{d}s\\ &\lesssim \int_0^{Q\log n}s\,(Q\log n)^{-\varphi}\,\mathrm{d}s+\int_{Q\log n}^\infty s^{-\varphi+1}\mathrm{d}s \leqslant n^{-2(\varphi-2)/\varphi}.\end{aligned}$$

Plugging the above two estimates into equation (A.14) and using $\varphi>4$, we get that

$$|\mathbb{E}Z_{ij}|=\mathrm{O}(n^{-3/2}),\quad \mathbb{E}|Z_{ij}|^2=1+\mathrm{O}(n^{-1}). \tag{A.15}$$

From the first estimate in equation (A.15), we can also get a bound on the operator norm:

$$\|\mathbb{E}Z\|=\mathrm{O}(n^{-1/2}). \tag{A.16}$$

We centralize and rescale $\widetilde{Z}$ as

$$\widehat{Z}:=(\mathbb{E}|\widetilde{Z}_{11}|^2)^{-1/2}(\widetilde{Z}-\mathbb{E}\widetilde{Z}). \tag{A.17}$$

Now $\widehat{Z}$ satisfies the assumptions of Lemma A.6 with bounded support Q, hence we get

$$(\sqrt{n}-\sqrt{p})^2-\mathrm{O}_\prec(n\cdot Q)\leqslant\lambda_p(\widehat{Z}^\top\widehat{Z})\leqslant\lambda_1(\widehat{Z}^\top\widehat{Z})\leqslant(\sqrt{n}-\sqrt{p})^2+\mathrm{O}_\prec(n\cdot Q).$$

Combining this estimate with (A.15) and (A.16), it is easy to show that (A.11) holds for the eigenvalues of $\widetilde{Z}^\top\widetilde{Z}$, which concludes the proof by (A.13). □

COROLLARY A.8. *Suppose $Z\in\mathbb{R}^{n\times p}$ is an $n\times p$ random matrix satisfying Assumption 2.1. Then there exists a high probability event, on which for any deterministic vector $v\in\mathbb{R}^p$,*

$$\left|\|Zv\|^2-n\|v\|^2\right|\prec n^{1/2}Q\|v\|^2,\quad \textit{for } Q=n^{2/\varphi}. \tag{A.18}$$

PROOF. As in the proof of Corollary A.7, we truncate Z as in (A.12) and define $\widehat{Z}$ as in (A.17). By (A.15) and (A.16), we see that to conclude (A.18), it suffices to show

$$\left|\|\widehat{Z}v\|^2-n\|v\|^2\right|\leqslant n^{1/2+c}Q\|v\|^2. \tag{A.19}$$

To prove equation (A.19), we first notice that $\widehat{Z}v\in\mathbb{R}^n$ is a random vector with i.i.d. entries of mean zero and variance $\|v\|^2$. Furthermore, using equation (A.3) from Lemma A.5, we get that

$$|(\widehat{Z}v)_i|\prec Q\max_{1\leqslant i\leqslant p}|v_i|+\|v\|\leqslant 2Q\|v\|.$$

Hence, $(\widehat{Z}v)/\|v\|$ consists of i.i.d random variables with zero mean, unit variance, and bounded support Q. Then applying equation (A.5), we get that

$$\left|\|\widehat{Z}v\|^2 - n\|v\|^2\right| = \|v\|^2 \left|\sum_i \left(\frac{|(\widehat{Z}v)_i|^2}{\|v\|^2} - \mathbb{E}\frac{|(\widehat{Z}v)_i|^2}{\|v\|^2}\right)\right| \prec Qn^{1/2}\|v\|^2.$$

Hence the proof is complete. □

From Lemma A.5, we immediately obtain the following concentration results for the noise vectors.

COROLLARY A.9. *Suppose $\varepsilon^{(1)}, \cdots, \varepsilon^{(t)} \in \mathbb{R}^n$ are independent random vectors satisfying Assumption 2.1. For any deterministic vector $v \in \mathbb{R}^n$, we have that*

$$|v^\top \varepsilon^{(i)}| \prec \sigma\|v\|, \quad i = 1, \cdots, t. \tag{A.20}$$

For any deterministic matrix $B \in \mathbb{R}^{n\times n}$, we have that

$$\left|(\varepsilon^{(i)})^\top B\varepsilon^{(j)} - \delta_{ij}\cdot\sigma^2\operatorname{Tr}(B)\right| \prec \sigma^2\|B\|_F, \quad i = 1, \cdots, t. \tag{A.21}$$

PROOF. Note that $\varepsilon^{(i)}/\sigma$ is a random vector with i.i.d. entries of zero mean, unit variance, and bounded moments up to any order. Then equation (A.20) is an immediate consequence of (A.7). Using equation (A.8), we obtain that for $i \neq j$,

$$\left|(\varepsilon^{(i)})^\top B\varepsilon^{(j)}\right| = \Big|\sum_{k,l=1}^n \varepsilon_k^{(i)}\varepsilon_l^{(j)}B_{kl}\Big| \prec \sigma^2\Big(\sum_{k,l=1}^n |B_{kl}|^2\Big)^{1/2} = \sigma^2\|B\|_F.$$

Using the two estimates (A.9) and (A.10), we obtain that

$$\begin{aligned}\left|(\varepsilon^{(i)})^\top B\varepsilon^{(i)} - \sigma^2\operatorname{Tr}[B]\right| &\leqslant \Big|\sum_k \left(|\varepsilon_k^{(i)}|^2 - \mathbb{E}|\varepsilon_k^{(i)}|^2\right)B_{ii}\Big| + \Big|\sum_{k\neq l}\varepsilon_k^{(i)}\varepsilon_l^{(i)}B_{kl}\Big| \\ &\prec \sigma^2\Big(\sum_k |B_{kk}|^2\Big) + \sigma^2\Big(\sum_{k\neq l}|B_{kl}|^2\Big)^{1/2} \lesssim \sigma^2\|B\|_F.\end{aligned}$$

Hence, we have shown equation (A.21). □

B. Proof of Lemma 2.2. In this section, we prove Lemma 2.2 using the estimates in Appendix A. For a fixed $a \in \mathbb{R}$, we expand $L(\hat{\beta}_2^{\text{HPS}}(a))$ as

$$L(\hat{\beta}_2^{\text{HPS}}(a)) = L_{\text{bias}}(a) + 2h_1(a) + 2h_2(a) + h_3(a) + h_4(a) + 2h_5(a), \tag{B.1}$$

where

$$\begin{aligned}h_1(a) &:= a(\varepsilon^{(1)})^\top X^{(1)}\hat{\Sigma}(a)^{-1}\Sigma^{(2)}\hat{\Sigma}(a)^{-1}(X^{(1)})^\top X^{(1)}(a\beta^{(1)} - a^2\beta^{(2)}),\\ h_2(a) &:= (\varepsilon^{(2)})^\top X^{(2)}\hat{\Sigma}(a)^{-1}\Sigma^{(2)}\hat{\Sigma}(a)^{-1}(X^{(1)})^\top X^{(1)}(a\beta^{(1)} - a^2\beta^{(2)}),\\ h_3(x) &:= a^2(\varepsilon^{(1)})^\top X^{(1)}\hat{\Sigma}(a)^{-1}\Sigma^{(2)}\hat{\Sigma}(a)^{-1}(X^{(1)})^\top\varepsilon^{(1)},\\ h_4(x) &:= (\varepsilon^{(2)})^\top X^{(2)}\hat{\Sigma}(a)^{-1}\Sigma^{(2)}\hat{\Sigma}(a)^{-1}(X^{(2)})^\top\varepsilon^{(2)},\\ h_5(x) &:= a(\varepsilon^{(1)})^\top X^{(1)}\hat{\Sigma}(a)^{-1}\Sigma^{(2)}\hat{\Sigma}(a)^{-1}(X^{(2)})^\top\varepsilon^{(2)}.\end{aligned}$$

Next, we estimate each term using Corollary A.9. In the proof, we will use the following estimates, which can be proved easily using (2.5) and Corollary A.7: on a high probability event Ξ_1,

$$\|X^{(1)}\| \lesssim \sqrt{n_1}, \quad \|X^{(2)}\| \lesssim \sqrt{n_2}, \quad \left\|\hat{\Sigma}(a)^{-1}\right\| \lesssim (a^2 n_1 + n_2)^{-1}, \tag{B.2}$$

and

$$L_{\text{bias}}(a) \sim \frac{n_1^2}{(a^2 n_1 + n_2)^2} \left\|a\beta^{(1)} - a^2\beta^{(2)}\right\|^2, \quad L_{\text{Var}}(a) \sim \frac{p\sigma^2}{a^2 n_1 + n_2}. \tag{B.3}$$

Throughout the following proof, we assume that event Ξ_1 holds.

For $h_1(a)$, using (A.20) we obtain that

$$\begin{aligned}
|h_1(a)| &\prec \sigma|a| \left\|X^{(1)}\hat{\Sigma}(a)^{-1}\Sigma^{(2)}\hat{\Sigma}(a)^{-1}(X^{(1)})^\top X^{(1)}(a\beta^{(1)} - a^2\beta^{(2)})\right\| \\
&\lesssim \sigma|a| \left\|a\beta^{(1)} - a^2\beta^{(2)}\right\| \cdot \left\|X^{(1)}\hat{\Sigma}(a)^{-1}\Sigma^{(2)}\hat{\Sigma}(a)^{-1}(X^{(1)})^\top X^{(1)}\right\| \\
&\lesssim \frac{\sigma|a| \left\|a\beta^{(1)} - a^2\beta^{(2)}\right\| n_1^{3/2}}{(a^2 n_1 + n_2)^2} \leqslant \frac{p^{-1/4} n_1 \left\|a\beta^{(1)} - a^2\beta^{(2)}\right\|}{a^2 n_1 + n_2} \cdot \frac{p^{1/4}\sigma}{(a^2 n_1 + n_2)^{1/2}} \\
&\leqslant \frac{p^{-1/2} n_1^2 \left\|a\beta^{(1)} - a^2\beta^{(2)}\right\|^2}{(a^2 n_1 + n_2)^2} + \frac{p^{1/2}\sigma^2}{a^2 n_1 + n_2} \lesssim p^{-1/2}\left[L_{\text{bias}}(a) + L_{\text{Var}}(a)\right],
\end{aligned}$$

where we use (2.5) and (B.2) in the third step, the AM-GM inequality in the fifth step, and (B.3) in the last step. Similarly, we can show that

$$|h_2(a)| \prec p^{-1/2}\left[L_{\text{bias}}(a) + L_{\text{Var}}(a)\right].$$

For $h_3(a)$, using (A.21) we obtain that

$$\begin{aligned}
&\left|h_3(a) - \sigma^2 a^2 \operatorname{Tr}\left[X^{(1)}\hat{\Sigma}(a)^{-1}\Sigma^{(2)}\hat{\Sigma}(a)^{-1}(X^{(1)})^\top\right]\right| \\
&\prec \sigma^2 a^2 \left\|X^{(1)}\hat{\Sigma}(a)^{-1}\Sigma^{(2)}\hat{\Sigma}(a)^{-1}(X^{(1)})^\top\right\|_F \lesssim \frac{\sigma^2 a^2}{(a^2 n_1 + n_2)^2} \left\|X^{(1)}(X^{(1)})^\top\right\|_F \\
&\lesssim \frac{\sigma^2 a^2 p^{1/2} n_1}{(a^2 n_1 + n_2)^2} \leqslant \frac{\sigma^2 p^{1/2}}{a^2 n_1 + n_2} \lesssim p^{-1/2} L_{\text{Var}}(a).
\end{aligned}$$

where we use (2.5) and (B.2) in the second step, and $\left\|X^{(1)}(X^{(1)})^\top\right\|_F \leqslant p^{1/2}\left\|X^{(1)}(X^{(1)})^\top\right\| \lesssim p^{1/2} n_1$ in the last step. Similarly, we can show that

$$\left|h_4(a) - \sigma^2 \operatorname{Tr}\left[X^{(2)}\hat{\Sigma}(a)^{-1}\Sigma^{(2)}\hat{\Sigma}(a)^{-1}(X^{(2)})^\top\right]\right| \prec p^{-1/2} L_{\text{Var}}(a),$$

and

$$|h_5(a)| \prec p^{-1/2} L_{\text{Var}}(a).$$

Combining the above estimates on $h_i(a)$, $i = 1, 2, 3, 4, 5$, and using that

$$\begin{aligned}
&\sigma^2 a^2 \operatorname{Tr}\left[X^{(1)}\hat{\Sigma}(a)^{-1}\Sigma^{(2)}\hat{\Sigma}(a)^{-1}(X^{(1)})^\top\right] + \sigma^2 \operatorname{Tr}\left[X^{(2)}\hat{\Sigma}(a)^{-1}\Sigma^{(2)}\hat{\Sigma}(a)^{-1}(X^{(2)})^\top\right] \\
&= \sigma^2 \operatorname{Tr}\left[\Sigma^{(2)}\hat{\Sigma}(a)^{-1}\right] = L_{\text{Var}}(a),
\end{aligned}$$

we conclude that on event Ξ_1,

$$L(\hat{\beta}_2^{\text{HPS}}(a)) = \left[1 + \mathrm{O}_\prec(p^{-1/2})\right] \cdot \left[L_{\text{bias}}(a) + L_{\text{Var}}(a)\right], \tag{B.4}$$

for any fixed $a \in \mathbb{R}$. With a similar argument, we can obtain (2.20) and (2.21).

Finally, it remains to extend the result (B.4) to all $a \in \mathbb{R}$ in a uniform way. Fix a large enough constant $C_0 > 0$, we consider a belonging to a discrete set

$$S := \left\{ a_k = \lceil k p^{-C_0} \rceil : -p^{2C_0} \leqslant k \leqslant p^{2C_0} \right\}.$$

Then combining (B.4) with a simple union bound, we obtain that for any small constant $c > 0$, the event

$$\Xi := \left\{ \left| L(\hat{\beta}_2^{\text{HPS}}(a) - L_{\text{bias}}(a) - L_{\text{Var}}(a) \right| \leqslant p^{-1/2+c} \left[L_{\text{bias}}(a) + L_{\text{Var}}(a) \right] \text{ for all } a \in S \right\} \cap \Xi_1$$

holds with high probability. Using (B.2) and (B.3), it is easy to check that on Ξ, for all $a_k \leqslant a \leqslant a_{k+1}$, $-p^{2C_0} \leqslant k \leqslant p^{2C_0} - 1$, the following deterministic estimates hold:

$$\left| L(\hat{\beta}_2^{\text{HPS}}(a)) - L(\hat{\beta}_2^{\text{HPS}}(a_k)) \right| \lesssim p^{-C_0} \left(\|\varepsilon^{(1)}\|^2 + \|\varepsilon^{(2)}\|^2 + \|\beta^{(1)}\|^2 + \|\beta^{(2)}\|^2 \right),$$

and

$$|L_{\text{bias}}(a) - L_{\text{bias}}(a_k)| + |L_{\text{Var}}(a) - L_{\text{Var}}(a_k)| \lesssim p^{-C_0} \left(\|\varepsilon^{(1)}\|^2 + \|\varepsilon^{(2)}\|^2 + \|\beta^{(1)}\|^2 + \|\beta^{(2)}\|^2 \right).$$

Similarly, we can also show that the above two estimates hold on Ξ if $a \geqslant p^{C_0}$ and $a_k = a_{\lceil p^{C_0} \rceil}$, or if $a \leqslant -p^{C_0}$ and $a_k = a_{-\lceil p^{C_0} \rceil}$. Now by triangle inequality, we get that on Ξ,

$$\begin{aligned} \left| L(\hat{\beta}_2^{\text{HPS}}(a)) - L_{\text{bias}}(a) - L_{\text{Var}}(a) \right| &\lesssim p^{-1/2+c} \left[L_{\text{bias}}(a) + L_{\text{Var}}(a) \right] \\ &\quad + p^{-C_0} \left(\|\varepsilon^{(1)}\|^2 + \|\varepsilon^{(2)}\|^2 + \|\beta^{(1)}\|^2 + \|\beta^{(2)}\|^2 \right), \end{aligned} \tag{B.5}$$

holds simultaneously for all $a \in \mathbb{R}$. Notice that

$$\|\varepsilon^{(i)}\|^2 = n\sigma^2 + \mathrm{O}_{\prec}(n^{1/2}\sigma^2), \quad i = 1, 2.$$

Plugging it into (B.5), we can conclude (2.19) since C_0 is arbitrary.

C. Proof of Theorem 5.1. In this section, we give the proof of Theorem 5.1. Note that $A^\top (AA^\top)^+ A$ is a projection onto the subspace spanned by the rows of A. For simplicity, we write it into the form

$$A^\top (AA^\top)^+ A = U_A U_A^\top,$$

where $U_A \in \mathbb{R}^{t \times r}$ is a $t \times r$ partial orthonormal matrix (i.e. $U_A^\top U_A = \mathrm{Id}_{r \times r}$). Hence we also denote the function $g(A)$ by $g(U_A)$. First, we can use the concentration estimate, Corollary A.9, to simplify the expression of $g(U_A)$. In this section, we always let $Q = n^{2/\varphi}$.

LEMMA C.1. *In the setting of Theorem 5.1, for any small constant $c > 0$ and large constant $C > 0$, there exists a high probability event Ξ, on which the following estimate holds:*

$$\begin{aligned} |g(U_A) - h(U_A)| &\leqslant Q n^{1/2+c} \left\| \Sigma^{1/2} B^\star (U_A U_A^\top - \mathrm{Id}_{t \times t}) \right\|_F^2 \\ &\quad + \sigma^2 n^{1/2+c} + p^{-C} \left\| \Sigma^{1/2} B^\star \right\|_F^2, \end{aligned} \tag{C.1}$$

uniformly in all rank-r partial orthonormal matrices $U_A \in \mathbb{R}^{t \times r}$. Here

$$h(U_A) := n \left\| \Sigma^{1/2} B^\star \left(U_A U_A^\top - \mathrm{Id}_{t \times t} \right) \right\|_F^2 + \sigma^2 (nt - pr). \tag{C.2}$$

PROOF. With Corollary A.7 and Corollary A.8, we can choose a high probability event Ξ_1 on which (A.18) holds and

$$C^{-1}n \leqslant \lambda_p(Z^\top Z) \leqslant \lambda_1(Z^\top Z) \leqslant Cn \tag{C.3}$$

for a large constant $C > 0$. Throughout the following proof, we assume that event Ξ_1 holds.

To facilitate the analysis, we introduce the following matrix notations. Denote

$$\mathcal{E} := [\varepsilon^{(1)}, \varepsilon^{(2)}, \cdots, \varepsilon^{(t)}], \quad \text{and} \quad \mathcal{W} := X(X^\top X)^{-1} X^\top \mathcal{E} U_A U_A^\top.$$

For any $j = 1, 2, \ldots, t$, let

$$H_j := B^\star \left(U_A U_A^\top - \mathrm{Id}_{t\times t} \right) e_j, \quad \text{and} \quad E_j := (\mathcal{W} - \mathcal{E}) e_j. \tag{C.4}$$

where e_j is the standard basis unit vector along j-th direction. Then plugging $Y = XB^\star + \mathcal{E}$ into (5.3), we can write the function $g(U_A)$ as

$$g(A) = \sum_{j=1}^{t} \|XH_j + E_j\|^2 .$$

We will divide $g(A)$ into three parts.

Part 1: The first part is

$$\sum_{j=1}^{t} \|XH_j\|^2 = \left\| XB^\star \left(U_A U_A^\top - \mathrm{Id}_{t\times t} \right) \right\|_F^2 .$$

Applying Corollary A.8 to $XH_j = Z\Sigma^{1/2} H_j$, we obtain that

$$\begin{aligned} \|XH_j\|^2 &= n \|\Sigma^{1/2} H_j\|^2 \cdot \left[1 + \mathrm{O}_\prec(n^{-1/2} Q) \right] \\ &= n \left\| \Sigma^{1/2} B^\star \left(U_A U_A^\top - \mathrm{Id}_{t\times t} \right) e_j \right\|^2 \cdot \left[1 + \mathrm{O}_\prec(n^{-1/2} Q) \right] . \end{aligned}$$

This implies that

$$\left| \sum_{j=1}^{t} \|XH_j\|^2 - n \left\| \Sigma^{1/2} B^\star \left(U_A U_A^\top - \mathrm{Id}_{t\times t} \right) \right\|_F^2 \right| \prec Q n^{1/2} \left\| \Sigma^{1/2} B^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}) \right\|^2 .$$

Part 2: The second part is the cross term

$$2 \sum_{j=1}^{t} \langle XH_j, E_j \rangle = 2 \langle XB^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}), \mathcal{W} - \mathcal{E} \rangle = -2 \langle XB^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}), \mathcal{E} \rangle .$$

Using (A.20), we obtain that

$$\begin{aligned} &|\langle XB^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}), \mathcal{E} \rangle| \prec \sigma \left\| XB^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}) \right\|_F \\ &\lesssim \sigma n^{1/2} \left\| \Sigma^{1/2} B^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}) \right\|_F \leqslant \sigma^2 n^{1/2} + n^{1/2} \left\| \Sigma^{1/2} B^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}) \right\|_F^2 . \end{aligned}$$

Here in the second step we use the fact that $X = Z\Sigma^{1/2}$ and (C.3), and in the third step we use the AM-GM inequality.

Part 3: The last part is

$$\sum_{j=1}^{t}\|E_j\|^2=\|\mathcal{W}-\mathcal{E}\|_F^2=\|\mathcal{E}\|_F^2-\langle\mathcal{W},\mathcal{E}\rangle, \tag{C.5}$$

where in the second step we use $\|\mathcal{W}\|_F^2=\langle\mathcal{W}A,\mathcal{E}\rangle$ by algebraic calculation. By (A.21), we have that

$$\Big|\|\varepsilon^{(i)}\|^2-\sigma^2 n\Big|\prec\sigma^2 n^{1/2}, \tag{C.6}$$

and

$$\begin{aligned}&\Big|(\varepsilon^{(i)})^\top X(X^\top X)^{-1}X^\top\varepsilon^{(j)}-\delta_{ij}\cdot p\sigma^2\Big|\\&=\Big|(\varepsilon^{(i)})^\top X(X^\top X)^{-1}X^\top\varepsilon^{(j)}-\delta_{ij}\cdot\sigma^2\operatorname{Tr}[X(X^\top X)^{-1}X^\top]\Big|\\&\prec\sigma^2\Big\|X(X^\top X)^{-1}X^\top\Big\|_F=\sigma^2\Big\{\operatorname{Tr}\Big[X(X^\top X)^{-1}X^\top\Big]\Big\}^{1/2}=\sigma^2p^{1/2},\end{aligned} \tag{C.7}$$

where we also use $\operatorname{Tr}[X(X^\top X)^{-1}X^\top]=\operatorname{Tr}[\mathrm{Id}_{p\times p}]=p$ in the above derivation. Summing (C.6) over i, we obtain that

$$\big|\|\mathcal{E}\|_F^2-\sigma^2 nt\big|\leqslant\sum_{i=1}^{t}\Big|\|\varepsilon^{(i)}\|^2-\sigma^2 n\Big|\prec\sigma^2 n^{1/2}. \tag{C.8}$$

Using (C.7), we can estimate the inner product between $\mathcal{W}$ and $\mathcal{E}$ as

$$\begin{aligned}\big|\langle\mathcal{W},\mathcal{E}\rangle-\sigma^2 pr\big|&=\Big|\operatorname{Tr}\Big[\Big(\mathcal{E}^\top U_X U_X^\top\mathcal{E}-p\sigma^2\cdot\mathrm{Id}_{t\times t}\Big)U_AU_A^\top\Big]\Big|\\&\leqslant\Big\|U_AU_A^\top\Big\|_F\cdot\Big\|\mathcal{E}^\top U_XU_X^\top\mathcal{E}-p\sigma^2\cdot\mathrm{Id}_{t\times t}\Big\|\prec\sigma^2n^{1/2}.\end{aligned} \tag{C.9}$$

Combining (C.8) and (C.9), we obtain that

$$\sum_{j=1}^{t}\|E_j\|^2=\sigma^2(nt-pr)+\mathrm{O}_\prec(\sigma^2n^{1/2}).$$

Combining the concentration error estimates for all three parts, we obtain that on event Ξ_1,

$$|g(U_A)-h(U_A)|\prec Qn^{1/2}\Big\|\Sigma^{1/2}B^\star(U_AU_A^\top-\mathrm{Id}_{t\times t})\Big\|_F^2+\sigma^2n^{1/2},$$

for any fixed U_A. Then using a similar ε-net argument as in Appendix B, we can obtain (C.1). We omit the details. □

From (C.2), it is easy to see that the global minimizer of $h(U_A)$ is the best rank-r approximation of $B^{\star\top}\Sigma B^\star$, $A^\star$, defined in (5.5). On the other hand, let $U_{\hat{A}}$ be the global minimizer of $g(U_A)$. We have the following characterization of $U_{\hat{A}}U_{\hat{A}}^\top$ based on Lemma C.1.

