Generalization Effects of Transferring in High-Dimensions using Multi-Task Learning

Anonymous Author(s)
Affiliation
Address
`email`

1 Introduction

Multi-task learning is an inductive learning mechanism to improve generalization performance using related task data. Many state-of-the-art results in computer vision and natural language processing are obtained using multi-task learning. On the other hand, unless the structures across task data are well-understood, applying multi-task learning on several different datasets often result in suboptimal models (or negative transfer in more technical terms).

[Todo: Motivate high-dimensional data.] The technical challenge is to deal with high-dimensional data, in particular when the size of the training set is only a small constant times the feature dimension. Prior theory is unable to explain a phase transition that occurs when comparing multi-task learning to single-task learning. In particular in Figure 1, we observe a shift from positive transfer to negative transfer as a parameter of task relatedness. The theory we develop will provide a precise explanation to such a phenomenon (and more).

To gain insight into the working of multi-task learning, we consider a simplified setting for learning multiple high-dimensional linear regression tasks. A typical process to do multi-task learning involves two steps: (i) Jointly learn a shared representation for all the tasks; (ii) Fine-tune the learnt model on a specific target task. We focus on a hard parameter sharing model proposed in [**?**] and identify conditions on when multi-task and transfer learning works, and when it doesn't. The high-dimensional linear regression setting where the target task data size is limited captures the intuition that the target task only contains limited labeled data.

Concretely, our input consists of k tasks (X_1, Y_1), (X_2, Y_2), $\dots$, (X_k, Y_k). We shall assume that each task data follows a linear model, i.e. $y_i = X_i\beta_i + \varepsilon_i$, $1 \leqslant i \leqslant k$. Here $\beta_i \in \mathbb{R}^p$ is the model parameter for the i-th task. Each row of $X_i \in \mathbb{R}^{n_i \times p}$ is assumed to be drawn i.i.d. from a fixed distribution with covariance matrix Σ_i. We use a shared body $B \in \mathbb{R}^{p\times r}$ for all tasks and a separate prediction head $\{W_i \in \mathbb{R}^r\}_{i=1}^k$ for each task. This corresponds to minimizing the following optimization objective.

$$f(B; W_1, \dots, W_k) = \sum_{i=1}^{k} \|X_i B W_i - Y_i\|^2. \tag{1.1}$$

Note that we consider the natural parameterization without reweighting the tasks above. The shared body B plays an important role because it allows information transfer between different task data. This is known as the hard parameter sharing architecture in the literature, where we control the capacity r of B, e.g. [**?**, **?**].

We focus on comparing the test performance on a target task using estimators from doing multi-task training and transfer learning. We compare the test performance of these estimators to the single-task baseline. The details are described in Algorithm 1.

By using random matrix theory, we can explain several phenomena that are not explained by the techniques of [**?**]. [Todo: list those here] We refine the task covariance part of [**?**] into three

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

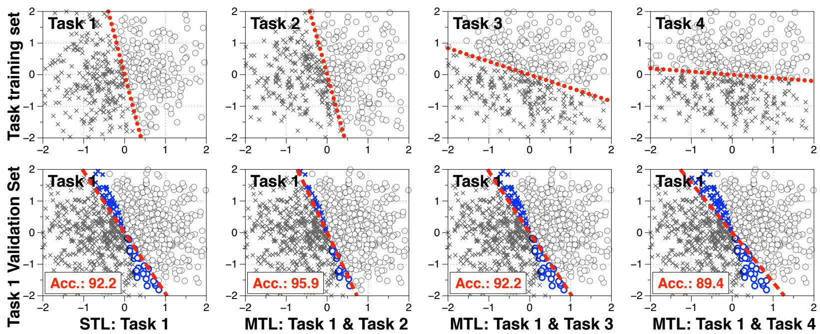

Figure 1: Phase transition as a parameter of model distance.

Algorithm 1 Multi-task learning using a hard-parameter sharing architecture

---

Input: Two regression tasks (X_1, Y_2), (X_2, Y_2).
Parameter: Shared body B, task-specific prediction heads W_1, W_2.

Training the shared body B.
Optimizing the task heads on the validation set
- Jointly optimizing both tasks: $\hat{\beta}_t^{\text{MTL}}$
- Optimizing the target task: $\hat{\beta}_t^{\text{TL-FT}}$
- Single-task training baseline: $\hat{\beta}_t^{\text{STL}}$

Problem statement: how can we compare the test error of the three estimators on the target task?

---

factors: model distance, covariate shift matrix and data ratio [**?**, **?**]. These are achieved through tight generalization bounds established in the high-dimensional regression setting. Our main results are described as follows.

Variance reduction from information transfer. We show that the benefit from doing multi-task or transfer learning stems from reducing the variance of the estimator for the target task through newly added source task data. We derive this result for the setting of two tasks with general inputs and k tasks with the same covariates for any $k \geqslant 2$. The latter setting is prevalent in applications of multi-task learning to image classification, where there are multiple prediction labels/tasks for every image [**?**]. On the other hand, the difference between task models causes a negative effect that we call the *model shift bias*. We show bounds on the trade-off between the amount of variance reduced and the amount of model shift bias incurred, which become tighter and tighter as the number of source task data points increases.

Insight 1: $\hat{\beta}_t^{\text{MTL}}$ vs $\hat{\beta}_t^{\text{STL}}$. With multi-task training, since the training objective balances the losses from both the source and target tasks, the trained model can have worse performance for the target task. In particular, if model shift is too large, we get negative transfer from multi-task training.

Insight 2: $\hat{\beta}_t^{\text{TL-FT}}$ vs $\hat{\beta}_t^{\text{MTL}}$. Finally, we show that the transfer learning estimator always improves over mutli-task training. The amount of improvement becomes more significant as the model distance becomes larger.

Covariate shift and data ratio. For the case of two tasks with general inputs, we further study the factors that determine the transfer rate. We identify two factors, the covariate shift matrix and the ratio between number of task data.

Insight 3: $\hat{\beta}_t^{\text{TL-FT}}$ vs $\hat{\beta}_t^{\text{STL}}$. Our result has implications on the following question. Is it better for two tasks to have the same covariance matrix or complementary covariance matrices. For our setting, we show that when the data ratio is large, having the same covariance matrix provably yields the lowest test performance on the target task. On the other hand, when data ratio is small, we find that there are cases when having complementary covariance matrices is better. The result provides insight into why the covariance alignment algorithm can help improve performance in [**?**].

Experimental results.

2 Preliminaries

We assume that for every row x of X_i, we have $\mathbb{E}\left[xx^\top\right] = \Sigma_i$. We also write $x = \Sigma_i^{1/2} z_i$, where z_i is a random vector with mean 0 and variance 1. We will designate the k-th task as the target. Our goal is to come up with an estimator $\hat{\beta}$ to provide accurate predictions for the target task, provided with the other auxiliary task data. Concretely, we focus on the test error for the target task:

$$\begin{aligned} te_k(\hat{\beta}) &:= \mathop{\mathbb{E}}_{x\sim\Sigma_k}\left[\mathop{\mathbb{E}}_{\varepsilon_i,\forall 1\leqslant i\leqslant k}\left[(x^\top\hat{\beta} - x^\top\beta_t)^2\right]\right] \\ &= \mathop{\mathbb{E}}_{\varepsilon_i,\forall 1\leqslant i\leqslant k}\left[(\hat{\beta}-\beta_t)^\top\Sigma_k(\hat{\beta}-\beta_t)\right]. \end{aligned}$$

[Todo: show that $te_k(\hat{\beta}_t^{\text{TL-FT}})$ is less than both $te_k(\hat{\beta}_t^{\text{MTL}})$ and $te_k(\hat{\beta}_t^{\text{STL}})$.]

Hypothesis on Heterogeneous Task Data Our hypothesis is that the heterogeneity among the multiple tasks can be categorized into two classes, *covariate shift* and *model shift*. We consider two natural questions within each category.

- **Model shift.** In general the single-task models can also be different across different tasks. We shall argue that in addition to the bias and variance terms of generalization error, model shift introduces a third term which is the bias caused by model shift.

  Under model shift, when do we get positive vs. negative transfer? How does the type of transfer depend on the number of data points, the distance of the task models etc?
- **Covariate shift.** The covariance matrices Σ_i may be be different across tasks, i.e. having different spectrum or singular vectors. This is also known as covariate shift in the literature. How does covariate shift affect the rate of information transfer? For example, is it better to have the same covariance matrix or not?

The High-Dimensional Setting. We would like to get insight on how covariate and model shifts affect the rate of transfer. We will consider the high-dimensional setting where for the target task, its number of data points is a small constant times p. This setting captures a wide range of applications of multi-task learning where we would like to use auxiliary task data to help train tasks with limited labeled data. Furthermore, this setting is particularly suited to our study since there is need for adding more data to help learn the target task.

For the case of two tasks, we can get precise rates using random matrix theory. For the sake of clarity, we call task 1 the source task and task 2 the target task, i.e. $\beta_1 = \beta_s$ and $\beta_2 = \beta_t$. We introduce the following notations for the high-dimensional setting

$$c_{n_1} := \frac{n_1}{p} \to c_1, \quad c_{n_2} := \frac{n_2}{p} \to c_2, \quad \text{as } n_1, n_2 \to \infty,$$

for some constants $c_1, c_2 \in (1, \infty)$. A crucial quantity is what we call the *covariate shift* matrix $M = \Sigma_1^{1/2}\Sigma_2^{-1/2}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ denote the singular values of M.

3 Illustrative Examples and Their Insights on Transfer

We illustrate our main results (to be presented in Section 4) by considering a few special cases, namely special settings of the task models $\{\beta_i\}_{i=1}^k$, covariance matrices $\{\Sigma_i\}_{i=1}^k$, and number of data points $\{n_i\}_{i=1}^k$. We show that our results explain several phenomenon that cannot be explained before. [Todo: list those here]

3.1 Model Distance

We compare the test error of $\hat{\beta}_t^{\text{MTL}}$ to that of $\hat{\beta}_t^{\text{STL}}$. For a simple example, we show that whether $\hat{\beta}_t^{\text{MTL}}$ performs better than $\hat{\beta}_t^{\text{STL}}$ is determined by the distance of the task models. We derive a sharp threshold when positive transfer transitions to negative transfer, as a ratio between the model distance and the noise level.

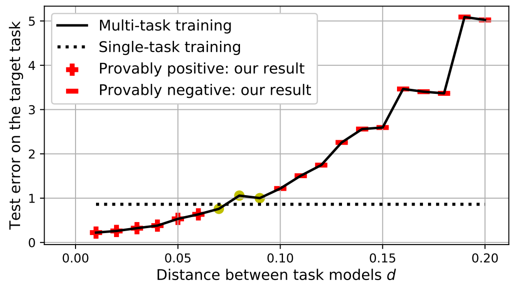

Figure 2: Positive vs negative transfer as a parameter of the task model distances.

Example. Consider a setting where $\Sigma_1 = \Sigma_2 = \mathrm{Id}$, in other words there is no covariate shift between the two tasks. For the task models, suppose that β_t has i.i.d. entries with mean zero and variance κ^2 and $\beta_s - \beta_t$ also has i.i.d. entries with mean 0 and variance d^2. We have $n = c \cdot p$ data points from each task.

We illustrate the example in a synthetic setting. We demonstrate our result with a simulation. ([Todo: uses the tighter bound Proposition A.3?]) We consider a setting where $p = 200$, $n_1 = 90p$, $n_2 = 30p$. [Todo: Fill in other params.] We fix the target task and vary the source task, by varying the task model distance parameter d. We show that Theorem 4.1 predicts whether we can get positive or negative transfer. Figure 2 shows the result.

Specifically, the transition threshold is derived in the following proposition.

Proposition 3.1. *In the example described above, whether $te(\hat{\beta}_t^{MTL})$ is lower than $te(\hat{\beta}_t^{STL})$ is determined by the ratio between the model distance and the noise level:*

- *If $p \cdot d^2/\sigma^2 < \frac{2(2c-1)^2}{(c-1)(2c+1)} - o(1)$, then whp we have that $te(\hat{\beta}_t^{MTL}) < te(\hat{\beta}_t^{STL})$.*
- *If $p \cdot d^2/\sigma^2 \geqslant \frac{2(2c-1)^2}{(c-1)(2c+1)} + o(1)$, then whp we have that $te(\hat{\beta}_t^{MTL}) \geqslant te(\hat{\beta}_t^{STL})$.*

We obtain the following insight from Proposition 3.1. First, adding the source task has the effect of reducing the variance of the estimator, independent of the model shift. On the other hand, model shift introduces an additional bias term, which scales with d^2, the distance of the two task models. Hence, the type of transfer is determined by the trade-off between the bias caused by model shift and the reduction of variance.

3.2 Data Ratio

3.3 Covariate Shift

Note that for the case of k tasks with the same covariates, since there is no covariate shift and the data ratio is always equal to one, the main factor is model distance.

A precise bound when there is no model shift. As Proposition **??** shows, if β_s and β_t are equal, then adding the source task dataset always helps learn the target task. The goal of this section is to understand how covariate shift affects the rate of transfer. [Todo: add conceptual msg]

A simple observation here is that when $\beta_s = \beta_t$, the optimal $\hat{w}$ for minimizing equation (**??**) is equal to 1. Based on this observation, we can get a more precise result than Theorem 4.1 on the improvement of adding the source task data that only depends on the covariance matrices Σ_1, Σ_2 and the number of data points n_1, n_2.

Proposition 3.2 (Transfer rate without model shift)**.** *Suppose $\beta_s = \beta_t$ and $\|\beta_t\|_2^2 \sim p\sigma^2$ (i.e. the l^2-norm of the vector β_t is of the same order as that of the error vector). Assume that the condition numbers of Σ_1, Σ_2 and $M := \Sigma_1^{1/2}\Sigma_2^{-1/2}$ are all bounded by a constant $C > 0$. Then we have that the optimal ratio for W_1/W_2 in equation* (**??**) *satisfies*

$$1 \leqslant \hat{w} \leqslant 1 + \mathrm{O}(p^{-1}).$$

Moreover, we have

$$te(\hat{\beta}_t^{\textit{TL-FT}}) = \sigma^2 \cdot \operatorname{Tr}\left[\left((n_1+n_2)a_1 M^\top M + (n_1+n_2)a_2 \operatorname{Id}\right)^{-1}\right] \cdot \left(1 + \mathrm{O}\left(p^{-1}\right)\right), \tag{3.1}$$

where a_1, a_2 *are the solutions to equations* (4.2).

Proof. We abbreviate $val(w_2\hat{B}(w)) := val(w)$. Note that $val(w) \leqslant val(-w)$ for $w \geqslant 0$. Hence we have $\hat{w} \geqslant 0$. Moreover, we notice that $val(w) < val(1)$ for all $0 \leqslant w < 1$. Thus we have $\hat{w} \geqslant 1$. It suffices to consider the case with $w > 1$. Under the assumption on β_s and β_t, we can write

$$\begin{aligned} val(w) = &\left(1 - \frac{1}{w}\right)^2 \left\|(M^\top Z_1^\top Z_1 M + w^{-2} Z_2^\top Z_2)^{-1} M^\top Z_1^\top Z_1 \Sigma_1^{1/2} \beta_t\right\|^2 \\ &+ \frac{\sigma^2}{w^2} \cdot \operatorname{Tr}\left[(M^\top Z_1^\top Z_1 M + w^{-2} Z_2^\top Z_2)^{-1}\right]. \end{aligned}$$

Since

$$\begin{aligned} &\left\|(M^\top Z_1^\top Z_1 M + w^{-2} Z_2^\top Z_2)^{-1} M^\top Z_1^\top Z_1 \Sigma_1^{1/2} \beta_t\right\|^2 \\ &= \operatorname{Tr}\left[(M^\top Z_1^\top Z_1 M + w^{-2} Z_2^\top Z_2)^{-2} M^\top Z_1^\top Z_1 \Sigma_1^{1/2} \beta_t \beta_t^\top \Sigma_1^{1/2} Z_1^\top Z_1 M\right] \end{aligned}$$

is increasing with respect to w, then the derivative of $val(w)$ can be bounded from below as

$$\begin{aligned} val'(w) \geqslant\ & 2\frac{w-1}{w^3} \left\|(M^\top Z_1^\top Z_1 M + w^{-2} Z_2^\top Z_2)^{-1} M^\top Z_1^\top Z_1 \Sigma_1^{1/2} \beta_t\right\|^2 \\ & - 2\frac{\sigma^2}{w^3} \cdot \operatorname{Tr}\left[(M^\top Z_1^\top Z_1 M + w^{-2} Z_2^\top Z_2)^{-1} M^\top Z_1^\top Z_1 M (M^\top Z_1^\top Z_1 M + w^{-2} Z_2^\top Z_2)^{-1}\right] \\ \geqslant\ & 2\frac{w-1}{w^3} \left\|(M^\top Z_1^\top Z_1 M + Z_2^\top Z_2)^{-1} M^\top Z_1^\top Z_1 \Sigma_1^{1/2} \beta_t\right\|^2 - 2\frac{\sigma^2}{w^3} \cdot \operatorname{Tr}\left[(M^\top Z_1^\top Z_1 M)^{-1}\right]. \end{aligned}$$

Hence $val'(w) \geqslant 0$ if $w > 1 + \varepsilon_0$, where

$$\varepsilon_0 := \frac{\sigma^2 \operatorname{Tr}\left[(M^\top Z_1^\top Z_1 M)^{-1}\right]}{\left\|(M^\top Z_1^\top Z_1 M + Z_2^\top Z_2)^{-1} M^\top Z_1^\top Z_1 \Sigma_1^{1/2} \beta_t\right\|^2}.$$

In other words, $val(w)$ is strictly increasing function on $[1+\varepsilon_0, \infty]$. Thus we get that $\hat{w}$ satisfies

$$1 \leqslant w \leqslant 1 + \varepsilon_0. \tag{3.2}$$

Using (B.3), we get that

$$\varepsilon_0 = \mathrm{O}(\sigma^2 / \|\beta_t\|_2^2) = \mathrm{O}(p^{-1}).$$

Finally, plugging (3.2) into the expression $te(\hat{\beta}_t^{\text{TL-FT}})$, we obtain (3.1). $\square$

As a remark, we see that Proposition **??** follows from Theorem 3.2. The amount of reduction on test error for the target task is given as

$$te(\hat{\beta}_t) - te(\hat{\beta}_{s,t}) = \sigma^2 \cdot \left(\frac{p}{n_2 - p} - \operatorname{Tr}\left[\left((n_1+n_2)a_1 \Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2} + (n_1+n_2)a_2 \operatorname{Id}\right)^{-1}\right]\right).$$

Because

$$\begin{aligned} te(\hat{\beta}_{s,t}) \leqslant te(\hat{\beta}_t) &\Leftarrow\ (n_2 - p)\Sigma_2 \preceq (n_1+n_2)a_1 \Sigma_1 + (n_1+n_2)a_2 \Sigma_2 \\ &\Leftrightarrow\ \mathbf{0} \preceq (n_1+n_2)a_1 \Sigma_1 + (n_1 - (n_1+n_2)\cdot a_1)\Sigma_2, \end{aligned}$$

which is true since $a_1 \leqslant n_1/(n_1+n_2)$ by equation (4.2). The proof for $te(\hat{\beta}_{s,t}) \leqslant te(\hat{\beta}_t)$ follows by multiplying $\Sigma_2^{-1/2}$ on both sides of the inequalities above.