LEMMA C.2. *Let In the setting of Theorem 5.1, we have that*

$$\Big\|U_{\hat{A}}U_{\hat{A}}^\top-A^\star A^{\star\top}\Big\|_F^2\lesssim n^{-1/2+c}\frac{Q\|B^{\star\top}\Sigma B^\star\|+\sigma^2}{\lambda_r-\lambda_{r+1}},$$

on the high probability event Ξ in Lemma C.1.

PROOF. Using equation (C.1) and triangle inequality, we upper bound the gap between $h(A^\star)$ and $h(U_{\hat{A}})$ as

$$h(U_{\hat{A}}) - h(A^\star) \leqslant (g(U_{\hat{A}}) - g(A^\star)) + |g(A^\star) - h(A^\star)| + \big|g(U_{\hat{A}}) - h(U_{\hat{A}})\big|$$

$$\leqslant |g(A^\star) - h(A^\star)| + \big|g(U_{\hat{A}}) - h(U_{\hat{A}})\big| \lesssim Qn^{1/2+c}\|\Sigma^{1/2}B^\star\|_F^2 + \sigma^2 n^{1/2+c}, \tag{C.10}$$

on event Ξ. Here in the second step, we use the fact that $\hat{A}$ is the global minimizer of $g(\cdot)$, so that $g(U_{\hat{A}}) \leqslant g(A^\star)$, and in the third step we use equation (C.1), $\|U_A U_A^\top - \mathrm{Id}_{t\times t}\| \leqslant 1$ and $p^{-C} \leqslant Qn^{1/2+c}$. Using the definition of $h(U_A)$ in (C.2), we can check that

$$h(U_{\hat{A}}) - h(A^\star) = n \operatorname{Tr}\left[B^{\star\top}\Sigma B^\star (A^\star A^{\star\top} - U_{\hat{A}}U_{\hat{A}}^\top)\right].$$

For $1 \leqslant i \leqslant t$, let λ_i be the i-th largest eigenvalue of $B^{\star\top}\Sigma B^\star$, and v_i be the corresponding eigenvector. Then we have $A^\star A^{\star\top} = \sum_{i=1}^r v_i v_i^\top$, and

$$\begin{aligned} h(U_{\hat{A}}) - h(A^\star) &= n\sum_{i=1}^r \lambda_i - n\sum_{i=1}^t \lambda_i \|U_{\hat{A}}^\top v_i\|^2 \\ &= n\sum_{i=1}^r \lambda_i \left(1 - \|U_{\hat{A}}^\top v_i\|^2\right) - n\sum_{i=r+1}^t \lambda_i \|U_{\hat{A}}^\top v_i\|^2 \\ &\geqslant n(\lambda_r - \lambda_{r+1}) \sum_{i=r+1}^t \|U_{\hat{A}}^\top v_i\|^2, \end{aligned} \tag{C.11}$$

where in the last step we use that

$$\sum_{i=1}^r \left(1 - \|U_{\hat{A}}^\top v_i\|^2\right) = r - \sum_{i=1}^r \|U_{\hat{A}}^\top v_i\|^2 = \sum_{i=r+1}^t \|U_{\hat{A}}^\top v_i\|^2.$$

From equations (C.10) and (C.11), we obtain that

$$\sum_{i=r+1}^t \|U_{\hat{A}}^\top v_i\|^2 \lesssim \frac{Qn^{-1/2+c}\left\|\Sigma^{1/2}B^\star\right\|_F^2 + n^{-1/2+c}\sigma^2}{\lambda_r - \lambda_{r+1}}. \tag{C.12}$$

On the other hand, we have

$$\left\|A^\star A^{\star\top} - U_{\hat{A}}U_{\hat{A}}^\top\right\|_F^2 = 2r - 2\left\langle A^\star A^{\star\top}, U_{\hat{A}}U_{\hat{A}}{}^\top\right\rangle = 2\sum_{i=r+1}^t \|U_{\hat{A}}^\top v_i\|^2.$$

Combining it with (C.12) and using $\|\Sigma^{1/2}B^\star\|_F^2 \lesssim \|B^{\star\top}\Sigma B^\star\|$, we conclude the proof. □

The last piece of the proof of Theorem 5.1 is the following concentration estimate on the prediction loss of $\hat{\beta}_i^{\mathrm{HPS}}(A)$. Its proof is similar to those of Lemma 2.2 and Lemma C.1.

LEMMA C.3. *In the setting of Theorem 5.1, let*

$$\hat{\beta}_i^{\mathrm{HPS}}(A) \equiv \hat{\beta}_i^{\mathrm{HPS}}(U_A) := \hat{B}(A)A = (X^\top X)^{-1}X^\top Y U_A U_A^\top,$$

and $a_i := U_A U_A^\top e_i$. Then for any small constant $c > 0$ and large constant $C > 0$, there exists a high probability event Ξ, on which the following estimate holds:

$$\begin{aligned} &\left|L_i(\hat{\beta}_i^{\mathrm{HPS}}(U_A)) - L_i(B^\star a_i) - \sigma^2\|a_i\|^2 \cdot \operatorname{Tr}\left[\Sigma(X^\top X)^{-1}\right]\right| \\ &\leqslant n^{-1/2+c}\left[L_i(B^\star a_i) + \sigma^2\|a_i\|^2\right] + p^{-C}\left\|\Sigma^{1/2}B^\star\right\|_F^2, \end{aligned} \tag{C.13}$$

uniformly in all rank-r partial orthonormal matrices $U_A \in \mathbb{R}^{t\times r}$, where L_i is defined in (5.4).

PROOF. We choose a high probability event Ξ_1 on which (C.3) holds. For a fixed U_A, the prediction loss of $\hat{\beta}_i^{\mathrm{HPS}}(A)$ for task i is

$$L(\hat{\beta}_i^{\mathrm{HPS}}(U_A)) = \left\|\Sigma^{1/2}\left((X^\top X)^{-1}X^\top Y a_i - \beta^{(i)}\right)\right\|^2 = \left\|\Sigma^{1/2}(H_i + R_i)\right\|^2,$$

where we denote $R_i = (X^\top X)^{-1}X^\top \mathcal{E} a_i$ and H_i is defined in (C.4). Then the rest of the proof is similar to that of Lemma C.1. We divide the prediction loss into three parts.

Part 1: The first part is the bias term $\|\Sigma^{1/2}H_i\|^2 = L_i(B^\star a_i)$.

Part 2: The second part is the cross term $2\langle \Sigma^{1/2}H_i, \Sigma^{1/2}R_i\rangle$. We can bound it as

$$\begin{aligned}\left|\langle \Sigma^{1/2}H_i, \Sigma^{1/2}R_i\rangle\right| &= \left|\langle X(X^\top X)^{-1}\Sigma H_i, \mathcal{E}a_i\rangle\right| \leqslant \sum_{j=1}^t |a_i(j)| \cdot \left|\langle X(X^\top X)^{-1}\Sigma H_i, \varepsilon^{(j)}\rangle\right| \\ &\prec \sum_{j=1}^t |a_i(j)| \cdot \sigma \left\|X(X^\top X)^{-1}\Sigma H_i\right\| \lesssim \frac{\|a_i\|\sigma}{n^{1/2}}\left\|\Sigma^{1/2}H_i\right\| \\ &\leqslant n^{-1/2}\sigma^2\|a_i\|^2 + n^{-1/2}L_i(B^\star a_i).\end{aligned}$$

Here in the second step we use $a_i(j)$ to denote the j-th coordinate of a_i, in the third step we use (A.20), in the fourth step we use (C.3), (2.5) and $\sum_j |a_i(j)| \leqslant \sqrt{t}\|a_i\|$ by Cauchy-Schwarz inequality, and in the last step we use AM-GM inequality and $\left\|\Sigma^{1/2}H_i\right\|^2 = L_i(B^\star a_i)$.

Part 3: The final part is

$$\begin{aligned}\|\Sigma^{1/2}R_i\|^2 &= \left\|\sum_{j=1}^t a_i(j)\Sigma^{1/2}(X^\top X)^{-1}X^\top \varepsilon^{(j)}\right\|^2 \\ &= \sum_{1\leqslant j,k\leqslant t} a_i(j)a_i(k){\varepsilon^{(j)}}^\top X(X^\top X)^{-1}\Sigma(X^\top X)^{-1}X^\top\varepsilon^{(k)}. \qquad \text{(C.14)}\end{aligned}$$

Using (A.21) and $\mathrm{Tr}[X(X^\top X)^{-1}\Sigma(X^\top X)^{-1}X^\top] = \mathrm{Tr}[\Sigma(X^\top X)^{-1}]$, we obtain that

$$\begin{aligned}&\left|{\varepsilon^{(j)}}^\top X(X^\top X)^{-1}\Sigma(X^\top X)^{-1}X^\top\varepsilon^{(k)} - \delta_{jk}\cdot\sigma^2\,\mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right]\right| \\ &\prec \sigma^2\left\|X(X^\top X)^{-1}\Sigma(X^\top X)^{-1}X^\top\right\|_F = \sigma^2\left\|\Sigma^{1/2}(X^\top X)^{-1}\Sigma^{1/2}\right\|_F \\ &\lesssim \sigma^2 p^{1/2}n^{-1} \leqslant \sigma^2 n^{-1/2}, \qquad \text{(C.15)}\end{aligned}$$

where we use (C.3) in the third step. Plugging (C.15) into (C.14), we obtain that

$$\begin{aligned}\left|\left\|\Sigma^{1/2}R_i\right\|^2 - \sigma^2\|\hat{a}_i\|^2\cdot\mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right]\right| &\prec \sigma^2 n^{-1/2}\sum_{1\leqslant j,k\leqslant t}|\hat{a}_i(j)||\hat{a}_i(k)| \\ &\lesssim n^{-1/2}\sigma^2\|\hat{a}_i\|^2.\end{aligned}$$

Combining the concentration error estimates for all three parts, we obtain that on event Ξ_1,

$$\left|L_i(\hat{\beta}_i^{\mathrm{HPS}}(U_A)) - L_i(B^\star a_i) - \sigma^2\|a_i\|^2\cdot\mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right]\right| \prec n^{-1/2}\left[L_i(B^\star a_i) + \sigma^2\|a_i\|^2\right],$$

for any fixed U_A. Then using a similar ε-net argument as in Appendix B, we can obtain (C.13). We omit the details. □

Provided with Lemmas C.1, C.2 and C.3, we are ready to prove Theorem 5.1.

PROOF OF THEOREM 5.1. The estimate (5.6) follows immediately from Lemma C.2. Using Lemma C.3 and applying Lemma 2.3 to $\mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right]$, we get that for $\hat{a}_i = U_{\hat{A}} U_{\hat{A}}^\top e_i$,

$$\begin{aligned}&\left|L_i(\hat{\beta}_i^{\mathrm{HPS}}(\hat{A})) - L_i(B^\star \hat{a}_i) - \frac{p}{n-p}\cdot \sigma^2 \|\hat{a}_i\|^2\right| \\ &\leqslant n^{-1/2+c}\left[L_i(B^\star \hat{a}_i) + \sigma^2\|\hat{a}_i\|^2\right] + p^{-C}\|\Sigma^{1/2}B^\star\|_F^2, \end{aligned} \tag{C.16}$$

with high probability. From this equation, we obtain that

$$\begin{aligned}&\left|L_i(\hat{\beta}_i^{\mathrm{HPS}}(\hat{A})) - L_i(B^\star a_i^\star) - \frac{p}{n-p}\cdot \sigma^2 \|a_i^\star\|^2\right| \\ &\leqslant n^{-1/2+c}\left[\|\Sigma^{1/2}B^\star\|^2 + \sigma^2\|a_i^\star\|^2\right] + \left(\|\Sigma^{1/2}B^\star\|^2 + \sigma^2\right)\|\hat{a}_i - a_i^\star\|. \end{aligned} \tag{C.17}$$

where we also use that $L_i(B^\star a_i^\star) \leqslant \|\Sigma^{1/2}B^\star\|^2$. On the other hand, using Lemma C.2 we can bound that

$$\|\hat{a}_i - a_i^\star\|^2 \lesssim n^{-1/2+c}\frac{Q\|B^{\star\top}\Sigma B^\star\| + \sigma^2}{\lambda_r - \lambda_{r+1}}.$$

Plugging it into (C.17), we obtain (5.7).

For (5.8), summing (C.16) over i and using Lemma 2.3, we obtain that

$$\begin{aligned}&\left|\sum_i L_i(\hat{\beta}_i^{\mathrm{HPS}}(\hat{A})) - \left\|\Sigma^{1/2}B^\star(U_{\hat{A}}U_{\hat{A}}^\top - \mathrm{Id}_{t\times t})\right\|_F^2 - \frac{p}{n-p}\cdot r\sigma^2\right| \\ &\leqslant n^{-1/2+c}\left[\left\|\Sigma^{1/2}B^\star(U_{\hat{A}}U_{\hat{A}}^\top - \mathrm{Id}_{t\times t})\right\|_F^2 + \sigma^2\right] + p^{-C}\|\Sigma^{1/2}B^\star\|_F^2, \end{aligned} \tag{C.18}$$

with high probability, where we use that $\sum_{i=1}^t \|\hat{a}_i\|^2 = r$ and

$$\sum_{i=1}^t L_i(B^\star \hat{a}_i) = \left\|\Sigma^{1/2}B^\star(U_{\hat{A}}U_{\hat{A}}^\top - \mathrm{Id}_{t\times t})\right\|_F^2.$$

With (C.2), we can write

$$\begin{aligned}\left\|\Sigma^{1/2}B^\star(U_{\hat{A}}U_{\hat{A}}^\top - \mathrm{Id}_{t\times t})\right\|_F^2 &= \frac{h(U_{\hat{A}}) - \sigma^2(nt-pr)}{n} \\ &\geqslant \frac{h(A^\star) - \sigma^2(nt-pr)}{n} = \left\|\Sigma^{1/2}B^\star(A^\star A^{\star\top} - \mathrm{Id}_{t\times t})\right\|_F^2, \end{aligned}$$

where in the second step we use that $A^\star$ is the global minimizer of h. On the other hand, using (C.10), we bound that

$$\frac{h(U_{\hat{A}}) - \sigma^2(nt-pr)}{n} \leqslant \frac{h(A^\star) - \sigma^2(nt-pr)}{n} + \mathrm{O}\left(n^{-1/2+c}\left(Q\|\Sigma^{1/2}B^\star\|^2 + \sigma^2\right)\right)$$

with high probability. Combining the above two estimates, we get

$$\begin{aligned}\left\|\Sigma^{1/2}B^\star(U_{\hat{A}}U_{\hat{A}}^\top - \mathrm{Id}_{t\times t})\right\|_F^2 &= \left\|\Sigma^{1/2}B^\star(A^\star A^{\star\top} - \mathrm{Id}_{t\times t})\right\|_F^2 \\ &\quad + \mathrm{O}\left(n^{-1/2+c}\left(Q\|\Sigma^{1/2}B^\star\|^2 + \sigma^2\right)\right), \end{aligned}$$

with high probability. Plugging it into (C.18), we obtain (5.8). □

D. Proof of Theorem 3.1. In the setting of Theorem 3.1, we have

$$L_{\text{bias}}(a) = \mathbf{v}_a^\top (Z^{(1)})^\top Z^{(1)} \left[a^2 (Z^{(1)})^\top Z^{(1)} + (Z^{(2)})^\top Z^{(2)}\right]^{-2} (Z^{(1)})^\top Z^{(1)} \mathbf{v}_a$$
$$= \mathbf{v}_a^\top \mathcal{Q}^{(1)} \left(a^2 \mathcal{Q}^{(1)} + \frac{n_2}{n_1} \mathcal{Q}^{(2)}\right)^{-2} \mathcal{Q}^{(1)} \mathbf{v}_a,$$

where we abbreviate

$$\mathbf{v}(a) := (\Sigma^{(1)})^{1/2} \left(a\beta^{(1)} - a^2 \beta^{(2)}\right), \quad \mathcal{Q}^{(1)} := \frac{1}{n_1} (Z^{(1)})^\top Z^{(1)}, \quad \mathcal{Q}^{(2)} := \frac{1}{n_2} (Z^{(2)})^\top Z^{(2)}.$$

Note that $\mathcal{Q}^{(1)}$ and $\mathcal{Q}^{(2)}$ are both Wishart matrices. Using the rotational invariance of the distributions of $\mathcal{Q}^{(1)}$ and $\mathcal{Q}^{(2)}$, we can simplify $L_{\text{bias}}(a)$ as in the following lemma.

LEMMA D.1. *In the setting of Theorem 3.1, we have*

$$L_{bias}(a) = \left[1 + \mathrm{O}_\prec(p^{-1/2})\right] \frac{\| \mathbf{v}_a \|^2}{p} \operatorname{Tr}\left[\left(a^2 \mathcal{Q}^{(1)} + \frac{n_2}{n_1} \mathcal{Q}^{(2)}\right)^{-2} (\mathcal{Q}^{(1)})^2\right]. \tag{D.1}$$

PROOF. It is easy to check that the distribution of the matrix $\mathcal{Q}^{(1)}(a^2\mathcal{Q}^{(1)} + \frac{n_2}{n_1}\mathcal{Q}^{(2)})^{-2}\mathcal{Q}^{(1)}$ is rotationally invariant. Hence for the eigenvalue decomposition

$$\mathcal{Q}^{(1)} \left(a^2 \mathcal{Q}^{(1)} + \frac{n_2}{n_1} \mathcal{Q}^{(2)}\right)^{-2} \mathcal{Q}^{(1)} = U^\top \Lambda U,$$

we have that U is a Haar distributed matrix that is independent of Λ. This shows that $\mathbf{u} := U \mathbf{v}_a / \| \mathbf{v}_a \|$ is a unit vector independent of Λ and uniformly distributed on the unit sphere in $\mathbb{R}^p$. Recall that a uniformly distributed unit vector has the same distribution as a normalized Gaussian vector $\mathbf{g}/\|\mathbf{g}\|$, where $\mathbf{g} = (g_1, \cdots, g_p)$ is a random vector with i.i.d. Gaussian entries of mean zero and variance one. Thus we have

$$\mathbf{v}_a^\top \mathcal{Q}^{(1)} \left(a^2 \mathcal{Q}^{(1)} + \frac{n_2}{n_1} \mathcal{Q}^{(2)}\right)^{-2} \mathcal{Q}^{(1)} \mathbf{v}_a \overset{d}{=} \frac{\| \mathbf{v}_a \|^2}{\|\mathbf{g}\|^2} \mathbf{g}^\top \Lambda \mathbf{g}. \tag{D.2}$$

Now using (A.21), we get that

$$\|\mathbf{g}\| = p + \mathrm{O}_\prec(p^{1/2}), \quad |\mathbf{g}^\top \Lambda \mathbf{g} - \operatorname{Tr} \Lambda| \prec \|\Lambda\|_F \prec p^{1/2}, \tag{D.3}$$

where we use Lemma A.6 to bound that $\|\Lambda\|_F \leqslant p^{1/2}\|\Lambda\| \lesssim 1$ with overwhelming probability. Plugging (D.3) into (D.2), we conclude that

$$L_{\text{bias}}(a) = \left[1 + \mathrm{O}_\prec(p^{-1/2})\right] \frac{\| \mathbf{v}_a \|^2}{p} \operatorname{Tr} \Lambda = \left[1 + \mathrm{O}_\prec(p^{-1/2})\right] \frac{\| \mathbf{v}_a \|^2}{p} \operatorname{Tr}[U^\top \Lambda U],$$

which implies (D.1). □

Without loss of generality, we assume that $a \neq 0$, since otherwise we trivially have $L_{\text{bias}}(a) = 0$. Then it is easy to see that

$$L_{\text{bias}}(a) = -\left[1 + \mathrm{O}_\prec(p^{-1/2})\right] \frac{\| \mathbf{v}_a \|^2}{a^4} \cdot \frac{\mathrm{d} f_\alpha(t)}{\mathrm{d}t}\bigg|_{t=0}, \tag{D.4}$$

where $\alpha = n_2/(n_1 a^2)$ and

$$f_\alpha(t) := \frac{1}{p} \operatorname{Tr}\left[\frac{1}{\mathcal{Q}^{(1)} + t(\mathcal{Q}^{(1)})^2 + \alpha \mathcal{Q}^{(2)}}\right].$$

Hence to obtain the asymptotic limit of $L_{\text{bias}}(a)$, we only need to calculate the values of $f_\alpha(t)$ for t around 0.

Our calculation of $f_\alpha(t)$ is based on the Stieltjes transform method in random matrix theory and free additive convolution (or free addition) in free probability theory. We briefly describe the basic concepts that are needed for the proof, and refer the interested readers to classical texts such as [2, 19, 14, 17] for a more thorough introduction. For any probability measure μ supported on $\mathbb{R}$, the Stieltjes transform of μ is a complex function defined as

$$m_\mu(z) := \int_0^\infty \frac{\mathrm{d}\mu(x)}{x-z}, \quad \text{for any } z \in \mathbb{C} \setminus \operatorname{supp}(\mu). \tag{D.5}$$

For any $p \times p$ symmetric matrix M, let $\mu_M := p^{-1}\sum_i \delta_{\lambda_i(M)}$ denote the empirical spectral distribution (ESD) of M, where $\lambda_i(M)$ denotes the i-th eigenvalues of M and $\delta_{\lambda_i(M)}$ is the point mass measure at $\lambda_i(M)$. Then it is easy to see that the Stieltjes transform of μ is equal to

$$m_{\mu_M}(z) := \frac{1}{p}\sum_{i=1}^p \frac{1}{\lambda_i(M)-z} = p^{-1}\operatorname{Tr}\left[(M - z\,\mathrm{Id})^{-1}\right].$$

Given two $p \times p$ matrices A_p and B_p, suppose that their ESD μ_{A_p} and μ_{B_p} converge weakly to two probability distributions, say μ_A and μ_B. Let U_p be a sequence of Haar distributed orthogonal matrices. Then it is known in free probability that the ESD of $A_p + U_p B_p U_p^\top$ converges to the *free addition* of μ_A and μ_B, denoted by $\mu_A \boxplus \mu_B$.