Now we apply Theorem 3.2 to show how covariate shift affects the rate of transfer.

Example 3.3 (**When** $\Sigma_1 = \Sigma_2/\lambda$). In this case, equations (4.2) become

$$a_1 + a_2 = 1 - p/(n_1+n_2), a_1 + \frac{p}{n_1+n_2} \cdot \frac{a_1}{a_1 + \lambda^2 a_2} = \frac{n_1}{n_1+n_2}.$$

By solving these, we can get the test errors (the estimation error behaves similarly). Figure 3 shows how they grow as we increase the number of source task data points. Here $n_2 = 4p$ and n_1 ranges from p to $20p$. We can see that the smaller λ is, the lower the test errors will be.

Example 3.4 (**When** Σ_1 **and** Σ_2 **are complementary**). We now consider another case when Σ_1 and Σ_2 have complementary eigenspaces. Suppose Σ_1 and Σ_2 have the eigendecomposition

$$\Sigma_1^{1/2} = 1 + U\Lambda U^\top, \quad \Sigma_2^{1/2} = 1 + V\Lambda V,$$

where

$$\Lambda = \mathrm{Diag}(\widetilde{\lambda}_1, \cdots, \widetilde{\lambda}_{p/2}), \quad U = (u_1, \cdots, u_{p/2}), \quad V = (v_1, \cdots, v_{p/2}).$$

If $V = U_\perp$, i.e. the vectors $v_1, \cdots, v_{p/2}$ are perpendicular to the vectors $u_1, \cdots, u_{p/2}$, then

$$M = \Sigma_1^{1/2}\Sigma_2^{-1/2} = (1+\Lambda)UU^\top + (1+\Lambda)^{-1}VV^\top.$$

As a concrete example we consider the case where $\widetilde{\lambda}_1 = \cdots = \widetilde{\lambda}_{p/2}$ and we denote $\lambda := 1 + \widetilde{\lambda}_1$. Thus for M, the first $p/2$ singular values are equal to λ and the rest ones are equal to λ^{-1}. In this case, equations in (4.2) become

$$a_1 + a_2 = 1 - \frac{p}{n_1+n_2}, \;\; a_1 + \frac{p}{2(n_1+n_2)} \cdot \left(\frac{a_1}{a_1+\lambda^2 a_2} + \frac{a_1}{a_1 + \frac{a_2}{\lambda^2}}\right) = \frac{n_1}{n_1+n_2}. \tag{3.3}$$

It's not hard to verify that there is only one valid solution (a_1, a_2) to (3.3). After solving these, we get the test error for the target task as follows.

$$te(\lambda) = \frac{p}{2(n_1+n_2)} \cdot \left(\frac{1}{\frac{a_1}{\lambda^2} + a_2} + \frac{1}{a_1\lambda^2 + a_2}\right). \tag{3.4}$$

In Figure 4, we plot the test error of the target task for $n_2 = 4p$ and n_1 ranging from p to $20p$. First we notice that the curves in Figure 4 all cross at the point $n_1 = n_2$. In fact, if $n_1 = n_2$, then it is easy to observe that $a_1 = a_2 = (1-\gamma)/2$ is the solution to equation (3.3), where we denote $\gamma = p/(n_1+n_2)$. Then for any λ, the test error in (3.4) takes the value

$$te(\lambda) = \frac{\gamma}{2}\frac{1}{(1-\gamma)/2} = \frac{p}{n_1+n_2-p}.$$

Second, we observe the following two phases as we increase n_1/p.

- When $n_1 \leqslant n_2$, having complementary covariance matrices leads to lower test error compared to the case when $\Sigma_1 = \Sigma_2$.
- When $n_1 > n_2$, having complementary covariance matrices leads to higher test error compared to the case when $\Sigma_1 = \Sigma_2$.

This phenomenon can be also explained using our theory. With (3.3), we can write

$$te(\lambda) = \frac{\gamma}{2} \cdot \left(\frac{1}{\frac{a_1}{\lambda^2} + (1-\gamma-a_1)} + \frac{1}{a_1\lambda^2 + (1-\gamma-a_1)}\right).$$

We can compute that

$$\begin{aligned} te(\lambda) - te(1) &= \frac{\gamma}{2(1-\gamma)}(\lambda^2-1)a_1 \cdot \left(\frac{1}{-a_1(\lambda^2-1)+(1-\gamma)\lambda^2} - \frac{1}{a_1(\lambda^2-1)+(1-\gamma)}\right) \\ &= \frac{\gamma}{2(1-\gamma)}(\lambda^2-1)^2 a_1 \cdot \frac{2a_1 - (1-\gamma)}{[-a_1(\lambda^2-1)+(1-\gamma)\lambda^2][a_1(\lambda^2-1)+(1-\gamma)]}. \end{aligned}$$

If $n_1 > n_2$, we have $a_1 > (1-\gamma)/2$ (because $a_1 > a_2$ as observed from the equation (3.3)), and hence $te(\lambda) > te(1)$. Otherwise if $n_1 < n_2$, we have $a_1 < (1-\gamma)/2$, and hence $te(\lambda) < te(1)$.

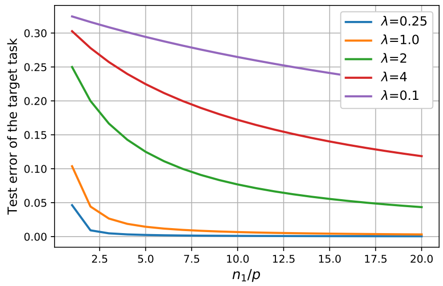

Figure 3: When $\Sigma_1 = \Sigma_2/\lambda$.

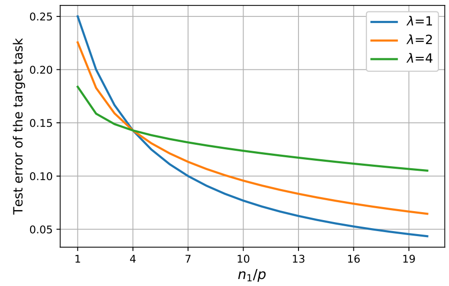

Figure 4: When Σ_1 and Σ_2 are complementary. The number of target task data points is $n_2 = 4p$.

We can extend the above observation for Example 3.4 to more general cases with arbitrary covariate shift. To compare different choices of M, we need to fix a scaling for them, because otherwise aM always achieves a smaller error than M for $a > 1$. For this purpose, we introduce the following condition

$$\det(M^\top M) = a^p \Leftrightarrow \prod_{i=1}^p \lambda_i = a^p \tag{3.5}$$

for some constant $a > 0$, and compare different choices of M under this constraint. The next proposition shows that, roughly speaking, as long as there are sufficiently many source task datas, then $M = a\,\mathrm{Id}$ always gives (approximately) the smallest test error.

Proposition 3.5. *We have that the test error satisfies*

$$te(M)\left(1 - \frac{n_2}{n_1 - p}\frac{1}{\lambda_p^2 + \frac{n_2}{n_1 - p}}\right) \leqslant \frac{\sigma^2}{n_1 + n_2}\operatorname{Tr}\left(\frac{1}{a_1 M^\top M + a_2}\right) \leqslant te(M), \tag{3.6}$$

where λ_p is the smallest singular value of p and

$$te(M) := \frac{\sigma^2}{a_1(n_1 + n_2)}\operatorname{Tr}\left(\frac{1}{M^\top M}\right).$$

Moreover, for all M satisfying (3.5)*, the minimum of $te(M)$ is attained when $M = a\,\mathrm{Id}$.*

Proof. From equation (4.2), we get

$$a_1 \geqslant \frac{n_1 - p}{n_1 + n_2}, \quad a_2 \leqslant \frac{n_2}{n_1 + n_2}.$$

With these two bounds, we can easily conclude (3.6).

For the second statement, we find that we need to minimize

$$\operatorname{Tr}\left(\frac{1}{M^\top M}\right) = \sum_{i=1}^p \frac{1}{\lambda_i}, \quad \text{under } \prod_{i=1}^p \lambda_i = a^p.$$

Then using AM-GM inequality, we conclude that the sum $\sum_{i=1}^p \lambda_i^{-1}$ is smallest when $\lambda_1 = \cdots = \lambda_p = a$. □

From the above proposition, we see that when $n_1 \gg n_2$, we have that the test error is approximately given by $te(M)$ (as long as we impose a proper lower bound on λ_p). Moreover, $te(M)$ is minimized when Σ_1 and Σ_2 are proportional to each other, i.e. there is no covariate shift between the source task data and target task data. This provides a theoretical evidence that in general covariate shift is unfavored in transfer learning if we have enough source task data, although Example 3.4 show that this may not be true when the number of source task data is small.

Extending the intuition to the general case. When there is model shift, i.e. $\beta_s = \beta_t$, we can still use Theorem 4.1 (and Proposition A.3) to get the result.