It is well-known that the ESD of $\mathcal{Q}^{(1)}$ and $\mathcal{Q}^{(2)}$ converge weakly to the famous Marchenko-Pastur (MP) law [16]: $\mu_{\mathcal{Q}^{(1)}} \Rightarrow \mu^{(1)}$ and $\mu_{\mathcal{Q}^{(2)}} \Rightarrow \mu^{(2)}$, where $\mu^{(1)}$ and $\mu^{(2)}$ have densities

$$\rho^{(i)}(x) = \frac{1}{2\pi\xi_i x}\sqrt{\left(\lambda_+^{(i)} - x\right)\left(x - \lambda_-^{(i)}\right)}\mathbf{1}_{x\in[\lambda_-^{(i)},\lambda_+^{(i)}]}, \quad i = 1, 2,$$

where recall that $\xi_i = p/n_i$ and $\lambda_\pm^{(i)} := (1 \pm \sqrt{\xi_i})^2$. Moreover, the Stieltjes transforms of $\mu^{(1)}$ and $\mu^{(2)}$ satisfy the self-consistent equations

$$z\xi_i m_{\mu^{(i)}}^2 - (1 - \xi_i - z)m_{\mu^{(i)}} + 1 = 0, \quad i = 1, 2.$$

With this equation, we can check that $g_i(m_{\mu^{(i)}}(z)) = z$ for the functions

$$g_i(m) = \frac{1}{1+\xi_i m} - \frac{1}{m}, \quad i = 1, 2. \tag{D.6}$$

The sharp convergence rates of $\mu_{\mathcal{Q}^{(1)}}$ and $\mu_{\mathcal{Q}^{(2)}}$ has also been obtained in Theorem 3.3 of [18], that is, in the setting of Theorem 3.1 we have

$$d_K\left(\mu_{\mathcal{Q}^{(i)}}, \mu^{(i)}\right) \prec p^{-1}, \quad i = 1, 2, \tag{D.7}$$

where d_K denote the Kolmogorov distance between two probability measures:

$$d_K\left(\mu_{\mathcal{Q}^{(i)}}, \mu^{(i)}\right) := \sum_{x\in\mathbb{R}}\left|\mu_{\mathcal{Q}^{(i)}}\left((-\infty, x]\right) - \mu^{(i)}\left((-\infty, x]\right)\right|.$$

For any fixed $\alpha > 0$ and a sufficiently small $t \geqslant 0$, the ESD of $\alpha\mathcal{Q}^{(2)}$ and $\mathcal{Q}^{(1)} + t(\mathcal{Q}^{(1)})^2$ converges weakly two measures $\mu_\alpha^{(2)}$ and $\mu_t^{(1)}$ defined as follows:

$$\mu_\alpha^{(2)}((-\infty, x]) = \int \mathbf{1}_{\alpha y\in(-\infty,x]}\mathrm{d}\mu^{(2)}(y), \quad \mu_t^{(1)}((-\infty, x]) = \int \mathbf{1}_{y+ty^2\in(-\infty,x]}\mathrm{d}\mu^{(1)}(y).$$

Hence their Stieltjes transforms are given by

$$m_{\mu_\alpha^{(2)}}(z) = \frac{1}{\alpha}m_{\mu^{(2)}}\left(\frac{z}{\alpha}\right), \quad m_{\mu_t^{(1)}}(z) = \int \frac{\mathrm{d}\mu^{(1)}(x)}{x + tx^2 - z}. \tag{D.8}$$

Note that the eigenmatrices of $\mathcal{Q}^{(1)} + t(\mathcal{Q}^{(1)})^2$ and $\alpha\mathcal{Q}^{(2)}$ are independent Haar distributed orthogonal matrices. Hence the ESD of $\mathcal{Q}^{(1)} + t(\mathcal{Q}^{(1)})^2 + \alpha\mathcal{Q}^{(2)}$ converges weakly to the free addition $\mu_t^{(1)} \boxplus \mu_\alpha^{(2)}$. In particular, we will use the following almost sharp estimate on the difference between the Stieltjes transforms of $\mu_{\mathcal{Q}^{(1)}+t(\mathcal{Q}^{(1)})^2+\alpha\mathcal{Q}^{(2)}}$ and $\mu_t^{(1)} \boxplus \mu_\alpha^{(2)}$.

LEMMA D.2. *In the setting of Theorem 3.1, suppose $\alpha, t \in [0, C]$ for a constant $C > 0$. Then we have*

$$\begin{aligned} &\left| \frac{1}{p} \operatorname{Tr} \left[\frac{1}{\mathcal{Q}^{(1)} + t(\mathcal{Q}^{(1)})^2 + \alpha\mathcal{Q}^{(2)} - z} \right] - m_{\mu_t^{(1)} \boxplus \mu_\alpha^{(2)}}(z) \right| \\ &\prec \frac{1}{p} + d_K \left(\mu_{\mathcal{Q}^{(1)}+t(\mathcal{Q}^{(1)})^2}, \mu_t^{(1)} \right) + d_K \left(\mu_{\alpha\mathcal{Q}^{(2)}}, \mu_\alpha^{(2)} \right), \end{aligned} \tag{D.9}$$

for any fixed $z \in \mathbb{C}$ that is away from the support of $\mu_t^{(1)} \boxplus \mu_\alpha^{(2)}$ by a distance of order 1.

PROOF. This lemma is essentially a consequence of Theorem 2.5 of [4] and Theorem 2.4 of [5]. In fact, (D.9) is proved for $z = E + \mathrm{i}\eta$ with $E \in \operatorname{supp}(\mu_t^{(1)} \boxplus \mu_\alpha^{(2)})$ and $\eta > 0$ in [5], but the proofs there can be repeated almost verbatim in our setting where z is away from the support of $\mu_t^{(1)} \boxplus \mu_\alpha^{(2)}$ by a distance of order 1. We omit the details. □

With the above lemma, to calculate the right-hand side of (D.4), it suffices to calculate $\partial_t m_{\mu_t^{(1)} \boxplus \mu_\alpha^{(2)}}(z = 0)$ at $t = 0$.

LEMMA D.3. *We have*

$$\frac{\mathrm{d}}{\mathrm{d}t} m_{\mu_t^{(1)} \boxplus \mu_\alpha^{(2)}}(0) \Big|_{t=0} = -\frac{1 - 2f_1(\alpha)f_3(\alpha) + f_2(\alpha)f_3(\alpha)^2}{1 - \xi_2 f_2(\alpha) f_3(\alpha)^2}, \tag{D.10}$$

where the functions f_1, f_2 and f_3 are defined as

$$\begin{aligned} f_1(\alpha) &:= m_{\mu_0^{(1)} \boxplus \mu_\alpha^{(2)}}(0) \\ &= \frac{2}{\alpha(1-\xi_2) + (1-\xi_1) + \sqrt{[\alpha(1-\xi_2) + (1-\xi_1)]^2 + 4\alpha(\xi_1 + \xi_2 - \xi_1\xi_2)}}, \end{aligned} \tag{D.11}$$

$$f_2(\alpha) := \left[\frac{1}{f_1(\alpha, \xi_1, \xi_2)^2} - \frac{\xi_1}{[1 + \xi_1 f_1(\alpha, \xi_1, \xi_2)]^2} \right]^{-1}, \tag{D.12}$$

$$f_3(\alpha) := \frac{\alpha}{1 + \alpha\xi_2 f_1(\alpha, \xi_1, \xi_2)}. \tag{D.13}$$

PROOF. We calculate the Stieltjes transform of the free addition of $\mu_t^{(1)}$ and $\mu_\alpha^{(2)}$ using the following lemma.

LEMMA D.4 (Theorem 4.1 of [6] and Theorem 2.1 of [8]). *Given two probability measures, μ_1 and μ_2 on $\mathbb{R}$, there exist analytic functions $\omega_1, \omega_2 : \mathbb{C}^+ \to \mathbb{C}^+$, where $C^+ := \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$ is the upper half complex plane, such that the following equations hold: for all $z \in \mathbb{C}^+$,*

$$m_{\mu_1}(\omega_2(z)) = m_{\mu_2}(\omega_1(z)), \quad \omega_1(z) + \omega_2(z) - z = -\frac{1}{m_{\mu_1}(\omega_2(z))}. \tag{D.14}$$

Moreover, $m_{\mu_1}(\omega_2(z))$ is the Stieltjes transform of $\mu_1 \boxplus \mu_2$, that is,

$$m_{\mu_1 \boxplus \mu_2}(z) = m_{\mu_1}(\omega_2(z)).$$

We now solve the equation (D.14) for $\mu_1 = \mu_t^{(1)}$ and $\mu_2 = \mu_\alpha^{(2)}$ for $z \to 0$:

$$m_{\mu_\alpha^{(1)}}(\omega_2(\alpha,t)) = m_{\mu_\alpha^{(2)}}(\omega_1(\alpha,t)), \quad \omega_1(\alpha,t) + \omega_2(\alpha,t) = -\frac{1}{m_{\mu_1}(\omega_2(\alpha,t))}, \tag{D.15}$$

where, for simplicity, we omit the argument $z = 0$ from $\omega_1(z = 0, \alpha, t)$ and $\omega_2(z = 0, \alpha, t)$. From the definition of $m_{\mu_\alpha^{(2)}}$ in (D.8), we can verify that

$$\alpha g_2\left(\alpha m_{\mu_\alpha^{(2)}}(z)\right) = z, \tag{D.16}$$

where g_2 is defined in (D.6). Then applying (D.16) to the first equation of (D.15), we get

$$\omega_1 = \frac{\alpha}{1 + \xi_2 \alpha m_{\mu_t^{(1)}}(\omega_2)} - \frac{1}{m_{\mu_t^{(1)}}(\omega_2)}.$$

Plugging this equation into the second equation of (D.15), we get

$$\frac{\alpha}{1 + \xi_2 \alpha m_{\mu_t^{(1)}}(\omega_2)} + \omega_2 = 0 \;\Leftrightarrow\; \alpha + \omega_2\left[1 + \alpha\xi_2 m_{\mu_t^{(1)}}(\omega_2)\right] = 0. \tag{D.17}$$

This gives a self-consistent equation of $m_{\mu_t^{(1)} \boxplus \mu_\alpha^{(2)}}(0) = m_{\mu_t^{(1)}}(\omega_2)$.

Now we define the following quantities at $t = 0$:

$$f_1(\alpha) := m_{\mu_0^{(1)}}(\omega_2(\alpha,0)), \quad f_3(\alpha) := -\omega_2(\alpha,0),$$

and

$$f_2(\alpha) := \left.\frac{\mathrm{d}m_{\mu_0^{(1)}}(z)}{\mathrm{d}z}\right|_{z=\omega_2(\alpha,0)} = \int \frac{\mathrm{d}\mu^{(1)}(x)}{[x - \omega_2(\alpha,0)]^2}.$$

First, from (D.17) we obtain (D.13). Using the fact that g_1 in (D.6) is the inverse function of $m_{\mu_0^{(1)}}$, we can write equation (D.17) into an equation of f_1 only when $t = 0$:

$$\alpha + \left(\frac{1}{1 + \xi_1 f_1} - \frac{1}{f_1}\right)(1 + \alpha\xi_2 f_1) = 0.$$

This equation can be reduced to a quadratic equation:

$$\alpha\left(\xi_1 + \xi_2 - \xi_1\xi_2\right) f_1^2 + \left[\alpha(1 - \xi_2) + (1 - \xi_1)\right] f_1 - 1 = 0. \tag{D.18}$$

By definition, f_1 is the Stieltjes transform of $\mu_0^{(1)} \boxplus \mu_\alpha^{(2)}$ at $z = 0$. Since $\mu_0^{(1)} \boxplus \mu_\alpha^{(2)}$ is supported on $(0, \infty)$, f_1 must be positive by (D.5). Then it is not hard to see that the only positive solution of (D.18) is given by (D.11). Finally, calculating the derivative of $m_{\mu_0^{(1)}}$ using its inverse function, we obtain that $f_2(\alpha) = [g_1'(f_1)]^{-1}$, which gives (D.12).

To conclude the proof, we still need to calculate $\partial_t m_{\mu_t^{(1)}}(\omega_2)|_{t=0}$. Taking derivative of equation (D.17) with respect to t at $t = 0$, we get

$$\partial_t\omega_2(\alpha,0) \cdot [1 + \alpha\xi_2 f_1(\alpha)] - \alpha\xi_2 f_3(\alpha) \cdot \partial_t m_{\mu_t^{(1)}}(\omega_2(\alpha,t))\Big|_{t=0} = 0. \tag{D.19}$$

Using (D.8), we can calculate that

$$\partial_t m_{\mu_t^{(1)}}(\omega_2(\alpha,t)) = \partial_t \int \frac{\mathrm{d}\mu^{(1)}(x)}{x + tx^2 - \omega_2(\alpha,t)} = -\int \frac{[x^2 - \partial_t\omega(\alpha,t)]\mathrm{d}\mu^{(1)}(x)}{[x + tx^2 - \omega_2(\alpha,t)]^2}.$$

Taking $t = 0$ in the above equation, we get

$$\begin{aligned}
&\partial_t m_{\mu_t^{(1)}}(\omega_2(\alpha,t))\Big|_{t=0} = \partial_t \omega(\alpha,0) f_2(\alpha) \\
&- \int \frac{[(x-\omega_2(\alpha,0))^2 + 2\omega_2(\alpha,0)(x-\omega_2(\alpha,0)) + \omega_2(\alpha,0)^2]\mathrm{d}\mu^{(1)}(x)}{[x-\omega_2(\alpha,0)]^2} \\
&= \partial_t \omega(\alpha,0)\cdot f_2(\alpha) - 1 + 2f_1(\alpha)f_3(\alpha) - f_2(\alpha)f_3(\alpha)^2.
\end{aligned}$$

Then we can solve that

$$\partial_t \omega(\alpha,0) = \frac{1}{f_2(\alpha)}\left[\partial_t m_{\mu_t^{(1)}}(\omega_2(\alpha,t))\Big|_{t=0} + 1 - 2f_1(\alpha)f_3(\alpha) + f_3(\alpha)^2 f_2(\alpha)\right].$$

Inserting it into (D.19), we can solve that

$$\partial_t m_{\mu_t^{(1)}}(\omega_2(\alpha,t))\Big|_{t=0} = -\frac{1 - 2f_1(\alpha)f_3(\alpha) + f_2(\alpha)f_3(\alpha)^2}{1 - \frac{\alpha\xi_2 f_2(\alpha)f_3(\alpha)}{1+\alpha\xi_2 f_1(\alpha)}}.$$

Using $(1+\alpha\xi_2 f_1(\alpha))^{-1} = \alpha^{-1} f_3(\alpha)$ by equation (D.17), we conclude Lemma D.3. □

Now we are ready complete the proof of Theorem 3.1

PROOF OF THEOREM 3.1. We first consider the case $|a| > 1$. We can write the variance term (2.18) as

$$L_{\mathrm{Var}}(a) = \sigma^2 \operatorname{Tr}\left[\frac{1}{n_1 a^2 \mathcal{Q}^{(1)} + n_2 \mathcal{Q}^{(2)}}\right] = \frac{p\sigma^2}{n_1 a^2}\cdot\frac{1}{p}\operatorname{Tr}\left[\frac{1}{\mathcal{Q}^{(1)} + \alpha\mathcal{Q}^{(2)}}\right] \tag{D.20}$$

for $\alpha = n_2/(n_1 a^2)$. Using Lemma D.2 and (D.7), we get

$$\left|\frac{1}{p}\operatorname{Tr}\left[\frac{1}{\mathcal{Q}^{(1)} + \alpha\mathcal{Q}^{(2)}}\right] - m_{\mu_0^{(1)}\boxplus\mu_\alpha^{(2)}}(0)\right| \prec \frac{1}{p}.$$

Recall that $m_{\mu_0^{(1)}\boxplus\mu_\alpha^{(2)}}(0)$ is given by $f_1(\alpha)$ in (D.11). Inserting it into (D.20), we can conclude (3.1) for any fixed $a \in [-1,1]^c$.

For the bias limit, recall that it is given by (D.4). Taking $t = p^{-1/2}$, we can use Lemma A.6 to check that

$$\left|\frac{\mathrm{d}f_\alpha(t)}{\mathrm{d}t}\Big|_{t=0} - \frac{f_\alpha(t) - f_\alpha(0)}{t}\right| \lesssim t \quad \text{w.o.p.}$$

Similarly, we have

$$\left|\frac{\mathrm{d}m_{\mu_t^{(1)}\boxplus\mu_\alpha^{(2)}}(0)}{\mathrm{d}t}\Big|_{t=0} - \frac{m_{\mu_t^{(1)}\boxplus\mu_\alpha^{(2)}}(0) - m_{\mu_0^{(1)}\boxplus\mu_\alpha^{(2)}}(0)}{t}\right| \lesssim t \quad \text{w.o.p.}$$

On the other hand, using Lemma D.2 and (D.7), we get we have

$$\left|f_\alpha(0) - m_{\mu_0^{(1)}\boxplus\mu_\alpha^{(2)}}(0)\right| + \left|f_\alpha(t) - m_{\mu_t^{(1)}\boxplus\mu_\alpha^{(2)}}(0)\right| \prec \frac{1}{p}.$$

Combining the above three estimates, we obtain that

$$\begin{aligned}
\left|\frac{\mathrm{d}f_\alpha(t)}{\mathrm{d}t}\Big|_{t=0} - \frac{\mathrm{d}m_{\mu_t^{(1)}\boxplus\mu_\alpha^{(2)}}(0)}{\mathrm{d}t}\Big|_{t=0}\right| &\prec t + \frac{\left|f_\alpha(0) - m_{\mu_0^{(1)}\boxplus\mu_\alpha^{(2)}}(0)\right| + \left|f_\alpha(t) - m_{\mu_t^{(1)}\boxplus\mu_\alpha^{(2)}}(0)\right|}{t} \\
&\prec p^{-1/2}.
\end{aligned}$$

Now plugging this estimate and (D.10) into (D.4), after a straightforward calculation, we can obtain (3.2) for any fixed $a \in [-1,1]^c$

Finally, using a similar ε-net argument as in Appendix B, we can show that (3.1) and (3.2) hold uniformly for all $a \in [-1,1]^c$ on a high probability event. We omit the details.

It remains to consider the case $|a| \leqslant 1$. In this case, Lemma D.2 cannot be applied when $a \to 0$, in which case $\alpha \to \infty$. However, we can apply Lemma D.2 to

$$\frac{1}{p} \operatorname{Tr}\left[\frac{1}{a^2 \mathcal{Q}^{(1)} + t(\mathcal{Q}^{(1)})^2 + \mathcal{Q}^{(2)} - z}\right],$$

whose limit is given by $m_{\mu^{(1)}_{a^2,t} \boxplus \mu^{(2)}}(z)$, where $\mu^{(1)}_{a^2,t}$ is the limiting ESD of $a^2\mathcal{Q}^{(1)} + t(\mathcal{Q}^{(1)})^2$. Then we can calculate $m_{\mu^{(1)}_{a^2,t} \boxplus \mu^{(2)}}(0)$ and $\partial_t m_{\mu^{(1)}_{a^2,t} \boxplus \mu^{(2)}}(0)$ at $t = 0$ using the same argument as in the proof of Lemma D.3. Finally, repeating the above proof of the $|a| > 1$ case, we can conclude that (3.1) and (3.2) hold uniformly for all $a \in [-1,1]$ on a high probability event. We omit the details. □

E. Proof of Proposition 3.3. In this section, we give the proof of Proposition 3.3. First, for $g(a)$ in equation (2.12), we can calculate its partial expectation over $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ as

$$\begin{aligned}
\mathop{\mathbb{E}}_{\varepsilon^{(1)},\varepsilon^{(2)}}[g(a)] = &\left\| X^{(1)} \hat{\Sigma}(a)^{-1} (X^{(2)})^\top X^{(2)} (\beta^{(1)} - a\beta^{(2)}) \right\|^2 \\
&+ a^2 \left\| X^{(2)} \hat{\Sigma}(a)^{-1} (X^{(1)})^\top X^{(1)} (\beta^{(1)} - a\beta^{(2)}) \right\|^2 \\
&+ \sigma^2 \operatorname{Tr}\left[\left(a^2 X^{(1)} \hat{\Sigma}(a)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1 \times n_1}\right)^2\right] \\
&+ \sigma^2 \operatorname{Tr}\left[\left(X^{(2)} \hat{\Sigma}(a)^{-1} (X^{(2)})^\top - \mathrm{Id}_{n_2 \times n_2}\right)^2\right] \\
&+ a^2\sigma^2 \operatorname{Tr}\left[X^{(2)} \hat{\Sigma}(a)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(a)^{-1} (X^{(2)})^\top\right] \\
&+ a^2\sigma^2 \operatorname{Tr}\left[X^{(1)} \hat{\Sigma}(a)^{-1} (X^{(2)})^\top X^{(2)} \hat{\Sigma}(a)^{-1} (X^{(1)})^\top\right]. \qquad \text{(E.1)}
\end{aligned}$$

Using the following identity

$$\begin{aligned}
(X^{(1)})^\top X^{(1)} \hat{\Sigma}(a)^{-1} (X^{(2)})^\top X^{(2)} &= \left(a^2 [(X^{(2)})^\top X^{(2)}]^{-1} + [(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1} \\
&= (X^{(2)})^\top X^{(2)} \hat{\Sigma}(a)^{-1} (X^{(1)})^\top X^{(1)},
\end{aligned}$$

we can simplify (E.1) to

$$\mathop{\mathbb{E}}_{\varepsilon^{(1)},\varepsilon^{(2)}}[g(a)] = (\beta^{(1)} - a\beta^{(2)})^\top \mathcal{M}(a) (\beta^{(1)} - a\beta^{(2)}) + \sigma^2(n_1 + n_2 - p),$$

where we abbreviate

$$\mathcal{M}(a) := (X^{(1)})^\top X^{(1)} \hat{\Sigma}(a)^{-1} (X^{(2)})^\top X^{(2)}.$$

Furthermore, taking partial expectation over the task-specific components $\widetilde{\beta}^{(1)}$ and $\widetilde{\beta}^{(2)}$ of $\beta^{(1)}$ and $\beta^{(2)}$, we obtain that

$$\mathop{\mathbb{E}}_{\varepsilon^{(1)},\varepsilon^{(2)},\widetilde{\beta}^{(1)},\widetilde{\beta}^{(2)}}[g(a)] = h(a),$$

where

$$h(a) := (a-1)^2 \beta_0^\top \mathcal{M}(a) \beta_0 + (a^2+1)\frac{d^2}{p} \operatorname{Tr}\left[\mathcal{M}(a)\right] + \sigma^2(n_1+n_2-p).$$

Again, using the concentration bounds in Corollary A.9, we can show that $g(a)$ concentrates around $h(a)$ for all $a \in \mathbb{R}$.