- **The effect of covariate shift:**
- **The effect of data ratio:**

4 Main Results: A Model Shift Bias versus Variance Trade-off

In this part, we consider the case of two tasks to show establish the intuition that adding more data helps in multi-task learning by reducing the variance of the estimator. We achieve this through tight generalization bounds obtained from random matrix theory. For the case of two tasks, we identify three factors that determine the type of transfer between tasks: model distance, covariate shift matrix, and data ratio.

Recall that the test error of $\hat{\beta}_t^{\text{MTL}}$ consists of two parts

$$te(\hat{\beta}_t^{\text{MTL}}) = \hat{w}^2 \left\| \Sigma_2^{1/2} (\hat{w}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 (\beta_s - \hat{w}\beta_t) \right\|^2 + \sigma^2 \cdot \operatorname{Tr}\left[(\hat{w}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_2 \right]. \tag{4.1}$$

It is not hard to show that the variance of $\hat{\beta}_t^{\text{MTL}}$ is reduced compared to $\hat{\beta}_t^{\text{STL}}$ (following the argument of Proposition **??**), i.e.

$$\sigma^2 \cdot \operatorname{Tr}\left[(\hat{w}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_2 \right] \leqslant \sigma^2 \cdot \operatorname{Tr}\left[(X_2^\top X_2)^{-1} \Sigma_2 \right].$$

Because of model shift however, i.e. $\beta_s \neq \beta_t$. We can no longer guarantee that $te(\hat{\beta}_t^{\text{MTL}}) \leqslant te(\hat{\beta}_t^{\text{STL}})$. The main result of this part show deterministic conditions under which we get positive or negative transfer. And the conditions depend only on the covariate shift matrix M, the difference of the task models, and the number of per-task data points. In order to characterize $te(\hat{\beta}_t^{\text{MTL}})$ and $te(\hat{\beta}_t^{\text{STL}})$, the technical crux of our approach relies on deriving the limit of the trace of matrix inverse in the high-dimensional setting. To illustrate the idea, we observe that by using Lemma A.1, we have that

$$te(\hat{\beta}_t^{\text{STL}}) = \frac{\sigma^2}{n_2 - p} \operatorname{Tr}\left[\Sigma_2^{-1} \right].$$

We shall also derive the limit of $te(\hat{\beta}_t^{\text{MTL}})$. [Todo: write a brief technical overview]

Let $M = \hat{w}\Sigma_1^{1/2}\Sigma_2^{-1/2}$ denote the weighted covariate shift matrix. Denote by $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$ the singular values of $M^\top M$. Let (a_1, a_2) be the solutions to the following system of equations

$$a_1 + a_2 = 1 - \frac{p}{n_1 + n_2}, \quad a_1 + \sum_{i=1}^p \frac{a_1}{(n_1 + n_2)(a_1 + a_2/\lambda_i^2)} = \frac{n_1}{n_1 + n_2}. \tag{4.2}$$

After obtaining (a_1, a_2), we can solve the following linear equations to get (a_3, a_4):

$$\left(\frac{n_2}{a_2^2} - \sum_{i=1}^p \frac{1}{(a_2 + \lambda_i^2 a_1)^2} \right) a_3 - \left(\sum_{i=1}^p \frac{\lambda_i^2}{(a_2 + \lambda_i^2 a_1)^2} \right) a_4 = \sum_{i=1}^p \frac{1}{(a_2 + \lambda_i^2 a_1)^2}, \tag{4.3}$$

$$\left(\frac{n_1}{a_1^2} - \sum_{i=1}^p \frac{\lambda_i^4}{(a_2 + \lambda_i^2 a_1)^2} \right) a_4 - \left(\sum_{i=1}^p \frac{\lambda_i^2}{(a_2 + \lambda_i^2 a_1)^2} \right) a_3 = \sum_{i=1}^p \frac{\lambda_i^2}{(a_2 + \lambda_i^2 a_1)^2}. \tag{4.4}$$

Then we introduce the following matrix

$$Z = \frac{n_1^2}{(n_1 + n_2)^2} \cdot M \frac{(1 + a_3)\operatorname{Id} + a_4 M^\top M}{(a_2 + a_1 M^\top M)^2} M^\top,$$

which can be regarded as the asymptotic limit of $\hat{w}\Sigma_2^{1/2}(\hat{w}^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_1^{1/2}$. Finally we introduce

$$\delta := \left[\frac{n_1 \lambda_1}{(\sqrt{n_1} - \sqrt{p})^2 \lambda_p + (\sqrt{n_2} - \sqrt{p})^2} \right]^2 \cdot \|\Sigma_1^{1/2}(\beta_s - \hat{w}\beta_t)\|^2.$$

may be able to get a better bound, but the statement will be long

We now state our main result for two tasks with both covariate and model shift in the following theorem.

Theorem 4.1. *Let n_1, n_2 be the number of data points for the source, target task, respectively. Let $\hat{w}$ denote the optimal solution for the ratio w_1/w_2 in equation* (A.2). *The information transfer is solely determined by two deterministic quantities Δ_β and Δ_{var}, which show the change of model shift bias and variace, respectively. With high probability we have*

$$te(\hat{\beta}_t^{MTL}) \leqslant te(\hat{\beta}_t^{STL}) \text{ when: } \Delta_{var} - \Delta_\beta \geqslant \left(\left(1 + \sqrt{\frac{p}{n_1}}\right)^4 - 1\right)\delta \tag{4.5}$$

$$te(\hat{\beta}_t^{MTL}) \geqslant te(\hat{\beta}_t^{STL}) \text{ when: } \Delta_{var} - \Delta_\beta \leqslant -2\left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1}\right)\delta, \tag{4.6}$$

where

$$\Delta_{var} := \sigma^2 \left(\frac{p}{n_2 - p} - \frac{1}{n_1 + n_2} \operatorname{Tr}\left[(a_1 M^\top M + a_2 \operatorname{Id})^{-1}\right]\right)$$
$$\Delta_\beta := (\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} Z \Sigma_1^{1/2} (\beta_s - \hat{w}\beta_t).$$

Theorem 4.1 shows upper and lower bounds that guarantee positive transfer, which is determined by the change of variance Δ_{var} and a certain model shift bias parameter Δ_β determined by the covariate shift matrix and the model shift. The bounds get tighter and tighter as n_1/p increases.

4.1 Extension to Many Tasks of the Same Covariates

In this section we consider the setting with k many that have the same covariates. Since every task has the same number of data points as well as the same covariance, the only differences between different tasks are their models $\{\beta_i\}_{i=1}^k$. For this setting, we derive solutions for the multi-task training and the transfer learning setting that match our insights qualitatively from Section **??**.

5 Related Work

Adding a regularization over B, e.g. [**?**, **?**]. Moreover, [**?**] observed that controlling the capacity can outperform the implicit capacity control of adding regularization over B.

A Supplementary Materials for Section 3

From [?], we know that we need to explicitly restrict the capacity r of B so that there is transfer between the two tasks. for the rest of the section, we shall consider the case when $r = 1$ we are considering the case of two tasks. Here, equation (1.1) simplifies to the following

$$f(B; w_1, w_2) = \|X_1 B w_1 - Y_1\|^2 + \|X_2 B w_2 - Y_2\|^2, \tag{A.1}$$

where $B \in \mathbb{R}^p$ and w_1, w_2 are both real numbers. To solve the above, suppose that w_1, w_2 are fixed, by local optimality, we solve B as

$$\begin{aligned}
&\hat{B}(w_1, w_2)\\
=&(w_1^2 X_1^\top X_1 + w_2^2 X_2^\top X_2)^{-1}(w_1 X_1^\top Y_1 + w_2 X_2^\top Y_2)\\
=&\frac{1}{w_2}((\frac{w_1}{w_2})^2 X_1^\top X_1 + X_2^\top X_2)^{-1}(\frac{w_1}{w_2} X_1^\top Y_1 + X_2^\top Y_2)\\
=&\frac{1}{w_2}\left(\beta_t + ((\frac{w_1}{w_2})^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\left(X_1^\top X_1(\frac{w_1}{w_2}\beta_s - w^2\beta_t) + (\frac{w_1}{w_2} X_1^\top \varepsilon_1 + X_2^\top \varepsilon_2)\right)\right).
\end{aligned}$$

As a remark, when $w_1 = w_2 = 1$, we recover the linear regression estimator. The advantage of using $f(B; w_1, w_2)$ is that if θ_1 is a scaling of θ_2, then this case can be solved optimally using equation (A.1) [?].