CLAIM E.1. *In the setting of Proposition 3.3, for any small constant $c > 0$ and large constant $C > 0$, there exists a high probability event Ξ, on which the following estimates hold uniformly in all $a \in \mathbb{R}$:*

$$|g(a) - h(a)| \leqslant p^{-1/2+c} h(a) + p^{-C}\kappa^2 \|\beta_0\|^2.$$

PROOF. With Corollary A.9, the proof of this claim is very similar to that for Lemma 2.2 in Appendix B. We do not repeat all the arguments here. □

Now we give the proof of Proposition 3.3 based on Claim E.1.

PROOF OF PROPOSITION 3.3. By (2.5), (2.7) and Corollary A.7, there exists a high probability event Ξ_1, on which

$$\begin{aligned} &\lambda_1\left((X^{(1)})^\top X^{(1)}\right) \sim \lambda_p\left((X^{(1)})^\top X^{(1)}\right) \sim n_1 \sim n, \\ &\lambda_1\left((X^{(2)})^\top X^{(2)}\right) \sim \lambda_p\left((X^{(2)})^\top X^{(2)}\right) \sim n_2 \sim n, \end{aligned} \tag{E.2}$$

where we denote $n := n_1 + n_2$. Throughout the following proof, we assume that event Ξ_1 holds. Notice that using (E.2), we can bound

$$h(a) \lesssim n\left[\frac{(a-1)^2}{a^2+1}\|\beta_0\|^2 + \sigma^2 + d^2\right]. \tag{E.3}$$

Let $a^\star$ denote the global minimizer of $h(a)$. Our proof involves two steps: we first show that $|x^\star - 1| \leqslant d^2/\kappa^2$; then we use Claim E.1 to show that the global minimizers of $g(a)$ and $h(a)$ are close to each other.

For the first step, it is easy to observe that $h(a) < h(-a)$ for any positive x. Hence the minimum of $h(a)$ is achieved when a is positive. We first consider the case where $a \geqslant 1$. We define the matrix

$$\mathcal{M}(a) := (X^{(1)})^\top X^{(1)} \hat{\Sigma}(a)^{-1} (X^{(2)})^\top X^{(2)}$$

Notice that for any vector $\mathbf{v} \in \mathbb{R}^p$, the following function is an increasing function in a^2:

$$a^2 \mathbf{v}^\top \mathcal{M}(a)\, \mathbf{v} = \mathbf{v}^\top \left([(X^{(2)})^\top X^{(2)}]^{-1} + a^{-2}[(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1} \mathbf{v}.$$

Hence taking the derivative of $h(a)$, we obtain that

$$h'(a) \geqslant \frac{2(a-1)}{a^3}\beta_0^\top \left[a^2\mathcal{M}(a)\right]\beta_0 - 2\frac{d^2}{a^3}\cdot p^{-1}\operatorname{Tr}\left[a^2\mathcal{M}(a)\right] \tag{E.4}$$

By (E.2), we have

$$\beta_0^\top \left[a^2\mathcal{M}(a)\right]\beta_0 \sim \|\beta_0\|^2 \cdot p^{-1}\operatorname{Tr}\left[a^2\mathcal{M}(a)\right]$$

Hence if $a - 1 \gg d^2/\|\beta_0\|^2$, then $h'(a) > 0$. Then we consider the case where $a \leqslant 1$. Notice that for any vector $\mathbf{v} \in \mathbb{R}^p$, $\mathbf{v}^\top \mathcal{M}(a)\, \mathbf{v}$ is a decreasing function in a^2. Hence taking derivative of $h(a)$, we obtain that

$$h'(a) \leqslant -2(1-a)\beta_0^\top \mathcal{M}(a)\beta_0 + 2ad^2 \cdot p^{-1}\operatorname{Tr}\left[\mathcal{M}(a)\right] < 0, \tag{E.5}$$

if $1-a \gg d^2/\|\beta_0\|^2$. In sum, we see that there exists a constant $C_0>0$, such that

$$h'(a)\begin{cases}>0, & \text{if } a-1 \geqslant C_0 d^2/\|\beta_0\|^2\\ <0, & \text{if } 1-a \geqslant C_0 d^2/\|\beta_0\|^2\end{cases}. \tag{E.6}$$

This gives that $|a^\star - 1| = \mathrm{O}(d^2/\|\beta_0\|^2)$.

For the second step, we show that if $\hat{a}$ deviates too much from $1 \pm C_0 d^2/\|\beta_0\|^2$, then it is no longer a global minimum of $g(a)$. We first argue that $|\hat{a}-1| \leqslant 1$. In fact, if $|\hat{a}-1|>1$, then using (E.2) we can get that

$$h(\hat{a}) - \sigma^2(n-p) \gtrsim n\|\beta_0\|^2.$$

On the other hand, we have $h(1) = \sigma^2(n-p) + \mathrm{O}(nd^2)$. Hence under (3.4), we have

$$h(\hat{a}) - h(1) \gtrsim n\|\beta_0\|^2. \tag{E.7}$$

Then using Claim E.1 and the fact that $g(a^\star) \geqslant g(\hat{a})$, we get

$$\begin{aligned} h(\hat{a}) - h(1) &= [g(\hat{a}) - g(1)] + [h(\hat{a}) - g(\hat{a})] + [g(1) - h(1)] \\ &\leqslant p^{-1/2+c}\,[h(\hat{a}) + h(1)] + p^{-C}\|\beta_0\|^2 \\ &\lesssim p^{-1/2+c} n\left(\|\beta_0\|^2 + d^2 + \sigma^2\right) \ll n\|\beta_0\|^2, \end{aligned}$$

where we use (E.3) in the third step, and condition (3.4) in the last step as long as $c < c_0$. This contradicts (E.7). Hence we conclude that $|\hat{a}-1| \leqslant 1$.

Now suppose that

$$1 + p^{-C_1} + C_1\frac{d^2}{\|\beta_0\|^2} + p^{-1/4+c_1}\frac{d+\sigma}{\|\beta_0\|} \leqslant \hat{a} \leqslant 2, \tag{E.8}$$

for a small constant $c_1 > 0$ and a large constant $C_1 > 0$. Using (E.2) and equation (E.4), it is not hard to check that the following estimate holds for all $a \geqslant 1 + C_1 d^2/(2\|\beta_0\|^2)$ as long as C_1 is large enough:

$$h'(a) \gtrsim \frac{a-1}{a}\beta_0^\top \mathcal{M}(a)\beta_0 \gtrsim n_1\|\beta_0\|^2\frac{a-1}{a(a^2+1)}.$$

Therefore, under (E.8), we have

$$\begin{aligned} h(\hat{a}) - h(a^\star) \geqslant h(\hat{a}) - h\left(1 + \frac{C_1 d^2}{2\|\beta_0\|^2}\right) &\geqslant \int_{1+\frac{C_1 d^2}{2\|\beta_0\|^2}}^{\hat{a}} h'(a)\mathrm{d}a \\ &\gtrsim n\|\beta_0\|^2 \int_{1+\frac{C_1 d^2}{2\|\beta_0\|^2}}^{\hat{a}} \frac{a-1}{a(1+a^2)}\mathrm{d}a \gtrsim n|\hat{a}-1|^2\|\beta_0\|^2 \\ &\gtrsim n\left(|\hat{a}-1|^2\|\beta_0\|^2 + p^{-1/2+2c_1}d^2 + p^{-1/2+2c_1}\sigma^2\right), \end{aligned} \tag{E.9}$$

where we use the condition (3.4) in the last step. On the other hand, using Claim E.1 and the fact that $g(a^\star) \geqslant g(\hat{a})$, we get

$$\begin{aligned} h(\hat{a}) - h(a^\star) &= [g(\hat{a}) - g(a^\star)] + [h(\hat{a}) - g(\hat{a})] + [g(a^\star) - h(a^\star)] \\ &\leqslant p^{-1/2+c}\,[h(\hat{a}) + h(a^\star)] + p^{-C}\|\beta_0\|^2 \lesssim p^{-1/2+c} n\left(|\hat{a}-1|^2\|\beta_0\|^2 + d^2 + \sigma^2\right), \end{aligned}$$

where in the last step we use (E.3) to bound $h(\hat{a})$ and $h(a^\star)$. This contradicts (E.9) as long as $c < 2c_1$. On the other hand, suppose that

$$0 \leqslant \hat{a} \leqslant 1 - p^{-C_1} - C_1\frac{d^2}{\|\beta_0\|^2} - p^{-1/4+c_1}\frac{d+\sigma}{\|\beta_0\|}.$$

Then using equation (E.5) and a similar argument as above, we can also arrive at a contradiction. Hence we have shown that

$$|\hat{a} - 1| \leqslant p^{-C_1} + C_1 \frac{d^2}{\|\beta_0\|^2} + p^{-1/4+c_1} \frac{d+\sigma}{\|\beta_0\|},$$

which completes the proof of (3.5). □

F. Proof of Theorem 3.4 and Proposition 3.6. In this section, we give the proofs of Theorem 3.4 and Proposition 3.6. The central quantity of interest is the inverse of the sum of two sample covariance matrices $\hat{\Sigma}(a)^{-1}$. Assume that $M \equiv M(a) = a(\Sigma^{(1)})^{1/2}(\Sigma^{(2)})^{-1/2}$ has singular value decomposition

$$M = U\Lambda V^\top, \quad \text{where } \Lambda \equiv \Lambda(a) := \mathrm{diag}(\lambda_1(a), \ldots, \lambda_p(a)), \tag{F.1}$$

Then we can write the variance term in (2.18) as

$$\mathrm{Tr}[\Sigma^{(2)}\hat{\Sigma}(a)^{-1}] = \frac{1}{n}\mathrm{Tr}[W(a)^{-1}], \tag{F.2}$$

where we denote $n := n_1 + n_2$ and

$$W(a) := n^{-1}\left(\Lambda(a)U^\top(Z^{(1)})^\top Z^{(1)}U\Lambda(a) + V^\top(Z^{(2)})^\top Z^{(2)}V\right).$$

For $W(a)$, its *resolvent* or *Green's function* is defined as $(W(a) - z\,\mathrm{Id}_{p\times p})^{-1}$ for $z \in \mathbb{C}$. In this section, we will prove a local convergence of this resolvent with a sharp convergence rate, which is conventionally referred to as a "local law of the resolvent" [7, 14, 15].

F.1. *Resolvent and local law.* For $W \equiv W(a)$, we can write that $W = FF^\top$ for a $p \times n$ matrix

$$F := n^{-1/2}[\Lambda U^\top(Z^{(1)})^\top, V^\top(Z^{(2)})^\top]. \tag{F.3}$$

We introduce a convenient self-adjoint linearization trick to study such a matrix. It has been proved to be useful in studying the local laws of random matrices of Gram type [15, 1, 20, 21].

DEFINITION F.1 (Self-adjoint linearization and resolvent). We define the following $(p+n)\times(p+n)$ symmetric block matrix

$$H := \begin{pmatrix} 0 & F \\ F^\top & 0 \end{pmatrix}. \tag{F.4}$$

We define its resolvent as

$$G(z) \equiv G(Z^{(1)}, Z^{(2)}, z) := \left[H - \begin{pmatrix} z\,\mathrm{Id}_{p\times p} & 0 \\ 0 & \mathrm{Id}_{(n_1+n_2)\times(n_1+n_2)} \end{pmatrix}\right]^{-1}, \quad z \in \mathbb{C},$$

as long as the inverse exists. Furthermore, we define the following (weighted) partial traces

$$\begin{aligned} m(z) &:= \frac{1}{p}\sum_{i\in\mathcal{I}_0} G_{ii}(z), \quad m_0(z) := \frac{1}{p}\sum_{i\in\mathcal{I}_0} \lambda_i^2 G_{ii}(z), \\ m_1(z) &:= \frac{1}{n_1}\sum_{\mu\in\mathcal{I}_1} G_{\mu\mu}(z), \quad m_2(z) := \frac{1}{n_2}\sum_{\nu\in\mathcal{I}_2} G_{\nu\nu}(z), \end{aligned} \tag{F.5}$$

where $\mathcal{I}_i$, $i = 0, 1, 2$, are index sets defined as

$$\mathcal{I}_0 := [\![1, p]\!], \quad \mathcal{I}_1 := [\![p+1, p+n_1]\!], \quad \mathcal{I}_2 := [\![p+n_1+1, p+n_1+n_2]\!].$$

REMARK F.2. We will consistently use the latin letters $i, j \in \mathcal{I}_0$ and greek letters $\mu, \nu \in \mathcal{I}_1 \cup \mathcal{I}_2$. Correspondingly, the indices of the matrices $Z^{(1)}$ and $Z^{(2)}$ are labelled as

$$Z^{(1)} = \left[Z^{(1)}_{\mu i} : i \in \mathcal{I}_0, \mu \in \mathcal{I}_1 \right], \quad Z^{(2)} = \left[Z^{(2)}_{\nu i} : i \in \mathcal{I}_0, \nu \in \mathcal{I}_2 \right]. \tag{F.6}$$

Moreover, we define the set of all indices $\mathcal{I} := \mathcal{I}_0 \cup \mathcal{I}_1 \cup \mathcal{I}_2$, and we label the indices in $\mathcal{I}$ as $\mathfrak{a}$, $\mathfrak{b}$, $\mathfrak{c}$ and so on.

Using Schur complement formula for the inverse of a block matrix, it is easy to verify that

$$G(z) = \begin{pmatrix} (W - z\,\mathrm{Id})^{-1} & (W - z\,\mathrm{Id})^{-1} F \\ F^\top (W - z\,\mathrm{Id})^{-1} & z(F^\top F - z\,\mathrm{Id})^{-1} \end{pmatrix}. \tag{F.7}$$

In particular, the upper left block of G is exactly the resolvent of W which we are interested in. Compared with $(W - z\,\mathrm{Id})^{-1}$, it turns out that $G(z)$ is more convenient to deal with because H is a linear function in $Z^{(1)}$ and $Z^{(2)}$. This is why we have chosen to work with $G(z)$.

We define the matrix limit of $G(z)$ as

$$\mathfrak{G}(z) := \begin{pmatrix} [a_1(z)\Lambda^2 + (a_2(z) - z)]^{-1} & 0 & 0 \\ 0 & -\frac{n}{n_1} a_1(z)\,\mathrm{Id}_{n_1 \times n_1} & 0 \\ 0 & 0 & -\frac{n}{n_2} a_2(z)\,\mathrm{Id}_{n_2 \times n_2} \end{pmatrix}, \tag{F.8}$$

where $(a_1(z), a_2(z))$ is the unique solution to the following system of self-consistent equations

$$\begin{aligned} a_1(z) + a_2(z) &= 1 - \frac{1}{n_1 + n_2} \left(\sum_{i=1}^{p} \frac{\lambda_i^2 a_1(z) + a_2(z)}{\lambda_i^2 a_1(z) + a_2(z) - z} \right), \\ a_1(z) &+ \frac{1}{n_1 + n_2} \left(\sum_{i=1}^{p} \frac{\lambda_i^2 a_1(z)}{\lambda_i^2 a_1(z) + a_2(z) - z} \right) = \frac{n_1}{n_1 + n_2}, \end{aligned} \tag{F.9}$$

such that $\operatorname{Im} a_1(z) \leqslant 0$ and $\operatorname{Im} a_2(z) \leqslant 0$ whenever $\operatorname{Im} z > 0$. The existence and uniqueness of solutions to the above system will be shown in Lemma F.5.

We now state the main random matrix result—Theorem F.3—which shows that for z in a small neighborhood around 0, $G(z)$ converges to the limit $\mathfrak{G}(z)$ when p goes to infinity. Moreover, it also gives an almost convergence rate on $G(z)$. Such an estimate is conventionally called the *anisotropic local law* [15]. We define a domain of the spectral parameter z as

$$\mathbf{D} := \left\{ z = E + \mathrm{i}\eta \in \mathbb{C}_+ : |z| \leqslant (\log n)^{-1} \right\}. \tag{F.10}$$

THEOREM F.3. *Suppose $Z^{(1)}$, $Z^{(2)}$, ρ_1 and ρ_2 satisfy Assumption 2.1. Suppose that $Z^{(1)}$ and $Z^{(2)}$ satisfy the bounded support condition* (A.1) *with $Q = n^{2/\varphi}$. Suppose that the singular values of M satisfy that*

$$\lambda_p \leqslant \cdots \leqslant \lambda_2 \leqslant \lambda_1 \leqslant \tau^{-1}. \tag{F.11}$$

Then the following local laws hold on Ξ.

(1) **Averaged local law***: We have*

$$\left| p^{-1} \sum_{i \in \mathcal{I}_0} [G_{ii}(z) - \mathfrak{G}_{ii}(z)] \right| \prec (np)^{-1/2} Q, \tag{F.12}$$

and

$$\left| p^{-1} \sum_{i \in \mathcal{I}_0} \lambda_i^2 [G_{ii}(z) - \mathfrak{G}_{ii}(z)] \right| \prec (np)^{-1/2} Q, \tag{F.13}$$

uniformly in $z \in \mathbf{D}$.

(2) **Anisotropic local law**: *For any deterministic unit vectors* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{p+n}$*, we have*

$$\max_{z\in\mathbf{D}} \left|\mathbf{u}^\top [G(z) - \mathfrak{G}(z)]\mathbf{v}\right| \prec n^{-1/2}Q, \tag{F.14}$$

uniformly in $z \in \mathbf{D}$.

REMARK F.4. Here we state a result that works for the case $|a| \leqslant 1$, where the singular values of M are bounded from above but not from below as in (F.11). To extend the result to the case with $|a| \geqslant 1$, we only need to apply our result to the inverse of

$$\Lambda(a)^{-1}\hat{\Sigma}(a)\Lambda(a)^{-1} = U^\top (Z^{(1)})^\top Z^{(1)} U + \Lambda(a)^{-1} V^\top (Z^{(2)})^\top Z^{(2)} V \Lambda(a)^{-1},$$

so that the eigenvalues of $\Lambda(a)^{-1}$ are bounded from above.

With Theorem F.3, we can complete the proof of Theorem 3.4 and Proposition 3.6 with a standard cutoff argument.

PROOF OF THEOREM 3.4. We introduce the truncated matrices $\widetilde{Z}^{(1)}$ and $\widetilde{Z}^{(2)}$ with entries

$$\widetilde{Z}^{(1)}_{\mu i} := \mathbf{1}\left(|Z^{(1)}_{\mu i}| \leqslant Q\log n\right) \cdot Z^{(1)}_{\mu i}, \quad \widetilde{Z}^{(2)}_{\nu i} := \mathbf{1}\left(|Z^{(2)}_{\nu i}| \leqslant Q\log n\right) \cdot Z^{(2)}_{\nu i}, \tag{F.15}$$

for $Q = n^{2/\varphi}$. From equation (A.2), we get

$$\mathbb{P}(\widetilde{Z}^{(1)} = Z^{(1)}, \widetilde{Z}^{(2)} = Z^{(2)}) = 1 - \mathrm{O}((\log n)^{-\varphi}). \tag{F.16}$$

As in (A.15) and (A.16), we have that

$$\begin{aligned} |\mathbb{E}\widetilde{Z}^{(1)}_{\mu i}| = \mathrm{O}(n^{-3/2}), \quad |\mathbb{E}\widetilde{Z}^{(2)}_{\nu i}| = \mathrm{O}(n^{-3/2}), \\ \mathbb{E}|\widetilde{Z}^{(1)}_{\mu i}|^2 = 1 + \mathrm{O}(n^{-1}), \quad \mathbb{E}|\widetilde{Z}^{(2)}_{\mu i}|^2 = 1 + \mathrm{O}(n^{-1}), \end{aligned} \tag{F.17}$$

and

$$\|\mathbb{E}\widetilde{Z}^{(1)}\| = \mathrm{O}(n^{-1/2}), \quad \|\mathbb{E}\widetilde{Z}^{(2)}\| = \mathrm{O}(n^{-1/2}). \tag{F.18}$$

Then we centralize and rescale $\widetilde{Z}^{(1)}$ and $\widetilde{Z}^{(2)}$ as

$$\widehat{Z}^{(1)} := (\mathbb{E}|\widetilde{Z}^{(1)}_{\mu i}|^2)^{-1/2}(\widetilde{Z}^{(1)} - \mathbb{E}\widetilde{Z}^{(1)}), \quad \widehat{Z}^{(2)} := (\mathbb{E}|\widetilde{Z}^{(2)}_{\mu i}|^2)^{-1/2}(\widetilde{Z}^{(2)} - \mathbb{E}\widetilde{Z}^{(2)}).$$

Now $\widehat{Z}^{(1)}$ and $\widehat{Z}^{(2)}$ satisfy the assumptions of Theorem F.3. Moreover, notice that (F.11) holds when $|a| \leqslant 1$. Hence (F.12) hold for $G(\widehat{Z}^{(1)}, \widehat{Z}^{(2)}, z)$ for any fixed $a \in [-1, 1]$, where $G(\widehat{Z}^{(1)}, \widehat{Z}^{(2)}, z)$ is defined in the same way as $G(z)$ but with $(Z^{(1)}, Z^{(2)})$ replaced by $(\widehat{Z}^{(1)}, \widehat{Z}^{(2)})$. Note that by equations (F.17) and (F.18), we can bound that for $\alpha = 1, 2$,

$$\|\widehat{Z}^{(\alpha)} - \widetilde{Z}^{(\alpha)}\| \lesssim n^{-1}\|\widetilde{Z}^{(\alpha)}\| + \|\mathbb{E}\widetilde{Z}^{(\alpha)}\| \lesssim n^{-1/2} \quad \text{w.o.p.},$$

where we also used Lemma A.6 to bound the operator norm of $\widetilde{Z}^{(\alpha)}$. Together with estimate (F.48) below, this bound implies that w.o.p.,

$$\left|G_{ii}(\widehat{Z}^{(1)}, \widehat{Z}^{(2)}, z) - G_{ii}(\widetilde{Z}^{(1)}, \widetilde{Z}^{(2)}, z)\right| \lesssim n^{-1/2} \sum_{\alpha=1}^{2} \|\widehat{Z}^{(\alpha)} - \widetilde{Z}^{(\alpha)}\| \lesssim n^{-1}, \quad i \in \mathcal{I}_0.$$

Combining this estimate with the local law (F.12) for $G(\widehat{Z}^{(1)}, \widehat{Z}^{(2)}, z)$, we obtain that estimate (F.12) also holds for $G(z)$ on the event $\Xi_1 := \{\widetilde{Z}^{(1)} = Z^{(1)}, \widetilde{Z}^{(2)} = Z^{(2)}\}$.