Defining the multi-task learning estimator. Using a validation set that is sub-sampled from the original training dataset, we get a validation loss as follows

$$\begin{aligned}
val(\hat{B}; w_1, w_2) =& n_1 \cdot \left\|\Sigma_1^{1/2}(w^2 X_1^\top X_1 + X_2^\top X_2)^{-1} X_2^\top X_2(\beta_s - w\beta_t)\right\|^2\\
&+ n_1\sigma^2 \cdot \mathrm{Tr}\left[(w^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_1\right]\\
&+ n_2 \cdot w^2 \left\|\Sigma_2^{1/2}(w^2 X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1(\beta_s - w\beta_t)\right\|^2\\
&+ n_2 \cdot \sigma^2 \cdot w^2\, \mathrm{Tr}\left[(w^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_2\right]
\end{aligned} \tag{A.2}$$

Let $\hat{w}_1, \hat{w}_2$ be the global minimizer of $val(\hat{B}; w_1, w_2)$. We will define the multi-task learning estimator for the target task as

$$\hat{\beta}_t^{\mathrm{MTL}} = \hat{w}_2 \hat{B}(\hat{w}_1, \hat{w}_2).$$

The intuition for deriving $\hat{\beta}_t^{\mathrm{MTL}}$ is akin to performing multi-task training in practice. Let $\hat{v} = \hat{w}_1/\hat{w}_2$ for the simplicity of notation. The test loss of using $\hat{\beta}_t^{\mathrm{MTL}}$ for the target task is

$$\begin{aligned}
te(\hat{\beta}_t^{\mathrm{MTL}}) =& \hat{v}^2 \left\|\Sigma_2^{1/2}(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1(\beta_s - \hat{v}\beta_t)\right\|^2\\
&+ \sigma^2 \cdot \mathrm{Tr}\left[(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_2\right].
\end{aligned} \tag{A.3}$$

Our goal is to study under model and covariate shifts, whether multi-task learning helps learn the target task better than single-task learning. The baseline where we solve the target task with its own data is

$$te(\hat{\beta}_t^{\mathrm{STL}}) = \sigma^2 \cdot \mathrm{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right], \text{ where } \hat{\beta}_t^{\mathrm{STL}} = (X_2^\top X_2)^{-1} X_2^\top Y_2.$$

We first state several helper lemmas that can be viewed as corollaries of Theorem 4.1.

Lemma A.1. *[[Todo: ref?]] Let $X \in \mathbb{R}^{n\times p}$ that contains i.i.d. row vectors with mean 0 and covariance Σ. In the setting when $n = cp$ we have that as p goes to infinity,*

$$\mathrm{Tr}\left[(X^\top X)^{-1}\Sigma\right] = \frac{1}{c-1}.$$

Lemma A.2. *[Todo: restate the setting] In the setting of Theorem 4.1, we have with high probability $1 - \mathrm{o}(1)$,*

$$\mathrm{Tr}((X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_2) = \frac{1}{n_1 + n_2} \cdot \mathrm{Tr}\left[(a_1 M^\top M + a_2)^{-1}\right] + \mathrm{O}\left(n^{-1/2+\varepsilon}\right), \tag{A.4}$$

for any constant $\varepsilon > 0$.

We will give the proof of Lemma A.2 in Section C.

A.1 Proof of Proposition 3.1

Under the setting of Proposition 3.1, we can get a bound tighter than Theorem 4.1 as follows.

Proposition A.3. *In the setting of Theorem 4.1, assume that* $\Sigma_1 = \mathrm{Id}$, β_t *is i.i.d. with mean* 0 *and variance* κ^2 *and* $\beta_s - \beta_t$ *is i.i.d. with mean* 0 *and variance* d^2. *We set* $\Delta_\beta = \left((1-\hat{w})^2\kappa^2 + d^2)\right) \mathrm{Tr}\,[Z]$ *and we have*

$$te(\hat{\beta}_t^{MTL}) \leqslant te(\hat{\beta}_t^{STL}) \textit{ when: } \Delta_{var} \geqslant \left(1 + \sqrt{\frac{p}{n_1}}\right)^4 \Delta_\beta,$$
$$te(\hat{\beta}_t^{MTL}) \geqslant te(\hat{\beta}_t^{STL}) \textit{ when: } \Delta_{var} \leqslant \left(1 - \sqrt{\frac{p}{n_1}}\right)^4 \Delta_\beta.$$

We will give the proof of this proposition after proving Theorem 4.1 in Section B.1 .

Using Lemma A.1 and A.2, we can track the change of variance from $\hat{\beta}_t^{\mathrm{MTL}}$ to $\hat{\beta}_t^{\mathrm{STL}}$ as follows The proof will consist of two main steps.

- First, we show that $\hat{v}$ is close to 1.
- Second, we plug $\hat{v}$ back into $te(\hat{\beta}_t^{\mathrm{MTL}})$ to show the result.

Hence the task models have distance $d^2 \cdot p$ in expectation. We first consider $\Sigma_2 = \mathrm{Id}$. In this case, we can simplify Δ_β as follows

$$\Delta_\beta := d^2 \cdot \sum_{i=1}^{p} \frac{(1+a_3)\lambda_i^2 + a_4\lambda_i^4}{(a_1\lambda_i^2 + a_2)^2}. \tag{A.5}$$

Now we solve the equations (4.2), (4.3), (4.4) to get

$$a_1 = \frac{c_1(c_1+c_2-1)}{(c_1+c_2)^2}, a_2 = \frac{c_2(c_1+c_2-1)}{(c_1+c_2)^2}, a_3 = \frac{c_2}{(c_1+c_2)(c_1+c_2-1)}, a_4 = \frac{c_1}{(c_1+c_2)(c_1+c_2-1)}. \tag{A.6}$$

Then we obtain

$$\Delta_\beta = p \cdot d^2 \cdot \frac{c_1^2(c_1+c_2)}{(c_1+c_2-1)^3}, \Delta_{\text{var}} = \sigma^2 \cdot \frac{c_1}{(c_2-1)(c_1+c_2-1)}. \tag{A.7}$$

We denote

$$\begin{aligned} val(w) = {} & n_1\left[d^2 + (w-1)^2\kappa^2\right] \cdot \mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-2}(X_2^\top X_2)^2\right] \\ & + n_2\left[d^2 + (w-1)^2\kappa^2\right] \cdot \mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-2}w^2(X_1^\top X_1)^2\right] \\ & + (n_1 + n_2w^2)\sigma^2 \cdot \mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-1}\right]. \end{aligned}$$

Under the setting of Proposition A.3, using concentration of random vector with i.i.d. entries, we have that

$$val(\hat{B}; w_1, w_2) = val(w)\left(1 + \mathrm{o}(1)\right) \quad \text{with probability } 1 - \mathrm{o}(1).$$

Thus it suffices to study the behavior of $val(w)$. For the minimizer $\hat{w}$ of $val(w)$, we have a similar result as in Proposition 3.2.

Lemma A.4. *Suppose the assumptions of Proposition A.3 hold. Assume that* $\kappa^2 \sim pd^2 \sim \sigma^2$ *are of the same order. Then we have that the optimal ratio for* $val(w)$ *satisfies*

$$|\hat{w} - 1| = \mathrm{O}(p^{-1}).$$

Proof. The proof is also similar to the one for Proposition 3.2. First it is easy to observe that $val(w) \leqslant val(-w)$ for $w \geqslant 0$. Hence it suffices to consider the $w \geqslant 0$ case.

We first consider the case $w \geqslant 1$. We write

$$\begin{aligned} val(w) = n_1 &\left[\frac{d^2}{w^4} + \frac{(w-1)^2}{w^4}\kappa^2\right] \cdot \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}(X_2^\top X_2)^2\right] \\ &+ n_2 \left[\frac{d^2}{w^2} + \frac{(w-1)^2}{w^2}\kappa^2\right] \cdot \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ &+ \sigma^2 n_1 \cdot \mathrm{Tr}\left[(w^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\right] + \sigma^2 n_2 \cdot \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-1}\right]. \end{aligned}$$

Taking derivative of $val(w)$ with respect to w, we obtain that

$$\begin{aligned} val'(w) \geqslant n_1 &\left[\frac{2(w-1)(2-w)}{w^5}\kappa^2 - \frac{4d^2}{w^5}\right] \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}(X_2^\top X_2)^2\right] \\ &+ n_2 \left[\frac{2\,(w-1)}{w^3}\kappa^2 - \frac{2d^2}{w^3}\right] \cdot \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ &- 2n_1\frac{\sigma^2}{w^3} \cdot \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}X_1^\top X_1\right] = n_1 \,\mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}\mathcal{A}\right], \end{aligned}$$

where the matrix $\mathcal{A}$ is

$$\mathcal{A} := \left[\frac{2(w-1)(2-w)}{w^5}\kappa^2 - \frac{4d^2}{w^5}\right](X_2^\top X_2)^2 + \frac{n_2}{n_1}\left[\frac{2\,(w-1)}{w^3}\kappa^2 - \frac{2d^2}{w^3}\right](X_1^\top X_1)^2. - 2\frac{\sigma^2}{w^3}X_1^\top X_1$$

Using the estimate (B.3), we get that $\mathcal{A}$ is lower bounded as

$$\mathcal{A} \succeq -\frac{4d^2}{w^5}(\sqrt{n_2}+\sqrt{p})^4 + \frac{n_2}{n_1}\left[\frac{2\,(w-1)}{w^3}\kappa^2 - \frac{2d^2}{w^3}\right](\sqrt{n_1}-\sqrt{p})^4 - 2\frac{\sigma^2}{w^3}(\sqrt{n_1}+\sqrt{p})^2 \succ 0,$$

as long as

$$w > w_1 := 1 + \frac{d^2}{\kappa^2} + \frac{\sigma^2}{\kappa^2}\frac{n_1(\sqrt{n_1}+\sqrt{p})^2}{n_2(\sqrt{n_1}-\sqrt{p})^4} + \frac{2d^2}{\kappa^2}\frac{n_1(\sqrt{n_2}+\sqrt{p})^4}{n_2(\sqrt{n_1}-\sqrt{p})^4}.$$

Hence $val'(w) > 0$ on $(1+w_1, \infty)$, i.e. $val(w)$ is strictly increasing for $w > 1 + w_1$. Hence we must have $\hat{w} \leqslant 1 + w_1$. Note that under our assumptions, we have $w_1 = 1 + \mathrm{O}(p^{-1})$.