Now we are ready to prove (3.9) for the case $|a| \leqslant 1$. Then with (F.2), we have

$$L_{\mathrm{Var}}(a) = \frac{\sigma^2}{n} \sum_{i \in \mathcal{I}_0} G_{ii}(0).$$

We also notice that the equations in (F.9) reduce to the equations in (3.10) when $z = 0$, which shows that $a_1 = a_1(0)$ and $a_2 = a_2(0)$. Hence for $\mathfrak{G}$ in (F.8), we have

$$\sum_{i \in \mathcal{I}_0} \mathfrak{G}_{ii}(0) = \mathrm{Tr}\left[\frac{1}{a_1 \Lambda(a)^2 + a_2}\right] = \mathrm{Tr}\left[\frac{1}{a_1 M(a)^\top M(a) + a_2}\right].$$

Now applying (F.12) to $G(z)$ on Ξ_1, we conclude that on Ξ_1,

$$\left| L_{\mathrm{Var}}(a) - \frac{\sigma^2}{n_1 + n_2} \mathrm{Tr}\left[\frac{1}{a_1 M(a)^\top M(a) + a_2}\right] \right| \prec \frac{n^{2/\varphi}}{p^{1/2} n^{1/2}} \cdot \frac{p\sigma^2}{n_1 + n_2},$$

for any fixed $a \in [-1, 1]$. Then using a similar ε-net argument as in Appendix B, we can show that this estimate holds uniformly for all $a \in [-1, 1]$ on a high probability event. We omit the details.

It remains to consider the case $|a| \geqslant 1$. In this case, we can write

$$L_{\mathrm{Var}}(a) = \frac{\sigma^2}{n} \mathrm{Tr}\left[\Lambda(a)^{-2} \widetilde{W}(a)^{-1}\right],$$

where

$$\widetilde{W}(a) := U^\top (Z^{(1)})^\top Z^{(1)} U + \Lambda(a)^{-1} V^\top (Z^{(2)})^\top Z^{(2)} V \Lambda(a)^{-1}.$$

Then we can apply the averaged local law (F.13) to $\widetilde{W}(a)^{-1}$ on Ξ_1, with the role of λ_i replaced by λ_i^{-1}, and the roles of X_1 and X_2 exchanged. This concludes (3.9) for the case $|a| \geqslant 1$. We omit the details. □

PROOF OF PROPOSITION 3.6. We first prove an estimate on

$$\widetilde{L}_{\mathrm{bias}}(a) := n_1^2 \left\| (\Sigma^{(2)})^{1/2} \hat{\Sigma}(a)^{-1} \Sigma^{(1)} \left(a\beta^{(1)} - a^2 \beta^{(2)} \right) \right\|^2.$$

We claim that for any small constant $c > 0$ and large constant $C > 0$, there exists a high probability event Ξ, on which

$$\begin{aligned} &\left| \widetilde{L}_{\mathrm{bias}}(a) - (\beta^{(1)} - a\beta^{(2)})^\top (\Sigma^{(1)})^{1/2} \Pi(a) (\Sigma^{(1)})^{1/2} (\beta^{(1)} - a\beta^{(2)}) \right| \\ &\prec n^{-1/2} Q \left\| (\Sigma^{(1)})^{1/2} \left(\beta^{(1)} - a\beta^{(2)} \right) \right\|^2 + p^{-C} \left[\|\beta^{(1)}\|^2 + \|\beta^{(2)}\|^2 \right], \end{aligned} \tag{F.19}$$

holds uniformly for $a \in \mathbb{R}$.

First, we fix any $a \in [-1, 1]$. Define the vector $\mathbf{v} := V^\top (\Sigma^{(2)})^{-1/2} \Sigma^{(1)} \left(a\beta^{(1)} - a^2 \beta^{(2)} \right) \in \mathbb{R}^p$, and its embedding in $\mathbb{R}^{p+n}$, $\mathbf{w} = (\mathbf{v}^\top, \mathbf{0}_n)^\top$, where $\mathbf{0}_n$ is an n-dimensional zero row vector. Then we have

$$\widetilde{L}_{\mathrm{bias}}(a) = \mathbf{w}^\top \frac{n_1^2}{\left(\Lambda(a) U^\top (Z^{(1)})^\top Z^{(1)} U \Lambda(a) + V^\top (Z^{(2)})^\top Z^{(2)} V \right)^2} \mathbf{w} = \frac{n_1^2}{n^2} \mathbf{w}^\top G'(0) \mathbf{w},$$

where $\mathcal{G}'(0)$ denotes the derivative of $\mathcal{G}(z)$ with respect to z at $z = 0$. Now we introduce the truncated matrices $\widetilde{Z}^{(1)}$ and $\widetilde{Z}^{(2)}$ as in (F.15). Then with a similar argument as in the above proof of Theorem 3.4, we can show that (F.14) holds for $G(z)$ on the event $\Xi_1 := \{\widetilde{Z}^{(1)} =$

$Z^{(1)}, \widetilde{Z}^{(2)} = Z^{(2)}\}$. Now combining (F.14) with Cauchy's integral formula, we get that on Ξ_1,

$$\begin{aligned}\mathbf{w}^\top \mathcal{G}'(0)\,\mathbf{w} = \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{\mathbf{w}^\top \mathcal{G}(z)\,\mathbf{w}}{z^2} \mathrm{d}z = \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{\mathbf{w}^\top \mathfrak{G}(z)\,\mathbf{w}}{z^2} \mathrm{d}z + \mathrm{O}_\prec(n^{-\frac{1}{2}} Q \|\mathbf{w}\|^2) \\ = \mathbf{w}^\top \mathfrak{G}'(0)\,\mathbf{w} + \mathrm{O}_\prec(n^{-\frac{1}{2}} Q \|\mathbf{w}\|^2),\end{aligned} \tag{F.20}$$

where $\mathcal{C}$ is the contour $\{z \in \mathbb{C} : |z| = (\log n)^{-1}\}$. With (F.8), we can calculate the derivative $\mathbf{w}^\top \mathfrak{G}'(0)\,\mathbf{w}$ as

$$\mathbf{w}^\top \mathfrak{G}'(0)\,\mathbf{w} = \mathbf{v}^\top \frac{a_3 \Lambda^2 + (1+a_4)\,\mathrm{Id}_p}{(a_1 \Lambda^2 + a_2\,\mathrm{Id}_p)^2}\,\mathbf{v}, \tag{F.21}$$

where

$$a_3 := -\left.\frac{\mathrm{d}a_1(z)}{\mathrm{d}z}\right|_{z=0}, \quad a_4 := -\left.\frac{\mathrm{d}a_2(z)}{\mathrm{d}z}\right|_{z=0}.$$

Taking derivatives of the system of equations (F.9) with respect to z at $z = 0$, we can derive equation (3.13) for (a_3, a_4). Together with equation (F.20), this concludes the proof of (F.19) for any fixed $a \in [-1, 1]$. Then using a similar ε-net argument as in Appendix B, we can show that (F.19) holds uniformly for all $a \in [-1, 1]$ on a high probability event. We omit the details.

Next for the case $|a| > 1$, we use a similar argument as above, except that we apply (F.14) to the resolvent of $\Lambda(a)^{-1}\hat{\Sigma}(a)\Lambda(a)^{-1}$ instead as discussed in Remark F.4. This will conclude that (F.19) holds uniformly for all $a \in [-1, 1]^c$ on a high probability event. We omit the details.

Now with (F.19), to conclude (3.12) it remains to bound $|L_{\text{bias}}(a) - \widetilde{L}_{\text{bias}}(a)|$:

$$\begin{aligned}&L_{\text{bias}}(a) - \widetilde{L}_{\text{bias}}(a) \\ &= 2n_1 (a\beta^{(1)} - a^2\beta^{(2)})^\top (\Sigma^{(1)})^{1/2} \Delta \left[(\Sigma^{(1)})^{1/2} \hat{\Sigma}(a)^{-1} \Sigma^{(2)} \hat{\Sigma}(a)^{-1} (\Sigma^{(1)})^{1/2}\right] (\Sigma^{(1)})^{1/2} (a\beta^{(1)} - a^2\beta^{(2)}) \\ &+ \left\| (\Sigma^{(2)})^{1/2} \hat{\Sigma}(a)^{-1} (\Sigma^{(1)})^{1/2} \Delta (\Sigma^{(1)})^{1/2} (\beta_1 - a\beta_2) \right\|^2,\end{aligned}$$

where we abbreviate $\Delta = (Z^{(1)})^\top Z^{(1)} - n_1\,\mathrm{Id}_{p\times p}$. From this equation, we get that

$$\begin{aligned}&\left|L_{\text{bias}}(a) - \widetilde{L}_{\text{bias}}(a)\right| \\ &\leqslant a^2 \left[(n_1 + \|\Delta\|)^2 - n_1^2\right] \left\|(\Sigma^{(1)})^{1/2} \hat{\Sigma}(a)^{-1} \Sigma^{(2)} \hat{\Sigma}(a)^{-1} (\Sigma^{(1)})^{1/2}\right\| \left\|(\Sigma^{(1)})^{1/2} (a\beta^{(1)} - a^2\beta^{(2)})\right\|^2,\end{aligned}$$

Using Corollary A.7, we can bound that for any constant $c > 0$,

$$(n_1 + \|\Delta\|)^2 - n_1^2 \leqslant n_1^2 \left[\left(1 + \sqrt{\frac{p}{n_1}}\right)^4 - 1 + n_1^{-1/2+2/\varphi+c}\right],$$

and

$$\left\|(\Sigma^{(2)})^{1/2} \hat{\Sigma}(a)^{-1} (\Sigma^{(2)})^{1/2}\right\| \leqslant \frac{1 + n_1^{-1/2+2/\varphi+c}}{(\sqrt{n_1} - \sqrt{p})^2 \lambda_p^2 + (\sqrt{n_2} - \sqrt{p})^2},$$

with high probability. Combining the above three estimates and using $\|a(\Sigma^{(1)})^{1/2}(\Sigma^{(2)})^{-1/2}\| \leqslant \lambda_1$, we can obtain that with high probability,

$$\begin{aligned}&\left|L_{\text{bias}}(a) - \widetilde{L}_{\text{bias}}(a)\right| \\ &\leqslant \left[\left(1 + \sqrt{\frac{p}{n_1}}\right)^4 - 1 + n_1^{-1/2+2/\varphi+c}\right] \frac{n_1^2 \lambda_1^2 \left\|(\Sigma^{(1)})^{1/2} \left(\beta^{(1)} - a\beta^{(2)}\right)\right\|^2}{[(\sqrt{n_1} - \sqrt{p})^2 \lambda_p^2 + (\sqrt{n_2} - \sqrt{p})^2]^2}.\end{aligned}$$

Together with (F.19), this estimate concludes Proposition 3.6. □

F.2. *Self-consistent equations.* The rest of this section is devoted to the proof of Theorem F.3. In this section, we show that the limiting equation (F.9) has a unique solution $(a_1(z), a_2(z))$ for any $z \in \mathbf{D}$ in equation (F.10). Otherwise, Theorem F.3 will be a vacuous statement. For simplicity of notations, we define the following ratios

$$\gamma_n := \frac{p}{n_1+n_2}, \quad r_1 := \frac{n_1}{n_1+n_2}, \quad r_2 := \frac{n_2}{n_1+n_2}. \tag{F.22}$$

When $z = 0$, the system of equations in (F.9) reduces to the system of equations in (3.10). With (3.10), we can derive an equation of $a_1 \equiv a_1(0)$ only:

$$f(a_1) = r_1, \quad \text{with} \quad f(a_1) := a_1 + \frac{1}{n_1+n_2}\sum_{i=1}^{p} \frac{\lambda_i^2 a_1}{\lambda_i^2 a_1 + (1-\gamma_n-a_1)}. \tag{F.23}$$

We can calculate that

$$f'(a_1) = 1 + \frac{1}{n_1+n_2}\sum_{i=1}^{p} \frac{\lambda_i^2(1-\gamma_n)}{[\lambda_i^2 a_1 + (1-\gamma_n-a_1)]^2} > 0.$$

Hence f is strictly increasing on $[0, 1-\gamma_n]$. Moreover, we have $f(0) = 0 < r_1$, $f(1-\gamma_n) = 1 < r_1$, and $f(r_1) > r_1$ if $r_1 \leqslant 1-\gamma_n$. Hence by mean value theorem, there exists a unique solution a_1 to (F.23) satisfying $0 < a_1 < \min\{1-\gamma_n, r_1\}$. Furthermore, using that $f'(x) = \mathrm{O}(1)$ for any fixed $x \in (0, 1-\gamma_n)$, it is not hard to check that

$$r_1\tau \leqslant a_1 \leqslant \min\{1-\gamma_n, r_1\} \tag{F.24}$$

for a small constant $\tau > 0$. With (3.10), we can also derive a equation of $a_2 \equiv a_2(0)$ only. With a similar argument as above, we can get that for a small constant $\tau > 0$,

$$r_1\tau \leqslant a_2 \leqslant \min\{1-\gamma_n, r_2\}. \tag{F.25}$$

Next, we prove the existence and uniqueness of the solution to the self-consistent equation (F.9) for a general $z \in \mathbf{D}$. For the proof of Theorem F.3, it is better to use the following rescaled functions of $a_1(z)$ and $a_2(z)$:

$$m_{1c}(z) := -r_1^{-1}a_1(z), \quad m_{2c}(z) := -r_2^{-1}a_2(z), \tag{F.26}$$

which, as we will see in (F.90) below, denote the classical values (i.e. asymptotic limits) of $m_1(z)$ and $m_2(z)$. Moreover, it is more convenient to work with the following system of self-consistent equations of $(m_{1c}(z), m_{2c}(z))$, which can be shown to be equivalent to the system of equations (F.9):

$$\begin{aligned} \frac{1}{m_{1c}} &= \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{\lambda_i^2}{z + \lambda_i^2 r_1 m_{1c} + r_2 m_{2c}} - 1, \\ \frac{1}{m_{2c}} &= \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{1}{z + \lambda_i^2 r_1 m_{1c} + r_2 m_{2c}} - 1. \end{aligned} \tag{F.27}$$

When $z = 0$, usings (3.10), (F.24) and (F.25), we get that

$$\tau \leqslant -m_{1c}(0) \leqslant 1, \quad \tau \leqslant -m_{2c}(0) \leqslant 1, \quad -r_1 m_{1c}(0) - r_2 m_{2c}(0) = 1-\gamma_n. \tag{F.28}$$

Now we claim the following lemma, which gives the existence and uniqueness of the solution $(m_{1c}(z), m_{2c}(z))$ to the system of equations (F.27).

LEMMA F.5. *There exist constants $c_0, C_0 > 0$ depending only on τ in Assumption 2.1 and equation* (F.28) *such that the following statements hold. There exists a unique solution $(m_{1c}(z), m_{2c}(z))$ to equation* (F.27) *under the conditions*

$$|z| \leqslant c_0, \quad |m_{1c}(z) - m_{1c}(0)| + |m_{2c}(z) - m_{2c}(0)| \leqslant c_0. \tag{F.29}$$

Moreover, the solution satisfies

$$|m_{1c}(z) - m_{1c}(0)| + |m_{2c}(z) - m_{2c}(0)| \leqslant C_0|z|. \tag{F.30}$$

PROOF. The proof is a standard application of the contraction principle. First, it is easy to check that the system of equations in (F.27) are equivalent to

$$r_1 m_{1c} = -(1-\gamma_n) - r_2 m_{2c} - z\left(m_{2c}^{-1} + 1\right), \quad g_z(m_{2c}(z)) = 1, \tag{F.31}$$

where

$$g_z(m_{2c}) := -m_{2c} + \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{m_{2c}}{z - \lambda_i^2(1-\gamma_n) + (1-\lambda_i^2) r_2 m_{2c} - \lambda_i^2 z\left(m_{2c}^{-1}+1\right)}.$$

We first show that there exists a unique solution $m_{2c}(z)$ to the equation $g_z(m_{2c}(z)) = 1$ under the conditions in equation (F.29). We abbreviate $\delta(z) := m_{2c}(z) - m_{2c}(0)$. From equation (F.31), we obtain that

$$0 = \left[g_z(m_{2c}(z)) - g_0(m_{2c}(0)) - g_z'(m_{2c}(0))\delta(z)\right] + g_z'(m_{2c}(0))\delta(z),$$

which gives that

$$\delta(z) = -\frac{g_z(m_{2c}(0)) - g_0(m_{2c}(0))}{g_z'(m_{2c}(0))} - \frac{g_z(m_{2c}(0)+\delta(z)) - g_z(m_{2c}(0)) - g_z'(m_{2c}(0))\delta(z)}{g_z'(m_{2c}(0))}.$$

Inspired by this equation, we define iteratively a sequence $\delta^{(k)}(z) \in \mathbb{C}$ such that $\delta^{(0)} = 0$, and

$$\begin{aligned} \delta^{(k+1)} = &- \frac{g_z(m_{2c}(0)) - g_0(m_{2c}(0))}{g_z'(m_{2c}(0))} \\ &- \frac{g_z(m_{2c}(0)+\delta^{(k)}) - g_z(m_{2c}(0)) - g_z'(m_{2c}(0))\delta^{(k)}}{g_z'(m_{2c}(0))}. \end{aligned} \tag{F.32}$$

Then equation (F.32) defines a mapping $h_z : \mathbb{C} \to \mathbb{C}$, which maps $\delta^{(k)}$ to $\delta^{(k+1)} = h_z(\delta^{(k)})$.

With direct calculation, we obtain that

$$g_z'(m_{2c}(0)) = -1 - \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{\lambda_i^2(1-\gamma_n) - z\left[1 - \lambda_i^2\left(2m_{2c}^{-1}(0)+1\right)\right]}{\left[z - \lambda_i^2(1-\gamma_n) + (1-\lambda_i^2) r_2 m_{2c}(0) - \lambda_i^2 z\left(m_{2c}^{-1}(0)+1\right)\right]^2}.$$

Using (F.11) and (F.28), it is easy to check that

$$\left|z - \lambda_i^2(1-\gamma_n) + (1-\lambda_i^2) r_2 m_{2c}(0) - \lambda_i^2 z\left(m_{2c}^{-1}(0)+1\right)\right| \geqslant c_\tau - c_\tau^{-1}|z|,$$

for a constant $c_\tau > 0$ depending only on τ. Then as long as we choose $c_0 \leqslant c_\tau^2/2$, we have

$$\left|z - \lambda_i^2(1-\gamma_n) + (1-\lambda_i^2) r_2 m_{2c}(0) - \lambda_i^2 z\left(m_{2c}^{-1}(0)+1\right)\right| \geqslant c_\tau/2. \tag{F.33}$$

With this estimate, it is not hard to check that there exist constants $\widetilde{c}_\tau, \widetilde{C}_\tau > 0$ depending only on τ such that the following estimates hold: for all z, δ_1 and δ_2 such that $|z| \leqslant \widetilde{c}_\tau$, $|\delta_1| \leqslant \widetilde{c}_\tau$ and $|\delta_2| \leqslant \widetilde{c}_\tau$,

$$\left|\frac{1}{g_z'(m_{2c}(0))}\right| \leqslant \widetilde{C}_\tau, \quad \left|\frac{g_z(m_{2c}(0)) - g_0(m_{2c}(0))}{g_z'(m_{2c}(0))}\right| \leqslant \widetilde{C}_\tau|z|, \tag{F.34}$$

and

$$\left|\frac{g_z(m_{2c}(0)+\delta_1)-g_z(m_{2c}(0)+\delta_2)-g_z'(m_{2c}(0))(\delta_1-\delta_2)}{g_z'(m_{2c}(0))}\right| \leqslant \widetilde{C}_\tau|\delta_1-\delta_2|^2. \tag{F.35}$$

Using equations (F.34) and (F.35), we find that there exists a sufficiently small constant $c_1>0$ depending only on $\widetilde{C}_\tau$, such that $h_z: B_d \to B_d$ is a self-mapping on the ball $B_d := \{\delta\in\mathbb{C}: |\delta|\leqslant d\}$, as long as $d\leqslant c_1$ and $|z|\leqslant c_1$. Now it suffices to prove that h_z restricted to B_d is a contraction, which then implies that $\delta := \lim_{k\to\infty}\delta^{(k)}$ exists and $m_{2c}(0)+\delta(z)$ is a unique solution to equation $g_z(m_{2c}(z))=1$ subject to the condition $\|\delta\|_\infty\leqslant d$.

From the iteration relation (F.32), using (F.35) one can readily check that

$$\delta^{(k+1)}-\delta^{(k)} = h_z(\delta^{(k)})-h_z(\delta^{(k-1)}) \leqslant \widetilde{C}_\tau|\delta^{(k)}-\delta^{(k-1)}|^2. \tag{F.36}$$

Hence as long as d is chosen to be sufficiently small such that $2d\widetilde{C}_\tau\leqslant 1/2$, then h is indeed a contraction mapping on B_d. This proves both the existence and uniqueness of the solution $m_{2c}(z)=m_{2c}(0)+\delta(z)$, if we choose c_0 in equation (F.29) as $c_0=\min\{c_\tau^2/2, c_1, (2\widetilde{C}_\tau)^{-1}\}$. After obtaining $m_{2c}(z)$, we can then solve $m_{1c}(z)$ directly using the first equation in (F.31).

Note that with equation (F.34) and $\delta^{(0)}=0$, we can obtain from equation (F.32) that $|\delta^{(1)}(z)|\leqslant\widetilde{C}|z|$. With the contraction mapping, we have the bound

$$|\delta|\leqslant\sum_{k=0}^{\infty}|\delta^{(k+1)}-\delta^{(k)}|\leqslant 2\widetilde{C}_\tau|z| \;\Rightarrow\; |m_{2c}(z)-m_{2c}(0)|\leqslant 2\widetilde{C}_\tau|z|. \tag{F.37}$$

Then using the first equation in equation (F.31), we immediately obtain the bound

$$|m_{1c}(z)-m_{1c}(0)|\leqslant C|z|$$

for a large constant $C>0$. We omit the details. □

As a byproduct of the above contraction mapping argument, we also obtain the following stability result that will be used in the proof of Theorem F.3. Roughly speaking, it states that if two analytic functions $m_1(z)$ and $m_2(z)$ satisfy the self-consistent equation (F.27) approximately up to some small errors, then $m_1(z)$ and $m_2(z)$ will be close to the solutions $m_{1c}(z)$ and $m_{2c}(z)$.

LEMMA F.6. *There exist constants $c_0, C_0>0$ depending only on τ in Assumption 2.1 and equation* (F.28) *such that the system of self-consistent equations in* (F.27) *is stable in the following sense. Suppose $|z|\leqslant c_0$, and m_{1c} and m_{2c} are analytic functions of z such that*

$$|m_1(z)-m_{1c}(0)|+|m_2(z)-m_{2c}(0)|\leqslant c_0. \tag{F.38}$$

Moreover, assume that (m_1,m_2) satisfies the system of equations

$$\begin{aligned}\frac{1}{m_1}+1-\frac{\gamma_n}{p}\sum_{i=1}^{p}\frac{\lambda_i^2}{z+\lambda_i^2 r_1 m_1+r_2 m_2}&=\mathcal{E}_1,\\ \frac{1}{m_2}+1-\frac{\gamma_n}{p}\sum_{i=1}^{p}\frac{1}{z+\lambda_i^2 r_1 m_1+r_2 m_2}&=\mathcal{E}_2,\end{aligned} \tag{F.39}$$

for some (deterministic or random) errors such that $|\mathcal{E}_1|+|\mathcal{E}_2|\leqslant\theta(z)$, where $\theta(z)$ is a deterministic function of z satisfying that $\theta(z)\leqslant(\log n)^{-1}$. Then we have

$$|m_1(z)-m_{1c}(z)|+|m_2(z)-m_{2c}(z)|\leqslant C_0\theta(z). \tag{F.40}$$

PROOF. Under condition (F.38), we can obtain equation (F.31) approximately up to some small errors:

$$r_1 m_1 = -(1-\gamma_n) - r_2 m_2 - z\left(m_2^{-1}+1\right) + \widetilde{\mathcal{E}}_1(z), \quad g_z(m_2(z)) = 1 + \widetilde{\mathcal{E}}_2(z), \tag{F.41}$$

where the errors satisfy that $|\widetilde{\mathcal{E}}_1(z)| + |\widetilde{\mathcal{E}}_2(z)| = \mathrm{O}(\theta(z))$. Then we subtract equation (F.31) from equation (F.41), and consider the contraction principle for the function $\delta(z) := m_2(z) - m_{2c}(z)$. The rest of the proof is exactly the same as the one for Lemma F.5, so we omit the details. □

F.3. *Multivariate Gaussian matrices: Entrywise local law.* The main difficulty in the proof is due to the fact that the entries of $Z^{(1)}U\Lambda$ and $Z^{(2)}V$ are not independent. However, notice that if the entries of $Z^{(1)}$ and $Z^{(2)}$ are i.i.d. Gaussian, then by the rotational invariance of the multivariate Gaussian distribution, we have

$$Z^{(1)}U\Lambda \stackrel{d}{=} Z^{(1)}\Lambda, \quad Z^{(2)}V \stackrel{d}{=} (Z^{(2)}), \tag{F.42}$$

where "$\stackrel{d}{=}$" means "equal in distribution". In this case, the problem is reduced to proving the anisotropic local law for $G(z)$ with $U = \mathrm{Id}$ and $V = \mathrm{Id}$, such that the entries of $Z^{(1)}\Lambda$ and $Z^{(2)}$ are independent. In this case, we use the standard resolvent methods in [7, 21, 18] to prove the following result. Note that if the entries of $Z^{(1)}$ and $Z^{(2)}$ are Gaussian, then we have $\varphi = \infty$ in (2.3), which gives $Q = n^{2/\varphi} = 1$.

PROPOSITION F.7. *In the setting of Theorem F.3 , assume further that the entries of $Z^{(1)}$ and $Z^{(2)}$ are i.i.d. Gaussian random variables. Suppose U and V are identity. Then the estimates* (F.12)*,* (F.13) *and* (F.14) *hold uniformly in $z \in \mathbf{D}$ for $Q = 1$.*

The proof of Proposition F.7 is based on the following entrywise local law.

LEMMA F.8. *In the setting of Proposition F.7, the averaged local laws* (F.12) *and* (F.13)*, and the following entrywise local law hold uniformly in $z \in \mathbf{D}$ for $Q = 1$:*

$$\max_{\mathfrak{a},\mathfrak{b}\in\mathcal{I}} |G_{\mathfrak{a}\mathfrak{b}}(z) - \mathfrak{G}_{\mathfrak{a}\mathfrak{b}}(z)| \prec n^{-1/2}. \tag{F.43}$$

With Lemma F.8, we can complete the proof of Proposition F.7.

PROOF OF PROPOSITION F.7. With estimate (F.43), one can use the polynomialization method in Section 5 of [7] to get the anisotropic local law (F.14) with $Q = 1$. The proof is exactly the same, except for some minor differences in notations. Hence we omit the details. □

The rest of this subsection is devoted to the proof of Lemma F.8. In the setting of Lemma F.8, the resolvent G in Definition F.1 becomes

$$G(z) = \begin{pmatrix} -z\,\mathrm{Id}_{p\times p} & n^{-1/2}\Lambda(Z^{(1)})^\top & n^{-1/2}(Z^{(2)})^\top \\ n^{-1/2}Z^{(1)}\Lambda & -\mathrm{Id}_{n_1\times n_1} & 0 \\ n^{-1/2}Z^{(2)} & 0 & -\mathrm{Id}_{n_2\times n_2} \end{pmatrix}^{-1}. \tag{F.44}$$

To deal with the matrix inverse, we introduce the following resolvent minors of $G(z)$.

DEFINITION F.9 (Resolvent minors). For any $(p+n)\times(p+n)$ matrix $\mathcal{A}$ and $\mathfrak{c}\in\mathcal{I}$, the minor of $\mathcal{A}$ after removing the $\mathfrak{c}$-th row and column of $\mathcal{A}$ is denoted by $\mathcal{A}^{(\mathfrak{c})}:=[\mathcal{A}_{\mathfrak{a}\mathfrak{b}}:\mathfrak{a},\mathfrak{b}\in\mathcal{I}\setminus\{\mathfrak{c}\}]$ as a square matrix of dimension $p+n-1$. Note that we keep the names of indices when defining $\mathcal{A}^{(\mathfrak{c})}$, i.e. $\mathcal{A}^{(\mathfrak{c})}_{\mathfrak{a}\mathfrak{b}}=\mathcal{A}_{\mathfrak{a}\mathfrak{b}}$ for $\mathfrak{a},\mathfrak{b}\neq\mathfrak{c}$. Correspondingly, we define the resolvent minor of $G(z)$ as

$$G^{(\mathfrak{c})}(z):=\left[\begin{pmatrix}-z\,\mathrm{Id}_{p\times p} & n^{-1/2}\Lambda(Z^{(1)})^\top & n^{-1/2}(Z^{(2)})^\top\\ n^{-1/2}Z^{(1)}\Lambda & -\mathrm{Id}_{n_1\times n_1} & 0\\ n^{-1/2}Z^{(2)} & 0 & -\mathrm{Id}_{n_2\times n_2}\end{pmatrix}^{(\mathfrak{c})}\right]^{-1}.$$

We define the partial traces $m^{(\mathfrak{c})}(z)$, $m_0^{(\mathfrak{c})}(z)$, $m_1^{(\mathfrak{c})}(z)$ and $m_2^{(\mathfrak{c})}(z)$ by replacing $G(z)$ with $G^{(\mathfrak{c})}(z)$ in equation (F.5). For convenience, we will adopt the convention that $G^{(\mathfrak{c})}_{\mathfrak{a}\mathfrak{b}}=0$ if $\mathfrak{a}=\mathfrak{c}$ or $\mathfrak{b}=\mathfrak{c}$.

The following resolvent identities are important tools for our proof. All of them can be proved directly using Schur's complement formula, cf. [15, Lemma 4.4].

LEMMA F.10. *We have the following resolvent identities.*

(i) *For $i\in\mathcal{I}_1$ and $\mu\in\mathcal{I}_1\cup\mathcal{I}_2$, we have*

$$\frac{1}{G_{ii}}=-z-\left(FG^{(i)}F^\top\right)_{ii},\quad \frac{1}{G_{\mu\mu}}=-1-\left(F^\top G^{(\mu)}F\right)_{\mu\mu}. \tag{F.45}$$

(ii) *For $i\in\mathcal{I}_1$, $\mu\in\mathcal{I}_1\cup\mathcal{I}_2$, $\mathfrak{a}\in\mathcal{I}\setminus\{i\}$ and $\mathfrak{b}\in\mathcal{I}\setminus\{\mu\}$, we have*

$$G_{i\mathfrak{a}}=-G_{ii}\left(FG^{(i)}\right)_{i\mathfrak{a}},\quad G_{\mu\mathfrak{b}}=-G_{\mu\mu}\left(F^\top G^{(\mu)}\right)_{\mu\mathfrak{b}}. \tag{F.46}$$

(iii) *For $\mathfrak{c}\in\mathcal{I}$ and $\mathfrak{a},\mathfrak{b}\in\mathcal{I}\setminus\{\mathfrak{c}\}$, we have*

$$G^{(\mathfrak{c})}_{\mathfrak{a}\mathfrak{b}}=G_{\mathfrak{a}\mathfrak{b}}-\frac{G_{\mathfrak{a}\mathfrak{c}}G_{\mathfrak{c}\mathfrak{b}}}{G_{\mathfrak{c}\mathfrak{c}}}. \tag{F.47}$$

We claim the following a priori estimate on the resolvent $G(z)$ for $z\in\mathbf{D}$.

LEMMA F.11. *In the setting of Theorem F.3, there exists a constant $C>0$ such that the following estimates hold uniformly in $z,z'\in\mathbf{D}$ with overwhelming probability:*

$$\|G(z)\|\leqslant C, \tag{F.48}$$

and

$$\left\|G(z)-G(z')\right\|\leqslant C|z-z'|. \tag{F.49}$$

PROOF. Our proof is a simple application of the spectral decomposition of G. Recall the matrix F defined in equation (F.3). Let

$$F=\sum_{k=1}^{p}\sqrt{\mu_k}\xi_k\zeta_k^\top,\quad \mu_1\geqslant\mu_2\geqslant\cdots\geqslant\mu_p\geqslant 0=\mu_{p+1}=\ldots=\mu_n, \tag{F.50}$$

be a singular value decomposition of F, where $\{\xi_k\}_{k=1}^p$ are the left-singular vectors and $\{\zeta_k\}_{k=1}^n$ are the right-singular vectors. Then using equation (F.7), we get that for $i,j\in\mathcal{I}_1$ and $\mu,\nu\in\mathcal{I}_1\cup\mathcal{I}_2$,

$$G_{ij}=\sum_{k=1}^{p}\frac{\xi_k(i)\xi_k^\top(j)}{\mu_k-z},\qquad G_{\mu\nu}=z\sum_{k=1}^{n}\frac{\zeta_k(\mu)\zeta_k^\top(\nu)}{\mu_k-z}, \tag{F.51}$$

and

$$G_{i\mu} = G_{\mu i} = \sum_{k=1}^{p} \frac{\sqrt{\mu_k}\xi_k(i)\zeta_k^\top(\mu)}{\mu_k - z}. \tag{F.52}$$

Using the fact $n^{-1}V^\top(Z^{(2)})^\top Z^{(2)}V \preceq FF^\top$ and Lemma A.6, we obtain that

$$\mu_p \geqslant \lambda_p\left(n^{-1}(Z^{(2)})^\top Z^{(2)}\right) \geqslant c_\tau \quad \text{with overwhelming probability,}$$

for a constant $c_\tau > 0$ depending only on τ. This further implies that

$$\inf_{z\in\mathbf{D}} \min_{1\leqslant k\leqslant p} |\mu_k - z| \geqslant c_\tau - (\log n)^{-1}.$$

Combining this bound with (F.51) and (F.52), we can easily conclude (F.48) and (F.49). □

Now we are ready to give the proof of Lemma F.8.

PROOF OF LEMMA F.8. Recall that in the setting of Lemma F.8, we have

$$F \stackrel{d}{=} n^{-1/2}[\Lambda(Z^{(1)})^\top, (Z^{(2)})^\top], \tag{F.53}$$

and it suffices to consider the resolvent in equation (F.44) throughout the whole proof. The proof is divided into four steps.

Step 1: Large deviation estimates. In this step, we prove some sharp large deviation estimates on the off-diagonal entries of G, and on the following $\mathcal{Z}$ variables. In analogy to Section 3 of [11] and Section 5 of [15], we introduce the $\mathcal{Z}$ variables

$$\mathcal{Z}_{\mathfrak{a}} := (1 - \mathbb{E}_{\mathfrak{a}})\left[\left(G_{\mathfrak{a}\mathfrak{a}}\right)^{-1}\right], \quad \mathfrak{a} \in \mathcal{I},$$

where $\mathbb{E}_{\mathfrak{a}}[\cdot] := \mathbb{E}[\cdot \mid H^{(\mathfrak{a})}]$ denotes the partial expectation over the entries in the $\mathfrak{a}$-th row and column of H. Now using equation (F.45), we get that for $i \in \mathcal{I}_0$,

$$\begin{aligned}\mathcal{Z}_i = \frac{\lambda_i^2}{n}\sum_{\mu,\nu\in\mathcal{I}_1} G_{\mu\nu}^{(i)}\left(\delta_{\mu\nu} - Z_{\mu i}^{(1)}Z_{\nu i}^{(1)}\right) + \frac{1}{n}\sum_{\mu,\nu\in\mathcal{I}_2} G_{\mu\nu}^{(i)}\left(\delta_{\mu\nu} - Z_{\mu i}^{(2)}Z_{\nu i}^{(2)}\right) \\ - 2\frac{\lambda_i}{n}\sum_{\mu\in\mathcal{I}_1,\nu\in\mathcal{I}_2} Z_{\mu i}^{(1)}Z_{\nu i}^{(2)}G_{\mu\nu}^{(i)},\end{aligned} \tag{F.54}$$

and for $\mu \in \mathcal{I}_1$ and $\nu \in \mathcal{I}_2$,

$$\mathcal{Z}_\mu = \frac{1}{n}\sum_{i,j\in\mathcal{I}_0} \lambda_i\lambda_j G_{ij}^{(\mu)}\left(\delta_{ij} - Z_{\mu i}^{(1)}Z_{\mu j}^{(1)}\right), \quad \mathcal{Z}_\nu = \frac{1}{n}\sum_{i,j\in\mathcal{I}_0} G_{ij}^{(\nu)}\left(\delta_{ij} - Z_{\nu i}^{(2)}Z_{\nu j}^{(2)}\right). \tag{F.55}$$

Moreover, we introduce the random error

$$\Lambda_o := \max_{\mathfrak{a}\neq\mathfrak{b}} \left|G_{\mathfrak{a}\mathfrak{a}}^{-1}G_{\mathfrak{a}\mathfrak{b}}\right|, \tag{F.56}$$

which controls the size of the off-diagonal entries. The following lemma gives the desired large deviation estimate on Λ_o and $\mathcal{Z}$ variables.

LEMMA F.12. *Under the assumptions of Proposition F.7, the following estimate holds uniformly in all $z \in \mathbf{D}$:*

$$\Lambda_o + \max_{a\in\mathcal{I}} |\mathcal{Z}_a| \prec n^{-1/2}. \tag{F.57}$$

PROOF. Note that for any $\mathfrak{a} \in \mathcal{I}$, $H^{(\mathfrak{a})}$ and $G^{(\mathfrak{a})}$ also satisfy the assumptions in Lemma F.11. Hence the estimates (F.48) and (F.49) also hold for $G^{(\mathfrak{a})}$ with overwhelming probability. For any $i \in \mathcal{I}_0$, since $G^{(i)}$ is independent of the entries in the i-th row and column of H, we can apply (A.8), (A.9) and (A.10) to (F.54) to obtain that

$$|\mathcal{Z}_i| \lesssim \frac{1}{n}\sum_{\alpha=1}^{2}\left|\sum_{\mu,\nu\in\mathcal{I}_\alpha} G^{(i)}_{\mu\nu}\left(\delta_{\mu\nu} - Z^{(1)}_{\mu i}Z^{(1)}_{\nu i}\right)\right| + \frac{1}{n}\left|\sum_{\mu\in\mathcal{I}_1,\nu\in\mathcal{I}_2} Z^{(1)}_{\mu i}Z^{(2)}_{\nu i}G^{(i)}_{\mu\nu}\right|$$
$$\prec \frac{1}{n}\left(\sum_{\mu,\nu\in\mathcal{I}_1\cup\mathcal{I}_2}|G^{(i)}_{\mu\nu}|^2\right)^{1/2} \prec n^{-1/2}.$$

Here in the last step we use (F.48) to get that for any $\mu \in \mathcal{I}_1 \cup \mathcal{I}_2$,

$$\sum_{\nu\in\mathcal{I}_1\cup\mathcal{I}_2}|G^{(i)}_{\mu\nu}|^2 \leqslant \sum_{\mathfrak{a}\in\mathcal{I}}|G^{(i)}_{\mu\mathfrak{a}}|^2 = \left[G^{(i)}(G^{(i)})^*\right]_{\mu\mu} = \mathrm{O}(1), \tag{F.58}$$

with overwhelming probability. Here $(G^{(i)})^*$ denotes the complex conjugate transpose of $G^{(i)}$. Similarly, applying (A.8), (A.9) and (A.10) to $\mathcal{Z}_\mu$ and $\mathcal{Z}_\nu$ in equation (F.55) and using (F.48), we can obtain the same bound. This gives that $\max_{a\in\mathcal{I}}|\mathcal{Z}_a| \prec n^{-1/2}$.

Next we prove the off-diagonal estimate on Λ_o. For $i \in \mathcal{I}_1$ and $\mathfrak{a} \in \mathcal{I}\setminus\{i\}$, using equations (F.46), (A.8) and (F.48), we can obtain that

$$\left|G^{-1}_{ii}G_{i\mathfrak{a}}\right| \lesssim n^{-1/2}\left|\sum_{\mu\in\mathcal{I}_1} Z^{(1)}_{\mu i}G^{(i)}_{\mu\mathfrak{a}}\right| + n^{-1/2}\left|\sum_{\mu\in\mathcal{I}_2} Z^{(2)}_{\mu i}G^{(i)}_{\mu\mathfrak{a}}(z)\right|$$
$$\prec n^{-1/2}\left(\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}|G^{(i)}_{\mu\mathfrak{a}}|^2\right)^{1/2} \prec n^{-1/2}.$$

We can get the same estimate for $\left|G^{-1}_{\mu\mu}G_{\mu\mathfrak{b}}\right|$, $\mu \in \mathcal{I}_1 \cup \mathcal{I}_2$ and $\mathfrak{b} \in \mathcal{I}\setminus\{\mu\}$, using a similar argument. This gives that $\Lambda_o \prec n^{-1/2}$. □

Note that combining (F.57) with the estimate $\max_{\mathfrak{a}}|G_{\mathfrak{a}\mathfrak{a}}| = \mathrm{O}(1)$ w.o.p. by (F.48), we immediately conclude (F.43) for the off-diagonal entries with $a \neq b$.

Step 2: Self-consistent equations. In this step, we show that $(m_1(z), m_2(z))$ satisfies the approximate self-consistent equations in (F.39) for some small errors $\mathcal{E}_1$ and $\mathcal{E}_2$. Later in Step 3, we will apply Lemma F.6 to show that $(m_1(z), m_2(z))$ is close to $(m_{1c}(z), m_{2c}(z))$.

Note that by (F.30), for $z \in \mathbf{D}$ the following estimates hold:

$$|m_{1c}(z) - m_{1c}(0)| \lesssim (\log n)^{-1}, \quad |m_{2c}(z) - m_{2c}(0)| \lesssim (\log n)^{-1}.$$

Together with the estimates in equation (F.28), we obtain that

$$|m_{1c}| \sim |m_{2c}| \sim 1, \quad |z + \lambda_i^2 r_1 m_{1c} + r_2 m_{2c}| \sim 1, \quad \text{uniformly in } z \in \mathbf{D}. \tag{F.59}$$

Moreover, using equation (F.27) we get

$$|1 + \gamma_n m_c(z)| = |m_{2c}^{-1}(z)| \sim 1, \quad |1 + \gamma_n m_{0c}(z)| = |m_{1c}^{-1}(z)| \sim 1, \tag{F.60}$$

uniformly in $z \in \mathbf{D}$, where we abbreviated

$$m_c(z) := -\frac{1}{p}\sum_{i=1}^{p}\frac{1}{z + \lambda_i^2 r_1 m_{1c}(z) + r_2 m_{2c}(z)}, \tag{F.61}$$

$$m_{0c}(z) := -\frac{1}{p}\sum_{i=1}^{p}\frac{\lambda_i^2}{z+\lambda_i^2 r_1 m_{1c}(z)+r_2 m_{2c}(z)}. \tag{F.62}$$

In fact, we will see that $m_c(z)$ and $m_{0c}(z)$ are the asymptotic limits of $m(z)$ and $m_0(z)$, respectively. Applying (F.59) to (F.8) (recall (F.26)), we get that

$$|\mathfrak{G}_{\mathfrak{a}\mathfrak{a}}(z)| \sim 1 \text{ uniformly in } z \in \mathbf{D} \text{ and } \mathfrak{a} \in \mathcal{I}. \tag{F.63}$$

We define the following z-dependent event

$$\Xi(z) := \left\{ |m_1(z) - m_{1c}(0)| + |m_2(z) - m_{2c}(0)| \leqslant (\log n)^{-1/2} \right\}. \tag{F.64}$$

With equation (F.59), we immediately get that on $\Xi(z)$,

$$|m_1(z)| \sim |m_2(z)| \sim 1, \quad |z+\lambda_i^2 r_1 m_1(z) + r_2 m_2(z)| \sim 1. \tag{F.65}$$

Then we prove the following key lemma, which shows that $(m_1(z), m_2(z))$ satisfies equation (F.39) approximately on $\Xi(z)$.