Then we consider the case $w \leqslant 1$. Taking derivative of $val(w)$ with respect to w, we obtain that

$$\begin{aligned} val'(w) \leqslant n_1 &\left[2\,(w-1)\,\kappa^2\right] \cdot \mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-2}(X_2^\top X_2)^2\right] \\ &+ n_2\left[2d^2 w + 2w\,(w-1)\,(2w-1)\kappa^2\right] \cdot \mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ &+ 2n_2 w\sigma^2 \cdot \mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-2}X_2^\top X_2\right] = n_1\,\mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-1}\mathcal{B}\right], \end{aligned}$$

where the matrix $\mathcal{B}$ is

$$\mathcal{B} = 2\,(w-1)\,\kappa^2(X_2^\top X_2)^2 + \frac{n_2}{n_1}\left[2d^2 w + 2w\,(w-1)\,(2w-1)\kappa^2\right](X_1^\top X_1)^2 + 2\frac{n_2}{n_1}w\sigma^2 X_2^\top X_2.$$

Using the estimate (B.3), we get that $\mathcal{A}$ is lower bounded as

$$\mathcal{A} \succeq -\frac{4d^2}{w^5}(\sqrt{n_2}+\sqrt{p})^4 + \frac{n_2}{n_1}\left[\frac{2\,(w-1)}{w^3}\kappa^2 - \frac{2d^2}{w^3}\right](\sqrt{n_1}-\sqrt{p})^4 - 2\frac{\sigma^2}{w^3}(\sqrt{n_1}+\sqrt{p})^2 \succ 0,$$

as long as

$$w > w_1 := 1 + \frac{d^2}{\kappa^2} + \frac{\sigma^2}{\kappa^2}\frac{n_1(\sqrt{n_1}+\sqrt{p})^2}{n_2(\sqrt{n_1}-\sqrt{p})^4} + \frac{2d^2}{\kappa^2}\frac{n_1(\sqrt{n_2}+\sqrt{p})^4}{n_2(\sqrt{n_1}-\sqrt{p})^4}.$$

Hence $val'(w) > 0$ on $(1+w_1, \infty)$, i.e. $val(w)$ is strictly increasing for $w > 1 + w_1$. Hence we must have $\hat{w} \leqslant 1 + w_1$. Note that under our assumptions, we have $w_1 = 1 + \mathrm{O}(p^{-1})$. $\square$

B Supplementary Materials for Section 4

B.1 Proof of Theorem 4.1

[Todo: A proof outline; including the following key lemma.] To prove Theorem 3.2, we study the spectrum of the random matrix model:

$$Q = \Sigma_1^{1/2} Z_1^\top Z_1 \Sigma_1^{1/2} + \Sigma_2^{1/2} Z_2^\top Z_2 \Sigma_2^{1/2},$$

where $\Sigma_{1,2}$ are $p \times p$ deterministic covariance matrices, and $X_1 = (x_{ij})_{1 \leqslant i \leqslant n_1, 1 \leqslant j \leqslant p}$ and $X_2 = (x_{ij})_{n_1+1 \leqslant i \leqslant n_1+n_2, 1 \leqslant j \leqslant p}$ are $n_1 \times p$ and $n_2 \times p$ random matrices, respectively, where the entries x_{ij}, $1 \leqslant i \leqslant n_1 + n_2 \equiv n$, $1 \leqslant j \leqslant p$, are real independent random variables satisfying

$$\mathbb{E} z_{ij} = 0, \qquad \mathbb{E}|z_{ij}|^2 = 1. \tag{B.1}$$

The proof of Theorem 4.1 involves two parts.

Part I: Bounding the bias from model shift. We relate the first term in equation (4.1) to Δ_β.

Proposition B.1. *In the setting of Theorem 4.1, denote by $K = (\hat{w}^2 X_1^\top X_1 + X_2^\top X_1)^{-1}$, and*

$$\begin{aligned}
\delta_1 &= \hat{w}^2 \left\| \Sigma_2^{1/2} K X_1^\top X_1 (\beta_s - \hat{w}\beta_t) \right\|^2, \\
\delta_2 &= n_1^2 \cdot \hat{w}^2 \left\| \Sigma_2^{1/2} K \Sigma_1 (\beta_s - \hat{w}\beta_t) \right\|, \\
\delta_3 &= n_1^2 \cdot \hat{w}^2 \left\| \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1^{1/2} \right\| \cdot \left\| \Sigma_1^{1/2} (\beta_s - \hat{w}\beta_t) \right\|^2.
\end{aligned}$$

We have that

$$-2n_1^2 \left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1} \right) \delta_3 \leqslant \delta_1 - \delta_2 \leqslant n_1^2 \left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1} \right) \left(2 + 2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1} \right) \delta_3.$$

For the special case when $\Sigma_1 = \mathrm{Id}$ and $\beta_s - \beta_t$ is i.i.d. with mean 0 and variance d^2, we further have

$$\left(1 - \sqrt{\frac{p}{n_1}} \right)^4 \Delta_\beta \leqslant \left\| \Sigma_2^{1/2} (X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 (\beta_s - \beta_t) \right\|^2.$$

Proof. The proof follows by applying equation (B.3). Recall that $X_1^\top X_1 = \Sigma_1^{1/2} Z_1^\top Z_1 \Sigma_1^{1/2}$. Denote by $\mathcal{E} = Z_1^\top Z_1 - n_1 \,\mathrm{Id}$. Let We have

$$\delta_1 = \delta_2 + 2\hat{w}^2 n_1 (\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t) + \hat{w}^2 \left\| \Sigma_2^{1/2} K \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} (\beta_s - \hat{w}\beta_t) \right\|^2 \tag{B.2}$$

Here we use the following on the second term in equation (B.2)

$$\begin{aligned}
&\left| (\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t) \right| \\
=& \left| \mathrm{Tr} \left[\mathcal{E} \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t)(\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} \right] \right| \\
\leqslant& \|\mathcal{E}\| \cdot \left\| \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t)(\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} \right\|_\star \\
\leqslant& n_1 \left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1} \right) \cdot \left\| \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t)(\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} \right\|_\star \qquad \text{(by equation (B.3))} \\
\leqslant& n_1 \left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1} \right) \left\| \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1^{1/2} \right\| \cdot \left\| \Sigma_1^{1/2} (\beta_s - \hat{w}\beta_t) \right\|^2 \\
& \text{(since the matrix inside is rank 1)}
\end{aligned}$$

The third term in equation (B.2) can be bounded with

$$\left\| \Sigma_2^{1/2} K \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} (\beta_s - \hat{w}\beta_t) \right\|^2 \leqslant n_1^2 \left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1} \right)^2 \left\| \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1^{1/2} \right\| \cdot \left\| \Sigma_1^{1/2} (\beta_s - \hat{w}\beta_t) \right\|^2.$$

Combined together we have shown the right direction for $\delta_1 - \delta_2$. For the left direction, we simply note that the third term in equation (B.2) is positive. And the second term is bigger than $-2n_1^2 (2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1})\alpha$ using equation (B.3). □

Part II: The limit of $\left\|\Sigma_2^{1/2}(\hat{w}^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_1(\beta_s - \hat{w}\beta_t)\right\|^2$ using random matrix theory. We consider the same setting as in previous subsection:

$$X_1^\top X_1 := \Sigma_1^{1/2} Z_1^T Z_1 \Sigma_1^{1/2}, \quad X_2^\top X_2 = \Sigma_2^{1/2} Z_2^T Z_2 \Sigma_2^{1/2},$$

where z_{ij}, $1 \leqslant i \leqslant n_1 + n_2 \equiv n$, $1 \leqslant j \leqslant p$, are real independent random variables satisfying (B.1). For now, we assume that the random variables z_{ij} are i.i.d. Gaussian, but we know that universality holds for generally distributed entries. Assume that p/n_1 is a small number such that $Z_1^T Z_1$ is roughly an isometry, that is, under (B.1), If we assume the variances of the entries of Z_1 are 1, then we have

$$-n_1\left(2\sqrt{\frac{p}{n_1}} - \frac{p}{n_1}\right) \leqslant Z_1^T Z_1 - n_1 \operatorname{Id} \leqslant n_1\left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1}\right). \tag{B.3}$$

Lemma B.2. *In the setting of Theorem 4.1, we have with high probability* $1 - \mathrm{o}(1)$,

$$\begin{aligned}
&\widehat{w}^2(n_1+n_2)^2\left\|\Sigma_2^{1/2}(\hat{w}X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_1(\beta_s - w\beta_t)\right\|^2 \\
&= (\beta_s - w\beta_t)^\top \Sigma_1^{1/2} M \frac{(1+a_3)\operatorname{Id} + a_4 M^\top M}{(a_2 + a_1 M^\top M)^2} M^\top \Sigma_1^{1/2}(\beta_s - \hat{w}\beta_t) + \mathrm{O}(n^{-1/2+\varepsilon}),
\end{aligned} \tag{B.4}$$

for any constant $\varepsilon > 0$.

We will give the proof of this lemma in Section C.

Proof of Proposition A.3. the proof for tighter bound .... $\square$

add some arguments with ε-net.

B.2 Proof of Theorem ???

Extension to Many Tasks of the Same Covariates In this section we consider the setting with k many that have the same covariates. Since every task has the same number of data points as well as the same covariance, the only differences between different tasks are their models $\{\beta_i\}_{i=1}^k$. For this setting, we derive solutions for the multi-task training and the transfer learning setting that match our insights qualitatively from Section **??**.