LEMMA F.13. *In the setting of Lemma F.8, the following estimates hold uniformly in* $z \in \mathbf{D}$*:*

$$\mathbf{1}(\Xi)\left|\frac{1}{m_1}+1-\frac{\gamma_n}{p}\sum_{i=1}^{p}\frac{\lambda_i^2}{z+\lambda_i^2 r_1 m_1 + r_2 m_2}\right| \prec n^{-1} + n^{-1/2}\Theta + |[\mathcal{Z}]_0| + |[\mathcal{Z}]_1|, \tag{F.66}$$

and

$$\mathbf{1}(\Xi)\left|\frac{1}{m_2}+1-\frac{\gamma_n}{p}\sum_{i=1}^{p}\frac{1}{z+\lambda_i^2 r_1 m_1 + r_2 m_2}\right| \prec n^{-1} + n^{-1/2}\Theta + |[\mathcal{Z}]| + |[\mathcal{Z}]_2|, \tag{F.67}$$

where we denote

$$\Theta := |m_1(z) - m_{1c}(z)| + |m_2(z) - m_{2c}(z)|, \tag{F.68}$$

and

$$[\mathcal{Z}] := \frac{1}{p}\sum_{i\in\mathcal{I}_0}\frac{\mathcal{Z}_i}{(z+\lambda_i^2 r_1 m_{1c} + r_2 m_{2c})^2}, \quad [\mathcal{Z}]_0 := \frac{1}{p}\sum_{i\in\mathcal{I}_0}\frac{\lambda_i^2\mathcal{Z}_i}{(z+\lambda_i^2 r_1 m_{1c} + r_2 m_{2c})^2}, \tag{F.69}$$

$$[\mathcal{Z}]_1 := \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\mathcal{Z}_\mu, \quad [\mathcal{Z}]_2 := \frac{1}{n_2}\sum_{\nu\in\mathcal{I}_2}\mathcal{Z}_\nu. \tag{F.70}$$

PROOF. With equations (F.45), (F.54) and (F.55), we obtain that

$$\begin{aligned}\frac{1}{G_{ii}} &= -z - \frac{\lambda_i^2}{n}\sum_{\mu\in\mathcal{I}_1} G_{\mu\mu}^{(i)} - \frac{1}{n}\sum_{\mu\in\mathcal{I}_2} G_{\mu\mu}^{(i)} + \mathcal{Z}_i \\ &= -z - \lambda_i^2 r_1 m_1 - r_2 m_2 + \mathcal{E}_i, \quad \text{for } i \in \mathcal{I}_0, \end{aligned} \tag{F.71}$$

$$\frac{1}{G_{\mu\mu}} = -1 - \frac{1}{n}\sum_{i\in\mathcal{I}_0}\lambda_i^2 G_{ii}^{(\mu)} + \mathcal{Z}_\mu = -1 - \gamma_n m_0 + \mathcal{E}_\mu, \quad \text{for } \mu \in \mathcal{I}_1, \tag{F.72}$$

$$\frac{1}{G_{\nu\nu}} = -1 - \frac{1}{n}\sum_{i\in\mathcal{I}_0} G_{ii}^{(\nu)} + \mathcal{Z}_\nu = -1 - \gamma_n m + \mathcal{E}_\nu, \quad \text{for } \nu \in \mathcal{I}_2, \tag{F.73}$$

where we denote (recall (F.5) and Definition F.9)

$$\mathcal{E}_i := \mathcal{Z}_i + \lambda_i^2 r_1 \left(m_1 - m_1^{(i)}\right) + r_2 \left(m_2 - m_2^{(i)}\right),$$

and

$$\mathcal{E}_\mu := \mathcal{Z}_\mu + \gamma_n(m_0 - m_0^{(\mu)}), \quad \mathcal{E}_\nu := \mathcal{Z}_\nu + \gamma_n(m - m^{(\nu)}).$$

Using equations (F.47), (F.56) and (F.57), we can bound that

$$|m_1 - m_1^{(i)}| \leqslant \frac{1}{n_1} \sum_{\mu \in \mathcal{I}_1} \left| \frac{G_{\mu i} G_{i\mu}}{G_{ii}} \right| \leqslant |\Lambda_o|^2 |G_{ii}| \prec n^{-1}. \tag{F.74}$$

where we also use bound (F.48) in the last step. Similarly, we can also obtain that

$$|m_2 - m_2^{(i)}| \prec n^{-1}, \quad |m_0 - m_0^{(\mu)}| \prec n^{-1}, \quad |m - m^{(\nu)}| \prec n^{-1}, \tag{F.75}$$

for any $i \in \mathcal{I}_0$, $\mu \in \mathcal{I}_1$ and $\nu \in \mathcal{I}_2$. Together with (F.57), we obtain the bound

$$\max_{i \in \mathcal{I}_0} |\mathcal{E}_i| + \max_{\mu \in \mathcal{I}_1 \cup \mathcal{I}_2} |\mathcal{E}_\mu| \prec n^{-1/2}. \tag{F.76}$$

From equation (F.71), we obtain that on Ξ,

$$\begin{aligned} G_{ii} &= -\frac{1}{z + \lambda_i^2 r_1 m_1 + r_2 m_2} - \frac{\mathcal{E}_i}{(z + \lambda_i^2 r_1 m_1 + r_2 m_2)^2} + \mathrm{O}_\prec\left(n^{-1}\right) \\ &= -\frac{1}{z + \lambda_i^2 r_1 m_1 + r_2 m_2} - \frac{\mathcal{Z}_i}{(z + \lambda_i^2 r_1 m_{1c} + r_2 m_{2c})^2} + \mathrm{O}_\prec\left(n^{-1} + n^{-1/2}\Theta\right). \end{aligned} \tag{F.77}$$

where in the first step we use (F.76) and (F.65) on Ξ, and in the second step we use (F.68), (F.74) and (F.75). Plugging (F.77) into the definitions of m and m_0 in (F.5) and using (F.69), we get that on Ξ,

$$m = -\frac{1}{p} \sum_{i \in \mathcal{I}_0} \frac{1}{z + \lambda_i^2 r_1 m_1 + r_2 m_2} - [\mathcal{Z}] + \mathrm{O}_\prec\left(n^{-1} + n^{-1/2}\Theta\right), \tag{F.78}$$

$$m_0 = -\frac{1}{p} \sum_{i \in \mathcal{I}_0} \frac{\lambda_i^2}{z + \lambda_i^2 r_1 m_1 + r_2 m_2} - [\mathcal{Z}]_0 + \mathrm{O}_\prec\left(n^{-1} + n^{-1/2}\Theta\right). \tag{F.79}$$

As a byproduct, comparing these two equations with (F.61) and (F.62), we obtain that

$$|m(z) - m_c(z)| + |m_0(z) - m_{0c}(z)| \lesssim (\log n)^{-1/2}, \quad \text{w.o.p. on } \Xi. \tag{F.80}$$

Together with equation (F.60), we get that

$$|1 + \gamma_n m(z)| \sim 1, \quad |1 + \gamma_n m_0(z)| \sim 1, \quad \text{w.o.p. on } \Xi. \tag{F.81}$$

Now with a similar argument as above, from equations (F.72), (F.73), we obtain that on Ξ,

$$G_{\mu\mu} = -\frac{1}{1 + \gamma_n m_0} - \frac{\mathcal{Z}_\mu}{(1 + \gamma_n m_0)^2} + \mathrm{O}_\prec\left(n^{-1} + n^{-1/2}\Theta\right), \quad \mu \in \mathcal{I}_1, \tag{F.82}$$

$$G_{\nu\nu} = -\frac{1}{1 + \gamma_n m} - \frac{\mathcal{Z}_\nu}{(1 + \gamma_n m)^2} + \mathrm{O}_\prec\left(n^{-1} + n^{-1/2}\Theta\right), \quad \nu \in \mathcal{I}_2, \tag{F.83}$$

where we use (F.76), (F.81), (F.68), (F.74) and (F.75) in the derivation. Taking average of (F.82) and (F.83) over $\mu \in \mathcal{I}_1$ and $\nu \in \mathcal{I}_2$, we get that on Ξ,

$$m_1 = -\frac{1}{1 + \gamma_n m_0} - \frac{[\mathcal{Z}]_1}{(1 + \gamma_n m_0)^2} + \mathrm{O}_\prec\left(n^{-1} + n^{-1/2}\Theta\right), \tag{F.84}$$

$$m_2 = -\frac{1}{1 + \gamma_n m} - \frac{[\mathcal{Z}]_2}{(1 + \gamma_n m)^2} + \mathrm{O}_\prec\left(n^{-1} + n^{-1/2}\Theta\right), \tag{F.85}$$

which further implies that on Ξ,

$$\frac{1}{m_1} + 1 + \gamma_n m_0 \prec n^{-1} + n^{-1/2}\Theta + |[\mathcal{Z}]_1|, \tag{F.86}$$

$$\frac{1}{m_2} + 1 + \gamma_n m \prec n^{-1} + n^{-1/2}\Theta + |[\mathcal{Z}]_2|. \tag{F.87}$$

Finally, plugging (F.78) and (F.79) into equations (F.86) and (F.87), we conclude equations (F.66) and (F.67). □

Step 3: Entrywise local law. In this step, we show that the event $\Xi(z)$ in (F.64) actually holds with overwhelming probability for all $z \in \mathbf{D}$. Once we have proved this fact, applying Lemma F.6 to equations (F.66) and (F.67) immediately shows that $(m_1(z), m_2(z))$ is close to $(m_{1c}(z), m_{2c}(z))$ up to an error of order $\mathrm{O}_\prec(n^{-1/2})$, with which we can conclude the entrywise local law (F.43).

We claim that it suffices to show that

$$|m_1(0) - m_{1c}(0)| + |m_2(0) - m_{2c}(0)| \prec n^{-1/2}. \tag{F.88}$$

In fact, notice that by (F.30) and (F.49) we have

$$|m_{1c}(z) - m_{1c}(0)| + |m_{2c}(z) - m_{2c}(0)| = \mathrm{O}((\log n)^{-1}),$$

and

$$|m_1(z) - m_1(0)| + |m_2(z) - m_2(0)| = \mathrm{O}((\log n)^{-1}),$$

with overwhelming probability for all $z \in \mathbf{D}$. Thus if (F.88) holds, using triangle inequality we can obtain from the above two estimates that

$$\sup_{z\in\mathbf{D}} \left(|m_1(z) - m_{1c}(0)| + |m_2(z) - m_{2c}(0)|\right) \lesssim (\log n)^{-1} \quad \text{w.o.p.} \tag{F.89}$$

The equation (F.89) shows that $\Xi(z)$ holds with overwhelming probability, and it also verifies the condition (F.38) of Lemma F.6. Now applying Lemma F.6 to equations (F.66) and (F.67), we obtain that

$$\begin{aligned}\Theta(z) &= |m_1(z) - m_{1c}(z)| + |m_2(z) - m_{2c}(z)| \\ &\prec n^{-1} + n^{-1/2}\Theta(z) + |[\mathcal{Z}]| + |[\mathcal{Z}]_0| + |[\mathcal{Z}]_1| + |[\mathcal{Z}]_2|,\end{aligned}$$

which implies that

$$\Theta(z) \prec n^{-1} + |[\mathcal{Z}]| + |[\mathcal{Z}]_0| + |[\mathcal{Z}]_1| + |[\mathcal{Z}]_2| \prec n^{-1/2}, \tag{F.90}$$

uniformly for all $z \in \mathbf{D}$. Here in the second step, we use (F.57). On the other hand, with equations (F.82)-(F.85), we obtain that

$$\max_{\mu\in\mathcal{I}_1} |G_{\mu\mu}(z) - m_1(z)| + \max_{\nu\in\mathcal{I}_2} |G_{\nu\nu}(z) - m_2(z)| \prec n^{-1/2}.$$

Combining this estimate with (F.90), we get that

$$\max_{\mu\in\mathcal{I}_1} |G_{\mu\mu}(z) - m_{1c}(z)| + \max_{\nu\in\mathcal{I}_2} |G_{\nu\nu}(z) - m_{2c}(z)| \prec n^{-1/2}. \tag{F.91}$$

Then plugging (F.90) into equation (F.77) and recalling (F.26), we obtain that

$$\max_{i\in\mathcal{I}_1} |G_{ii}(z) - \mathfrak{G}_{ii}(z)| \prec n^{-1/2}.$$

Together with equation (F.91), it gives the diagonal estimate

$$\max_{\mathfrak{a}\in\mathcal{I}} |G_{\mathfrak{a}\mathfrak{a}}(z) - \Pi_{\mathfrak{a}\mathfrak{a}}(z)| \prec n^{-1/2}. \tag{F.92}$$

Combining equation (F.92) with the off-diagonal estimate on Λ_o in equation (F.57), we conclude the entrywise local law (F.43).

Now we give the proof of (F.88). Using (F.48) and (F.51), we get that with overwhelming probability,

$$1 \gtrsim m(0) = \frac{1}{p}\sum_{i\in\mathcal{I}_0} G_{ii}(0) = \frac{1}{p}\sum_{i\in\mathcal{I}_0}\sum_{k=1}^{p} \frac{|\xi_k(i)|^2}{\mu_k} \geqslant \mu_1^{-1} \gtrsim 1,$$

where we use Lemma A.6 to bound μ_1. Similarly, we can also get that $m_0(0)$ is positive and has size $m_0(0) \sim 1$. Hence we have the estimates

$$1 + \gamma_n m(0) \sim 1, \quad 1 + \gamma_n m_0(0) \sim 1. \tag{F.93}$$

Combining these estimates with equations (F.72), (F.73) and (F.76), we obtain that (F.84) and (F.85) hold at $z = 0$ even without the indicator function $\mathbf{1}(\Xi)$, which further give that with overwhelming probability,

$$\left|\lambda_i^2 r_1 m_1(0) + r_2 m_2(0)\right| = \left|\frac{\lambda_i^2 r_1}{1+\gamma_n m_0(0)} + \frac{r_2}{1+\gamma_n m(0)} + \mathrm{O}_{\prec}(n^{-1/2})\right| \sim 1.$$

Then combining this estimate with (F.71) and (F.76), we obtain that (F.78) and (F.79) also hold at $z = 0$ even without the indicator function $\mathbf{1}(\Xi)$. Finally, plugging (F.78) and (F.79) into equations (F.86) and (F.87), we conclude that (F.66) and (F.67) hold at $z = 0$, that is,

$$\begin{aligned} &\left|\frac{1}{m_1(0)} + 1 - \frac{1}{n}\sum_{i=1}^{p} \frac{\lambda_i^2}{\lambda_i^2 r_1 m_1(0) + r_2 m_2(0)}\right| \prec n^{-1/2}, \\ &\left|\frac{1}{m_2(0)} + 1 - \frac{1}{n}\sum_{i=1}^{p} \frac{1}{\lambda_i^2 r_1 m_1(0) + r_2 m_2(0)}\right| \prec n^{-1/2}. \end{aligned} \tag{F.94}$$

Denoting $y_1 = -m_1(0)$ and $y_2 = -m_2(0)$, by (F.84) and (F.85) at $z = 0$ we have

$$y_1 = \frac{1}{1+\gamma_n m_0(0)} + \mathrm{O}_{\prec}(n^{-1/2}), \quad y_2 = \frac{1}{1+\gamma_n m(0)} + \mathrm{O}_{\prec}(n^{-1/2}).$$

Hence by (F.93), there exists a constant $c > 0$ such that

$$c \leqslant y_1 \leqslant 1, \quad c \leqslant y_2 \leqslant 1, \quad \text{with overwhelming probability.} \tag{F.95}$$

Also one can verify from equation (F.94) that $(r_1 y_1, r_2 y_2)$ satisfies approximately the same system of equations as equation (3.10):

$$r_1 y_1 + r_2 y_2 = 1 - \gamma_n + \mathrm{O}_{\prec}(n^{-1/2}), \quad f(r_1 y_1) = r_1 + \mathrm{O}_{\prec}(n^{-1/2}), \tag{F.96}$$

where recall that the function f was defined in equation (F.23). The first equation of (F.96) and equation (F.95) together imply that $y_1 \in [0, r_1^{-1}(1-\gamma_n)]$ with overwhelming probability. For the second equation of (F.96), we know that $y_1 = r_1^{-1} a_1$ is a solution. Moreover, it is easy to check that the function $f(r_1 y_1)$ is strictly increasing and has bounded derivative on $[0, r_1^{-1}(1-\gamma_n)]$. So by basic calculus, we obtain that

$$|m_1(0) - m_{1c}(0)| = |y_1 - r_1^{-1} a_1| \prec n^{-1/2}.$$

Plugging it into the first equation of (F.96), we get

$$|m_2(0) - m_{2c}(0)| = |y_2 - r_2^{-1}a_2| \prec n^{-1/2}.$$

The above two estimates conclude (F.88).

Step 4: Averaged local law. Finally, we prove the averaged local laws (F.12) and (F.13). For this purpose, we need to use the following *fluctuation averaging estimate*.

LEMMA F.14 (Fluctuation averaging). *In the setting of Lemma F.7, suppose the entrywise local law* (F.43) *holds uniformly in* $z \in \mathbf{D}$. *Then we have that*

$$|[\mathcal{Z}]| + |[\mathcal{Z}]_0| + |[\mathcal{Z}]_1| + |[\mathcal{Z}]_2| \prec (np)^{-1/2}, \tag{F.97}$$

uniformly in $z \in \mathbf{D}$.

PROOF. The proof is the same as the one for Theorem 4.7 of [13]. □

Plugging (F.90) and (F.97) into equations (F.66) and (F.67), and applying Lemma F.6, we obtain that

$$|m_1(z) - m_{1c}(z)| + |m_2(z) - m_{2c}(z)| \prec (np)^{-1/2}. \tag{F.98}$$

Now subtracting (F.61) from (F.78), and using (F.97) and (F.98), we obtain that

$$|m(z) - m_c(z)| \prec (np)^{-1/2}.$$

This is exactly the averaged local law (F.12) with $Q = 1$. The proof of (F.13) is similar. □

F.4. *Anisotropic local law.* In this section, we prove the anisotropic local law in Theorem F.3 by extending from the multivariate Gaussian random matrices to generally distributed random matrices. With Proposition F.7, it suffices is to prove that for $Z^{(1)}$ and $Z^{(2)}$ satisfying the assumptions in Theorem F.3, we have

$$\left|\mathbf{u}^\top \left(G(Z,z) - G(Z^{\text{Gauss}},z)\right)\mathbf{v}\right| \prec n^{-1/2}Q$$

for any deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{p+n}$ and $z \in \mathbf{D}$, where we abbreviated that

$$Z := \begin{pmatrix} Z^{(1)} \\ Z^{(2)} \end{pmatrix}, \quad \text{and} \quad Z^{\text{Gauss}} := \begin{pmatrix} (Z^{(1)})^{\text{Gauss}} \\ (Z^{(2)})^{\text{Gauss}} \end{pmatrix}.$$

Here $(Z^{(1)})^{\text{Gauss}}$ and $(Z^{(2)})^{\text{Gauss}}$ are Gaussian random matrices satisfying the assumptions in Proposition F.7. We will prove the above statement using a continuous comparison argument developed in [15]. Since the arguments are similar to the one in Sections 7 and 8 of [15] and Section 6 of [21], we will not write down all the details.

We define the following continuous sequence of interpolating matrices between Z^{Gauss} and Z.

DEFINITION F.15 (Interpolation). We denote $Z^0 := Z^{\text{Gauss}}$ and $Z^1 := Z$. Let $\rho^0_{\mu i}$ and $\rho^1_{\mu i}$ be the laws of $Z^0_{\mu i}$ and $Z^1_{\mu i}$, respectively, for $i \in \mathcal{I}_0$ and $\mu \in \mathcal{I}_1 \cup \mathcal{I}_2$. For any $\theta \in [0,1]$, we define the interpolated law $\rho^\theta_{\mu i} := (1-\theta)\rho^0_{\mu i} + \theta\rho^1_{\mu i}$. We shall work on the probability space consisting of triples (Z^0, Z^θ, Z^1) of independent $n \times p$ random matrices, where the matrix $Z^\theta = (Z^\theta_{\mu i})$ has law

$$\prod_{i \in \mathcal{I}_0} \prod_{\mu \in \mathcal{I}_1 \cup \mathcal{I}_2} \rho^\theta_{\mu i}(\mathrm{d}Z^\theta_{\mu i}). \tag{F.99}$$

For $\lambda \in \mathbb{R}$, $i \in \mathcal{I}_0$ and $\mu \in \mathcal{I}_1 \cup \mathcal{I}_2$, we define the matrix $Z^{\theta,\lambda}_{(\mu i)}$ through

$$\left(Z^{\theta,\lambda}_{(\mu i)}\right)_{\nu j} := \begin{cases} Z^{\theta}_{\mu i}, & \text{if } (j,\nu) \neq (i,\mu) \\ \lambda, & \text{if } (j,\nu) = (i,\mu) \end{cases},$$

that is, it replaces the (μ, i)-th entry of Z^θ with λ. We also abbreviate

$$G^\theta(z) := G\left(Z^\theta, z\right), \quad G^{\theta,\lambda}_{(\mu i)}(z) := G\left(Z^{\theta,\lambda}_{(\mu i)}, z\right).$$

We shall prove the anisotropic local law (F.14) through interpolating matrices Z^θ between Z^0 and Z^1. We have seen that (F.14) holds for $G(Z^0, z)$ by Proposition F.7. Using (F.99) and fundamental calculus, we get the following basic interpolation formula: for any differentiable $F : \mathbb{R}^{n\times p} \to \mathbb{C}$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}F(Z^\theta) = \sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}\left[\mathbb{E}F\left(Z^{\theta,Z^1_{\mu i}}_{(\mu i)}\right) - \mathbb{E}F\left(Z^{\theta,Z^0_{\mu i}}_{(\mu i)}\right)\right], \tag{F.100}$$

provided all the expectations exist. We shall apply equation (F.100) to the function $F(Z) := F^s_{\mathbf{u}\mathbf{v}}(Z, z)$ for any fixed $s \in 2\mathbb{N}$, where

$$F_{\mathbf{u}\mathbf{v}}(Z, z) := \left|\mathbf{u}^\top \left(G(Z, z) - \mathfrak{G}(z)\right)\mathbf{v}\right|. \tag{F.101}$$

The main part of the proof is to show the following self-consistent estimate for the right-hand side of (F.100): for any fixed $s \in 2\mathbb{N}$, any constant $c > 0$ and all $\theta \in [0, 1]$,

$$\sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}\left[\mathbb{E}F^s_{\mathbf{u}\mathbf{v}}\left(Z^{\theta,Z^1_{\mu i}}_{(\mu i)}, z\right) - \mathbb{E}F^s_{\mathbf{u}\mathbf{v}}\left(Z^{\theta,Z^0_{\mu i}}_{(\mu i)}, z\right)\right] \leqslant (n^c q)^s + C\mathbb{E}F^s_{\mathbf{u}\mathbf{v}}(Z^\theta, z), \tag{F.102}$$

for a constant $C > 0$. Here and throughout the following proof, we abbreviate

$$q := n^{-1/2}Q.$$

If (F.102) holds, then combining equation (F.100) with Grönwall's inequality we obtain that for any fixed $s \in 2\mathbb{N}$ and constant $c > 0$,

$$\mathbb{E}\left|\mathbf{u}^\top \left(G(Z^1, z) - \Pi(z)\right)\mathbf{v}\right|^s \lesssim (n^c q)^s. \tag{F.103}$$

Finally applying Markov's inequality and noticing that c can be chosen arbitrarily small, we conclude (F.14).

In order to prove equation (F.102), we compare $Z^{\theta,Z^0_{\mu i}}_{(\mu i)}$ and $Z^{\theta,Z^1_{\mu i}}_{(\mu i)}$ via a common $Z^{\theta,0}_{(\mu i)}$, i.e. we will prove that for any constant $c > 0$,

$$\sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}\left[\mathbb{E}F^s_{\mathbf{u}\mathbf{v}}\left(Z^{\theta,Z^\alpha_{\mu i}}_{(\mu i)}, z\right) - \mathbb{E}F^s_{\mathbf{v}}\left(Z^{\theta,0}_{(\mu i)}, z\right)\right] \lesssim (n^c q)^s + \mathbb{E}F^s_{\mathbf{u}\mathbf{v}}\left(Z^\theta, z\right), \tag{F.104}$$

for all $\alpha \in \{0, 1\}$ and $\theta \in [0, 1]$. Underlying the proof of the estimate (F.104) is an expansion approach which we describe now. We define the $\mathcal{I} \times \mathcal{I}$ matrix $\Delta^\lambda_{(\mu i)}$ as

$$\Delta^\lambda_{(\mu i)} := \lambda\begin{pmatrix} 0 & \mathbf{u}^{(\mu)}_i \mathbf{e}^\top_\mu \\ \mathbf{e}_\mu (\mathbf{u}^{(\mu)}_i)^\top & 0 \end{pmatrix}, \tag{F.105}$$

where we denote $\mathbf{u}^{(\mu)}_i := \Lambda U \mathbf{e}_i$ if $\mu \in \mathcal{I}_1$, and $\mathbf{u}^{(\mu)}_i := V\mathbf{e}_i$ if $\mu \in \mathcal{I}_2$. Here $\mathbf{e}_i$ and $\mathbf{e}_\mu$ denote the standard basis vectors along the i-th and μ-th directions. Then by the definition of H in equation (F.4), we have for any $\lambda, \lambda' \in \mathbb{R}$ and $K \in \mathbb{N}$,

$$G^{\theta,\lambda'}_{(\mu i)} = G^{\theta,\lambda}_{(\mu i)} + n^{-\frac{k}{2}}\sum_{k=1}^{K} G^{\theta,\lambda}_{(\mu i)}\left(\Delta^{\lambda-\lambda'}_{(\mu i)} G^{\theta,\lambda}_{(\mu i)}\right)^k + n^{-\frac{K+1}{2}} G^{\theta,\lambda'}_{(\mu i)}\left(\Delta^{\lambda-\lambda'}_{(\mu i)} G^{\theta,\lambda}_{(\mu i)}\right)^{K+1}. \tag{F.106}$$

Using this expansion and the a priori bound (F.48), it is easy to prove the following estimate: if y is a random variable satisfying $|y| \prec Q$ (specifically the entries of all the interpolating matrices Z^θ satisfy this bound), then

$$G^{\theta,y}_{(\mu i)} = \mathrm{O}(1), \quad i \in \mathcal{I}_1,\ \mu \in \mathcal{I}_2 \cup \mathcal{I}_3, \tag{F.107}$$

with overwhelming probability.