Concretely we will consider the following problem.

$$f(B; W_1, \dots, W_k) = \sum_{i=1}^k \|XBW_i - Y_i\|^2. \tag{B.5}$$

By fixing $W_1, W_2, \dots, W_k$, we can derive a closed form solution for B as

$$\begin{aligned}
\hat{B}(W_1, \dots, W_k) &= (X^\top X)^{-1} X^\top \left(\sum_{i=1}^k Y_i W_i^\top\right)(ZZ^\top)^{-1} \\
&= \sum_{i=1}^k \left(\beta_i W_i^\top\right)(ZZ^\top)^{-1} + (X^\top X)^{-1} X^\top \left(\sum_{i=1}^k \varepsilon_i W_i^\top\right)(ZZ^\top)^{-1}
\end{aligned}$$

where we denote $Z \in \mathbb{R}^{r \times k}$ as the k vectors $W_1, W_2, \dots, W_k$ stacked together. Similar to Section 2, we consider minimizing the validation loss over $W_1, W_2, \dots, W_k$ provided with $\hat{B}$.

Jointly minimizing over all tasks. Denote by $\varepsilon(W) = \sum_{i=1}^k \varepsilon_i W_i^\top$. We shall decompose the validation loss $val(\hat{B}; W_1, \dots, W_k)$ into two parts. The first part is the model shift bias, which is equal to

$$\sum_{j=1}^k \left(\left\|\Sigma^{1/2}\left(\sum_{i=1}^k (\beta_i W_i^\top)(ZZ^\top)^{-1} W_j - \beta_j\right)\right\|^2\right)$$

The second part is the variance, which is equal to

$$\sum_{j=1}^{k} \mathop{\mathbb{E}}_{\varepsilon_i, \forall i} \left[\left(\left(\sum_{i=1}^{k} \varepsilon_i W_i^\top \right) (ZZ^\top)^{-1} W_j \right)^2 \right] \cdot \operatorname{Tr}\left[\Sigma (X^\top X)^{-1}\right]$$
$$= \sigma^2 \cdot \operatorname{Tr}\left[\Sigma (X^\top X)^{-1}\right].$$

Therefore we shall focus on the minimizer for the model shift bias since the variance part does not depend the weights. Let us denote $Q \in \mathbb{R}^{k \times k}$ where the (i,j)-th entry is equal to $W_i^\top (ZZ^\top)^{-1} W_j$, for any $1 \leqslant i, j \leqslant k$. Let $B^\star = [\beta_1, \beta_2, \dots, \beta_k] \in \mathbb{R}^{p \times k}$ denote the true model parameters. We can now write the validation loss succinctly as follows.

$$val(\hat{B}; W_1, \dots, W_k) = \sum_{j=1}^{k} \left\| \Sigma^{1/2} \left(B^\star Q - \beta_j \right) \right\|^2 + \sigma^2 \cdot \operatorname{Tr}\left[\Sigma (X^\top X)^{-1}\right]$$

From the above we can solve for Q optimally as [Todo: this].

Minimizing over the target task alone. If we only minimize over the the validation loss for the target task, we shall get the following.

$$val_j(\hat{B} W_j) = \left\| \Sigma^{1/2} \left(\sum_{i=1}^{k} W_i^\top (ZZ^\top)^{-1} W_j \beta_i - \beta_j \right) \right\|^2 + \sigma^2 \cdot W_j^\top (ZZ^\top)^{-1} W_j \cdot \operatorname{Tr}\left[\Sigma (X^\top X)^{-1}\right].$$

From the above we can obtain three conceptual insights that are consistent with Section **??** and 3.

- The de-noising effect of multi-task learning.
- Multi-task training vs single-task training can be either positive or negative.
- Transfer learning is better than the other two. And the improvement over multi-task training increases as the model distances become larger.

C Proof of Lemma A.2 and Lemma B.2

We consider two $p \times p$ random sample covariance matrices $\mathcal{Q}_1 := \Sigma_1^{1/2} Z_1^\top Z_1 \Sigma_1^{1/2}$ and $\mathcal{Q}_2 := \Sigma_2^{1/2} Z_2^\top Z_2 \Sigma_2^{1/2}$, where Σ_1 and Σ_2 are $p \times p$ deterministic non-negative definite (real) symmetric matrices. We assume that $Z_1 = (z_{ij}^{(1)})$ and $Z_2 = (z_{ij}^{(2)})$ are $n_1 \times p$ and $n_2 \times p$ random matrix with (real) i.i.d. entries satisfying

$$\mathbb{E} z_{ij}^{(\alpha)} = 0, \qquad \mathbb{E} |z_{ij}^{(\alpha)}|^2 = n^{-1}, \tag{C.1}$$

where we denote $n := n_1 + n_2$. Here we have chosen the scaling that is more standard in the random matrix theory literature—under this $n^{-1/2}$ scaling, the eigenvalues of $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are all of order 1. Moreover, we assume that the fourth moment exists:

$$\mathbb{E} |\sqrt{n} z_{ij}^{(\alpha)}|^4 \leqslant C \tag{C.2}$$

for some constant $C > 0$. We assume that the aspect ratios $d_1 := p/n_1$ and $d_2 := p/n_2$ satisfy that

$$0 \leqslant d_1 \leqslant \tau^{-1}, \quad 1 + \tau \leqslant d_2 \leqslant \tau^{-1}, \tag{C.3}$$

for some small constant $0 < \tau < 1$. Here the lower bound $1 + \tau \leqslant d_2$ is to ensure that the covariance matrix $\mathcal{Q}_2$ for the target task is non-singular with high probability; see Lemma D.2 below.

We assume that Σ_1 and Σ_2 have eigendecompositions

$$\Sigma_1 = O_1 \Lambda_1 O_1^\top, \quad \Sigma_2 = O_2 \Lambda_2 O_2^\top, \quad \Lambda_1 = \operatorname{diag}(\sigma_1^{(1)}, \dots, \sigma_n^{(1)}), \quad \widetilde{\Sigma} = \operatorname{diag}(\sigma_1^{(2)}, \dots, \sigma_N^{(2)}), \tag{C.4}$$

where the eigenvalues satisfy that

$$\tau^{-1} \geqslant \sigma_1^{(1)} \geqslant \sigma_2^{(1)} \geqslant \dots \geqslant \sigma_p^{(1)} \geqslant 0, \quad \tau^{-1} \geqslant \sigma_1^{(2)} \geqslant \sigma_2^{(2)} \geqslant \dots \geqslant \sigma_p^{(2)} \geqslant \tau, \tag{C.5}$$

for some small constant $0 < \tau < 1$. We assume that $M := \Sigma_1^{1/2}\Sigma_2^{-1/2}$ has singular value decomposition

$$M = U\Lambda V^\top, \quad \Lambda = \mathrm{diag}(\sigma_1, \ldots, \sigma_p), \tag{C.6}$$

where the singular values satisfy that

$$\tau \leqslant \sigma_p \leqslant \sigma_1 \leqslant \tau^{-1} \tag{C.7}$$

for some small constant $0 < \tau < 1$.

We summarize our basic assumptions here for future reference.

Assumption C.1. *We assume that Z_1 and Z_2 are independent $n_1 \times p$ and $n_2 \times p$ random matrices with real $i.i.d.$ entries satisfying (C.1) and* (C.2)*, Σ_1 and Σ_2 are deterministic non-negative definite symmetric matrices satisfying* (C.4)-(C.7)*, and $d_{1,2}$ satisfy* (C.3)*.*

Before giving the main proof, we first introduce some notations and tools.

C.1 Notations

We will use the following notion of stochastic domination, which was first introduced in [**?**] and subsequently used in many works on random matrix theory, such as [**?**, **?**, **?**, **?**, **?**, **?**]. It simplifies the presentation of the results and their proofs by systematizing statements of the form "ξ is bounded by ζ with high probability up to a small power of n".

Definition C.2 (Stochastic domination)**.** *(i) Let*

$$\xi = \left(\xi^{(n)}(u) : n \in \mathbb{N}, u \in U^{(n)}\right), \quad \zeta = \left(\zeta^{(n)}(u) : n \in \mathbb{N}, u \in U^{(n)}\right)$$

be two families of nonnegative random variables, where $U^{(n)}$ is a possibly N-dependent parameter set. We say ξ is stochastically dominated by ζ, uniformly in u, if for any fixed (small) $\varepsilon > 0$ and (large) $D > 0$,

$$\sup_{u \in U^{(n)}} \mathbb{P}\left[\xi^{(n)}(u) > N^{\varepsilon}\zeta^{(n)}(u)\right] \leqslant N^{-D}$$

for large enough $n \geqslant n_0(\varepsilon, D)$, and we shall use the notation $\xi \prec \zeta$. Throughout this paper, the stochastic domination will always be uniform in all parameters that are not explicitly fixed (such as matrix indices, and z that takes values in some compact set). If for some complex family ξ we have $|\xi| \prec \zeta$, then we will also write $\xi \prec \zeta$ or $\xi = \mathrm{O}_\prec(\zeta)$.

(ii) We say an event Ξ holds with high probability if for any constant $D > 0$, $\mathbb{P}(\Xi) \geqslant 1 - n^{-D}$ for large enough n. We say Ξ holds with high probability on an event Ω if for any constant $D > 0$, $\mathbb{P}(\Omega \setminus \Xi) \leqslant n^{-D}$ for large enough n

The following lemma collects basic properties of stochastic domination $\prec$, which will be used tacitly in the proof.

Lemma C.3 (Lemma 3.2 in [**?**])**.** *Let ξ and ζ be families of nonnegative random variables.*

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

(ii) If $\xi_1(u) \prec \zeta_1(u)$ and $\xi_2(u) \prec \zeta_2(u)$ uniformly in $u \in U$, then $\xi_1(u)\xi_2(u) \prec \zeta_1(u)\zeta_2(u)$ uniformly in u.