In the following proof, for simplicity of notations, we denote

$$f_{(\mu i)}(\lambda) := F^s_{\mathbf{uv}}\left(Z^{\theta,\lambda}_{(\mu i)}\right) = \left|\mathbf{u}^\top\left(G\left(Z^{\theta,\lambda}_{(\mu i)}, z\right) - \mathfrak{G}(z)\right)\mathbf{v}\right|^s.$$

We use $f^{(r)}_{(\mu i)}$ to denote the r-th derivative of $f_{(\mu i)}$. By equation (F.107), it is easy to see that for any fixed $r \in \mathbb{N}$, $f^{(r)}_{(\mu i)}(y) = \mathrm{O}(1)$ with overwhelming probability for any random variable y satisfying $|y| \prec Q$. Then the Taylor expansion of $f_{(\mu i)}$ gives

$$f_{(\mu i)}(y) = \frac{1}{n^{r/2}} \sum_{r=0}^{s+4} \frac{y^r}{r!} f^{(r)}_{(\mu i)}(0) + \mathrm{O}_\prec\left(q^{s+4}\right). \tag{F.108}$$

Therefore we have for $\alpha \in \{0,1\}$,

$$\begin{aligned} &\mathbb{E}F^s_{\mathbf{uv}}\left(Z^{\theta,Z^\alpha_{\mu i}}_{(\mu i)}\right) - \mathbb{E}F^s_{\mathbf{uv}}\left(Z^{\theta,0}_{(\mu i)}\right) = \mathbb{E}\left[f_{(\mu i)}\left(Z^\alpha_{\mu i}\right) - f_{(\mu i)}(0)\right] \\ &= \mathbb{E}\, f_{(\mu i)}(0) + \frac{1}{2n}\mathbb{E}\, f^{(2)}_{(\mu i)}(0) + \sum_{r=3}^{s+4} \frac{n^{-r/2}}{r!}\mathbb{E}\, f^{(r)}_{(\mu i)}(0)\,\mathbb{E}\left(Z^\alpha_{\mu i}\right)^r + \mathrm{O}_\prec(q^{s+4}). \end{aligned} \tag{F.109}$$

Here to illustrate the idea in a more concise way, we assume the extra condition

$$\mathbb{E}(Z^1_{\mu i})^3 = 0, \quad 1 \leqslant \mu \leqslant n,\ \ 1 \leqslant i \leqslant p. \tag{F.110}$$

Hence the $r = 3$ term in the Taylor expansion (F.109) vanishes. However, this condition is not necessary as we will explain at the end of the proof.

Recall that the entries of $Z^1_{\mu i}$ have finite fourth moment as given by (2.3). Combining it with the bounded support condition, we have

$$\left|\mathbb{E}\left(Z^a_{\mu i}\right)^r\right| \prec Q^{r-4}, \quad r \geqslant 4. \tag{F.111}$$

Thus to show (F.104) under (F.110), we only need to prove that for $r = 4, \cdots, s+4$,

$$n^{-2}q^{r-4}\sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}\left|\mathbb{E}\, f^{(r)}_{(\mu i)}(0)\right| \lesssim (n^c q)^s + \mathbb{E}F^s_{\mathbf{u}\,\mathbf{v}}(Z^\theta, z). \tag{F.112}$$

In order to get a self-consistent estimate in terms of the matrix Z^θ on the right-hand side of (F.112), we want to replace $Z^{\theta,0}_{(\mu i)}$ in $f_{(\mu i)}(0) = F^s_{\mathbf{u}\,\mathbf{v}}(Z^{\theta,0}_{(\mu i)})$ with $Z^\theta \equiv Z^{\theta,Z^\theta_{\mu i}}_{(\mu i)}$.

LEMMA F.16. *Suppose that*

$$n^{-2}q^{r-4}\sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}\left|\mathbb{E}\, f^{(r)}_{(\mu i)}(Z^\theta_{\mu i})\right| \lesssim (n^c q)^s + \mathbb{E}F^s_{\mathbf{v}}(Z^\theta) \tag{F.113}$$

holds for $r = 4, \cdots, s+4$. Then (F.112) holds for $r = 4, \cdots, s+4$.

PROOF. The proof is the same as the one for [15, Lemma 7.16]. □

What remains now is to prove (F.113). For simplicity of notations, we shall abbreviate $Z^\theta \equiv Z$ in the following proof. For any $k \in \mathbb{N}$, we denote

$$A_{\mu i}(k) := \left(\frac{\partial}{\partial Z_{\mu i}}\right)^k \mathbf{u}^\top (G - \mathfrak{G})\, \mathbf{v}. \tag{F.114}$$

The derivative on the right-hand side can be calculated using the expansion equation (F.106). In particular, it is easy to verify that it satisfies the following bound

$$|A_{\mu i}(k)| \prec \begin{cases} (\mathcal{R}_i^{(\mu)})^2 + \mathcal{R}_\mu^2, & \text{if } k \geqslant 2 \\ \mathcal{R}_i^{(\mu)} \mathcal{R}_\mu, & \text{if } k = 1 \end{cases}, \tag{F.115}$$

where for $i \in \mathcal{I}_1$ and $\mu \in \mathcal{I}_1 \cup \mathcal{I}_2$, we denote

$$\mathcal{R}_i^{(\mu)} := |\mathbf{u}^\top G \mathbf{u}_i^{(\mu)}| + |\mathbf{v}^\top G \mathbf{u}_i^{(\mu)}|, \quad \mathcal{R}_\mu := |\mathbf{u}^\top G \mathbf{e}_\mu| + |\mathbf{v}^\top G \mathbf{e}_\mu|. \tag{F.116}$$

Then we can calculate the derivative

$$f_{(\mu i)}^{(r)}(Z_{\mu i}) = \left(\frac{\partial}{\partial Z_{\mu i}}\right)^r F_{\mathbf{u}\mathbf{v}}^s(Z) = \sum_{k_1 + \cdots + k_s = r} \prod_{t=1}^{s/2} \left(A_{\mu i}(k_t) \overline{A_{\mu i}(k_{t+s/2})} \right).$$

Then to prove (F.113), it suffices to show that

$$n^{-2} q^{r-4} \sum_{i \in \mathcal{I}_0} \sum_{\mu \in \mathcal{I}_1 \cup \mathcal{I}_2} \left| \mathbb{E} \prod_{t=1}^{s/2} A_{\mu i}(k_t) \overline{A_{\mu i}(k_{t+s/2})} \right| \lesssim (n^c q)^s + \mathbb{E} F_{\mathbf{u}\mathbf{v}}^s(Z, z), \tag{F.117}$$

for $4 \leqslant r \leqslant s + 4$ and $(k_1, \cdots, k_s) \in \mathbb{N}^s$ satisfying $k_1 + \cdots + k_s = r$. Treating zero k_t's separately (note $A_{\mu i}(0) = (G_{\mathbf{u}\mathbf{v}} - \Pi_{\mathbf{u}\mathbf{v}})$ by definition), we find that it suffices to prove

$$n^{-2} q^{r-4} \sum_{i \in \mathcal{I}_0} \sum_{\mu \in \mathcal{I}_1 \cup \mathcal{I}_2} \mathbb{E} \, |A_{\mu i}(0)|^{s-l} \prod_{t=1}^{l} |A_{\mu i}(k_t)| \lesssim (n^c q)^s + \mathbb{E} F_{\mathbf{u}\mathbf{v}}^s(Z, z) \tag{F.118}$$

for $4 \leqslant r \leqslant s + 4$ and $1 \leqslant l \leqslant s$. Here without loss of generality, we assume that $k_t = 0$ for $l + 1 \leqslant t \leqslant s$, $k_t \geqslant 1$ for $1 \leqslant t \leqslant l$, and $\sum_{t=1}^l k_t = r$.

For (F.118), notice that there are at least one non-zero k_t in all cases, while in the case $r \leqslant 2l - 2$, by pigeonhole principle, there exist at least two k_t's with $k_t = 1$. Therefore with (F.115), we have that

$$\prod_{t=1}^{l} |A_{\mu i}(k_t)| \prec \mathbf{1}(r \geqslant 2l - 1) \left[(\mathcal{R}_i^{(\mu)})^2 + \mathcal{R}_\mu^2 \right] + \mathbf{1}(r \leqslant 2l - 2) (\mathcal{R}_i^{(\mu)})^2 \mathcal{R}_\mu^2. \tag{F.119}$$

Using (F.48) and a similar argument as in (F.58), we can obtain that

$$\sum_{i \in \mathcal{I}_0} (\mathcal{R}_i^{(\mu)})^2 = \mathrm{O}(1), \quad \sum_{\mu \in \mathcal{I}_1 \cup \mathcal{I}_2} \mathcal{R}_\mu^2 = \mathrm{O}(1), \quad \text{w.o.p.} \tag{F.120}$$

Using (F.120) and $n^{-1/2} \leqslant n^{-1/2} Q = q$, we get that

$$\begin{aligned} & n^{-2} q^{r-4} \sum_{i \in \mathcal{I}_0} \sum_{\mu \in \mathcal{I}_1 \cup \mathcal{I}_2} |A_{\mu i}(0)|^{s-l} \prod_{t=1}^{l} |A_{\mu i}(k_t)| \\ & \prec q^{r-4} F_{\mathbf{u}\mathbf{v}}^{s-l}(Z) \left[\mathbf{1}(r \geqslant 2l - 1) n^{-1} + \mathbf{1}(r \leqslant 2l - 2) n^{-2} \right] \\ & \leqslant F_{\mathbf{u}\mathbf{v}}^{s-l}(Z) \left[\mathbf{1}(r \geqslant 2l - 1) q^{r-2} + \mathbf{1}(r \leqslant 2l - 2) q^r \right]. \end{aligned} \tag{F.121}$$

If $r \leqslant 2l-2$, then we have $q^r \leqslant q^l$ by the trivial inequality $r \geqslant l$. On the other hand, if $r \geqslant 4$ and $r \geqslant 2l-1$, then $r \geqslant l+2$ and we get $q^r \leqslant q^{l+2}$. Thus with (F.121), we conclude that

$$n^{-2}q^{r-4}\sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}\mathbb{E}|A_{\mu i}(0)|^{s-l}\prod_{t=1}^{l}|A_{\mu i}(k_t)|$$

$$\prec \mathbb{E}F_{\mathbf{uv}}^{s-l}(Z)q^l \leqslant [\mathbb{E}F_{\mathbf{uv}}^{s}(Z)]^{\frac{s-l}{s}}q^l \lesssim \mathbb{E}F_{\mathbf{uv}}^{s}(Z)+q^s,$$

where we use Hölder's inequality in the second step, and Young's inequality in the last step. This gives (F.118), which concludes the proof of (F.113), and hence of (F.104), and hence of (F.102), which concludes (F.103) and completes the proof of the anisotropic local law (F.14) under the condition (F.110).

Finally, if the condition equation (F.110) does not hold, then there is also an $r=3$ term in the Taylor expansion (F.109):

$$\frac{1}{6n^{3/2}}\,\mathbb{E}\,f^{(3)}_{(\mu i)}(0)\,\mathbb{E}\left(Z^{\alpha}_{i\mu}\right)^3.$$

But the sum over i and μ in equation (F.104) provides a factor n^2, which cannot be cancelled by the $n^{-3/2}$ factor in the above equation. In fact, $\mathbb{E}\,f^{(3)}_{(\mu i)}(0)$ will provide an extra $n^{-1/2}$ factor to compensate the remaining $n^{1/2}$ factor. This follows from an improved self-consistent comparison argument for sample covariance matrices in [15, Section 8]. The argument for our setting is almost the same except for some notational differences, so we omit the details. This concludes the proof of (F.14) without the condition (F.110).

F.5. *Averaged local law.* Finally, in this subsection, we prove the averaged local law (F.12) in the setting of Theorem F.3. The proof of (F.13) is almost the same.

The proof of (F.12) for $G(Z,z)$ is similar to that for equation (F.14) in previous subsection, and we only explain the differences. In analogy to (F.101), we define

$$\widetilde{F}(Z,z) := \left|\frac{1}{p}\sum_{i\in\mathcal{I}_0}\left[G_{ii}(Z,z)-\mathfrak{G}_{ii}(z)\right]\right|.$$

In the notation of Definition F.15, we have proved that $\widetilde{F}(Z^0,z)\prec(np)^{-1/2}$ in Lemma F.8. To illustrate the idea, we assume the condition (F.110) holds in the following argument. Following the argument in Section F.4, analogous to (F.113), we only need to prove that for $q=n^{-1/2}Q$ and any small constant $c>0$,

$$n^{-2}q^{r-4}\sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}\left|\mathbb{E}\left(\frac{\partial}{\partial Z_{\mu i}}\right)^r\widetilde{F}^s(Z)\right|\lesssim\left(p^{-1/2+c}q\right)^s+\mathbb{E}\widetilde{F}^s(Z), \tag{F.122}$$

for all $r=4,\ldots,s+4$. Similar to (F.114), we denote

$$A_{j,\mu i}(k) := \left(\frac{\partial}{\partial Z_{\mu i}}\right)^k(G_{jj}-\mathfrak{G}_{jj}).$$

Analogous to (F.117), it suffices to prove that

$$n^{-2}q^{r-4}\sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_2}\left|\mathbb{E}\prod_{t=1}^{s/2}\left(\frac{1}{p}\sum_{j\in\mathcal{I}_0}A_{j,\mu i}(k_t)\right)\left(\frac{1}{p}\sum_{j\in\mathcal{I}_0}\overline{A_{j,\mu i}(k_{t+s/2})}\right)\right|$$

$$\lesssim\left(p^{-1/2+c}q\right)^s+\mathbb{E}\widetilde{F}^s(Z),$$

for all $4 \leqslant r \leqslant s+4$ and $(k_1, \cdots, k_s) \in \mathbb{N}^s$ satisfying $k_1 + \cdots + k_s = r$. Without loss of generality, we assume that $k_t = 0$ for $l+1 \leqslant t \leqslant s$, $k_t \geqslant 1$ for $1 \leqslant t \leqslant l$, and $\sum_{t=1}^{l} k_t = r$. Then it suffices to prove

$$n^{-2}q^{r-4} \sum_{i \in \mathcal{I}_0} \sum_{\mu \in \mathcal{I}_1 \cup \mathcal{I}_2} \mathbb{E}\, \widetilde{F}^{s-l}(Z) \prod_{t=1}^{l} \left| \frac{1}{p} \sum_{j \in \mathcal{I}_0} A_{j,\mu i}(k_t) \right| \lesssim \left(p^{-1/2+c} q \right)^s + \mathbb{E} \widetilde{F}^s(Z) \tag{F.123}$$

for $4 \leqslant r \leqslant s+4$ and $1 \leqslant l \leqslant s$.

First, using (F.48) and a similar argument as in (F.58), we can obtain that for $1 \leqslant t \leqslant l$,

$$\left| \sum_{j \in \mathcal{I}_0} A_{j,\mu i}(k_t) \right| \lesssim 1 \quad \text{w.o.p} \tag{F.124}$$

On the other hand, similar to (F.116) we define

$$\mathcal{R}_{j,i}^{(\mu)} := |\mathbf{u}^\top G \mathbf{u}_i^{(\mu)}| + |\mathbf{v}^\top G \mathbf{u}_i^{(\mu)}|, \quad \mathcal{R}_{j,\mu} := |\mathbf{u}^\top G \mathbf{e}_\mu| + |\mathbf{v}^\top G \mathbf{e}_\mu|.$$

As in (F.120), we have that

$$\sum_{i \in \mathcal{I}_0} (\mathcal{R}_{j,i}^{(\mu)})^2 = \mathrm{O}(1), \quad \sum_{\mu \in \mathcal{I}_1 \cup \mathcal{I}_2} \mathcal{R}_{j,\mu}^2 = \mathrm{O}(1), \quad \text{w.o.p.} \tag{F.125}$$

Now with (F.124) and a similar bound as for (F.119), we obtain that

$$\begin{aligned} \prod_{t=1}^{l} \left| \frac{1}{p} \sum_{j \in \mathcal{I}_0} A_{j,\mu i}(k_t) \right| &\prec \mathbf{1}(r \geqslant 2l-1) p^{-(l-1)} \frac{1}{p} \sum_{j \in \mathcal{I}_0} \left[(\mathcal{R}_{j,i}^{(\mu)})^2 + \mathcal{R}_{j,\mu}^2 \right] \\ &\quad + \mathbf{1}(r \leqslant 2l-2) p^{-(l-2)} \mathcal{R}_{j_1,i}^{(\mu)} \mathcal{R}_{j_2,i}^{(\mu)} \mathcal{R}_{j_1,\mu} \mathcal{R}_{j_2,\mu} \end{aligned}$$

Summing this equation over $i\mathcal{I}_0$ and $\mu \in \mathcal{I}_1 \cup \mathcal{I}_2$ and using (F.125), we get that

$$\begin{aligned} &n^{-2}q^{r-4} \sum_{i \in \mathcal{I}_0} \sum_{\mu \in \mathcal{I}_1 \cup \mathcal{I}_2} \mathbb{E}\, \widetilde{F}^{s-l}(Z) \prod_{t=1}^{l} \left| \frac{1}{p} \sum_{j \in \mathcal{I}_0} A_{j,\mu i}(k_t) \right| \\ &\prec q^{r-4}\, \mathbb{E}\, \widetilde{F}^{s-l}(Z) \left[\mathbf{1}(r \geqslant 2l-1) p^{-(l-1)} n^{-1} + \mathbf{1}(r \leqslant 2l-2) n^{-2} p^{-(l-2)} \right] \end{aligned}$$

We consider the following cases.

- If $r \geqslant 2l-1$ and $l \geqslant 2$, then we have $r \geqslant l+2$ and $l - 1 \geqslant l/2$, which gives
$$p^{-(l-1)} q^{r-4} n^{-1} \leqslant p^{-l/2} q^{r-2} \leqslant (p^{-1/2} q)^l.$$
- If $r \geqslant 2l-1$ and $l = 1$, then we have
$$p^{-(l-1)} q^{r-4} n^{-1} \leqslant n^{-1} \leqslant (p^{-1/2} q)^l.$$
- If $r \leqslant 2l-2$ and $l \geqslant 4$, then we have $r \geqslant l$ and $l - 2 \geqslant l/2$, which gives
$$p^{-(l-2)} q^{r-4} n^{-2} \leqslant p^{-l/2} q^r \leqslant (p^{-1/2} q)^l.$$
- If $r \leqslant 2l-2$ and $l < 4$, then we must have $l = 3$ (because $r \geqslant 4$), which gives
$$p^{-(l-2)} q^{r-4} n^{-2} \leqslant p^{-1} n^{-2} \leqslant (p^{-1/2} q)^l.$$

Combining the above cases, we see that

$$n^{-2}q^{r-4}\sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}\mathbb{E}\,\widetilde{F}^{s-l}(Z)\prod_{t=1}^{l}\left|\frac{1}{p}\sum_{j\in\mathcal{I}_0}A_{j,\mu i}(k_t)\right| \prec \mathbb{E}\,\widetilde{F}^{p-l}(X)\left(p^{-1/2}q\right)^{l}.$$

Applying Holder's inequality and Young's inequality, we then obtain (F.122), which completes the proof of the averaged local law (F.12) under condition (F.110).

Finally, even if the condition (F.110) does not hold, using the self-consistent comparison argument in [15, Section 9], we can still prove (F.12) for $G(Z,z)$. Again the arguments are almost the same as the ones in [15, Section 9], hence we omit the details.

REFERENCES

[1] ALT, J., ERDŐS, L. and KRÜGER, T. (2017). Local law for random Gram matrices. *Electron. J. Probab.* **22** 41 pp.

[2] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral analysis of large dimensional random matrices*, 2nd ed. *Springer Series in Statistics*. Springer, New York.

[3] BAI, Z. D. and SILVERSTEIN, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.* **26** 316–345.

[4] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2017). Local Law of Addition of Random Matrices on Optimal Scale. *Communications in Mathematical Physics* **349** 947–990.

[5] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2017). Convergence rate for spectral distribution of addition of random matrices. *Advances in Mathematics* **319** 251 - 291.

[6] BELINSCHI, S. T. and BERCOVICI, H. (2007). A new approach to subordination results in free probability. *Journal d'Analyse Mathématique* **101** 357–365.

[7] BLOEMENDAL, A., ERDŐS, L., KNOWLES, A., YAU, H. T. and YIN, J. (2014). Isotropic Local Laws for Sample Covariance and Generalized Wigner Matrices. *Electron. J. Probab.* **19** 1-53.

[8] CHISTYAKOV, G. P. and GÖTZE, F. (2011). The arithmetic of distributions in free probability theory. *Central European Journal of Mathematics* **9** 997–1050.

[9] DING, X. and YANG, F. (2018). A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *Ann. Appl. Probab.* **28** 1679–1738.

[10] ERDŐS, L., KNOWLES, A. and YAU, H. T. (2013). Averaging Fluctuations in Resolvents of Random Band Matrices. *Ann. Henri Poincaré* **14** 1837-1926.

[11] ERDŐS, L., KNOWLES, A., YAU, H. T. and YIN, J. (2013). Spectral statistics of Erdős-Rényi graphs I: Local semicircle law. *Ann. Probab.* **41** 2279-2375.

[12] ERDŐS, L., KNOWLES, A., YAU, H. T. and YIN, J. (2013). Delocalization and Diffusion Profile for Random Band Matrices. *Commun. Math. Phys.* **323** 367-416.

[13] ERDŐS, L., KNOWLES, A., YAU, H. T. and YIN, J. (2013). The local semicircle law for a general class of random matrices. *Electron. J. Probab.* **18** 1-58.

[14] ERDOS, L. and YAU, H.-T. (2017). A dynamical approach to random matrix theory. *Courant Lecture Notes in Mathematics* **28**.

[15] KNOWLES, A. and YIN, J. (2016). Anisotropic local laws for random matrices. *Probability Theory and Related Fields* 1–96.

[16] MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik* **1** 457.

[17] NICA, A. and SPEICHER, R. (2006). *Lectures on the combinatorics of free probability* **13**. Cambridge University Press.

[18] PILLAI, N. S. and YIN, J. (2014). Universality of covariance matrices. *Ann. Appl. Probab.* **24** 935–1001.

[19] TAO, T. (2012). *Topics in Random Matrix Theory* **132**. American Mathematical Soc.

[20] XI, H., YANG, F. and YIN, J. (2017). Local circular law for the product of a deterministic matrix with a random matrix. *Electron. J. Probab.* **22** 77 pp.

[21] YANG, F. (2019). Edge universality of separable covariance matrices. *Electron. J. Probab.* **24** 57 pp.