(iii) Suppose that $\Psi(u) \geqslant n^{-C}$ is deterministic and $\xi(u)$ satisfies $\mathbb{E}\xi(u)^2 \leqslant n^C$ for all u. Then if $\xi(u) \prec \Psi(u)$ uniformly in u, we have $\mathbb{E}\xi(u) \prec \Psi(u)$ uniformly in u.

Definition C.4 (Bounded support condition)**.** *We say a random matrix Z satisfies the bounded support condition with q, if*

$$\max_{i,j} |x_{ij}| \prec q. \tag{C.8}$$

Here $q \equiv q(N)$ is a deterministic parameter and usually satisfies $n^{-1/2} \leqslant q \leqslant n^{-\phi}$ for some (small) constant $\phi > 0$. Whenever (C.8) holds, we say that X has support q.

Our main goal is to study the following matrix inverse

$$(\mathcal{Q}_1 + \mathcal{Q}_2)^{-1} = \left(\Sigma_1^{1/2} Z_1^\top Z_1 \Sigma_1^{1/2} + \Sigma_2^{1/2} Z_2^\top Z_2 \Sigma_2^{1/2}\right)^{-1}.$$

Using (C.6), we can rewrite it as

$$\Sigma_2^{-1/2} V \left(\Lambda U^\top Z_1^\top Z_1 U \Lambda + V^\top Z_2^\top Z_2 V\right)^{-1} V^\top \Sigma_2^{-1/2}. \tag{C.9}$$

For this purpose, we shall study the following matrix for $z \in \mathbb{C}_+$,

$$\mathcal{G}(z) := \left(\Lambda U^\top Z_1^\top Z_1 U \Lambda + V^\top Z_2^\top Z_2 V - z\right)^{-1}, \quad z \in \mathbb{C}_+, \tag{C.10}$$

which we shall refer to as resolvent (or Green's function).

Now we introduce a convenient self-adjoint linearization trick. This idea dates back at least to Girko, see e.g., the works [**?**, **?**, **?**] and references therein. It has been proved to be useful in studying the local laws of random matrices of the Gram type [**?**, **?**, **?**, **?**]. We define the following $(p+n) \times (p+n)$ self-adjoint block matrix, which is a linear function of X:

$$H \equiv H(Z_1, Z_2) := \begin{pmatrix} 0 & \Lambda U^\top Z_1^\top & V^\top Z_2^\top \\ Z_1 U \Lambda & 0 & 0 \\ Z_2 V & 0 & 0 \end{pmatrix}. \tag{C.11}$$

Then we define its resolvent (Green's function) as

$$G \equiv G(Z_1, Z_2, z) := \left[H(Z_1, Z_2) - \begin{pmatrix} z I_{p\times p} & 0 & 0 \\ 0 & I_{n_1 \times n_1} & 0 \\ 0 & 0 & I_{n_2 \times n_2} \end{pmatrix} \right]^{-1}, \quad z \in \mathbb{C}_+. \tag{C.12}$$

For simplicity of notations, we define the index sets

$$\mathcal{I}_1 := [\![1, p]\!], \quad \mathcal{I}_2 := [\![p+1, p+n_1]\!], \quad \mathcal{I}_3 := [\![p+n_1+1, p+n_1+n_2]\!], \quad \mathcal{I} := \mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3.$$

We will consistently use the latin letters $i, j \in \mathcal{I}_1$ and greek letters $\mu, \nu \in \mathcal{I}_2 \cup \mathcal{I}_3$. Moreover, we shall use the notations $\mathfrak{a}, \mathfrak{b} \in \mathcal{I} := \cup_{i=1}^3 \mathcal{I}_i$. We label the indices of the matrices according to

$$Z_1 = (z_{\mu i} : i \in \mathcal{I}_1, \mu \in \mathcal{I}_2), \quad Z_2 = (z_{\nu i} : i \in \mathcal{I}_1, \nu \in \mathcal{I}_3).$$

Then we denote the $\mathcal{I}_1 \times \mathcal{I}_1$ block of $G(z)$ by $\mathcal{G}_L(z)$, the $\mathcal{I}_1 \times (\mathcal{I}_2 \cup \mathcal{I}_3)$ by $\mathcal{G}_{LR}$, the $(\mathcal{I}_2 \cup \mathcal{I}_3) \times \mathcal{I}_1$ block by $\mathcal{G}_{RL}$, and the $(\mathcal{I}_2 \cup \mathcal{I}_3) \times (\mathcal{I}_2 \cup \mathcal{I}_3)$ block by $\mathcal{G}_R$. For simplicity, we abbreviate $Y_1 := Z_1 U \Lambda$, $Y_2 := Z_2 V$ and $W := (Y_1^\top, Y_2^\top)$. By Schur complement formula, one can find that (recall (C.10))

$$\mathcal{G}_{11} = \left(WW^\top - z\right)^{-1} = \mathcal{G}, \quad \mathcal{G}_{LR} = \mathcal{G}_{RL}^\top = \mathcal{G}W, \quad \mathcal{G}_R := \begin{pmatrix} \mathcal{G}_{22} & \mathcal{G}_{23} \\ \mathcal{G}_{32} & \mathcal{G}_{33} \end{pmatrix} = z\left(W^\top W - z\right)^{-1}. \tag{C.13}$$

Thus a control of G yields directly a control of the resolvent $\mathcal{G}$. We also introduce the following random quantities (some partial traces and weighted partial traces):

$$\begin{aligned} m(z) &:= \frac{1}{p} \sum_{i \in \mathcal{I}_1} G_{ii}(z), \quad m_1(z) := \frac{1}{p} \sum_{i \in \mathcal{I}_1} \sigma_i^2 G_{ii}(z), \\ m_2(z) &:= \frac{1}{n_1} \sum_{\mu \in \mathcal{I}_2} G_{\mu\mu}(z), \quad m_3(z) := \frac{1}{n_2} \sum_{\mu \in \mathcal{I}_3} G_{\mu\mu}(z). \end{aligned} \tag{C.14}$$

Next we introduce the spectral decomposition of G. Let

$$W = \sum_{k=1}^p \sqrt{\lambda_k} \xi_k \zeta_k^\top,$$

be a singular value decomposition of W, where

$$\lambda_1 \geqslant \Lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0 = \lambda_{p+1} = \ldots = \lambda_n,$$

$\{\xi_k\}_{k=1}^p$ are the left-singular vectors, and $\{\zeta_k\}_{k=1}^n$ are the right-singular vectors. Then using (C.13), we can get that for $i, j \in \mathcal{I}_1$ and $\mu, \nu \in \mathcal{I}_2$,

$$\begin{aligned} G_{ij} &= \sum_{k=1}^{p} \frac{\xi_k(i)\xi_k^\top(j)}{\lambda_k - z}, \quad G_{\mu\nu} = z\sum_{k=1}^{p} \frac{\zeta_k(\mu)\zeta_k^\top(\nu)}{\lambda_k - z} - \sum_{k=p+1}^{n} \zeta_k(\mu)\zeta_k^\top(\nu), \\ G_{i\mu} &= \sum_{k=1}^{p} \frac{\sqrt{\lambda_k}\xi_k(i)\zeta_k^\top(\mu)}{\lambda_k - z}, \quad G_{\mu i} = \sum_{k=1}^{p} \frac{\sqrt{\lambda_k}\zeta_k(\mu)\xi_k^\top(i)}{\lambda_k - z}. \end{aligned} \tag{C.15}$$

We now define the deterministic limit of $\mathcal{G}(z)$. We first define the deterministic limits of $(m_2(z), m_3(z))$, that is $(m_{2c}(z), m_{3c}(z))$, as the (unique) solution to the following system of self-consistent equations

$$\frac{1}{m_{2c}} = -1 + \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{\sigma_i^2}{z + \sigma_i^2 r_1 m_{2c} + r_2 m_{3c}}, \quad \frac{1}{m_{3c}} = -1 + \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{1}{z + \sigma_i^2 r_1 m_{2c} + r_2 m_{3c}}, \tag{C.16}$$

such that $(m_{2c}(z), m_{3c}(z)) \in \mathbb{C}_+$ for $z \in \mathbb{C}_+$, where, for simplicity, we introduce the parameters

$$\gamma_n := \frac{p}{n}, \quad r_1 \equiv r_1(n) := \frac{n_1}{n}, \quad r_2 \equiv r_2(n) := \frac{n_2}{n}. \tag{C.17}$$

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

$$\Pi(z) := \begin{pmatrix} -(z + r_1 m_{2c}\Lambda^2 + r_2 m_{3c})^{-1} & 0 & 0 \\ 0 & m_{2c}(z) I_{n_1} & 0 \\ 0 & 0 & m_{3c}(z) I_{n_2} \end{pmatrix}. \tag{C.18}$$

In particular, the matrix limit of $\mathcal{G}(z)$ is given by $-(z + r_1 m_{2c}\Lambda^2 + r_2 m_{3c})^{-1}$.

If $z = 0$, then the equations (C.16) is reduced to

$$r_1 b_2 + r_2 b_3 = 1 - \gamma_n, \quad b_2 + \frac{1}{n}\sum_{i=1}^{p} \frac{\sigma_i^2 b_2}{\sigma_i^2 r_1 b_2 + (1 - \gamma_n - r_1 b_2)} = 1. \tag{C.19}$$

where $b_2 := -m_{2c}(0)$ and $b_3 := -m_{3c}(0)$. Note that the function

$$f(b_2) := b_2 + \frac{1}{n}\sum_{i=1}^{p} \frac{\sigma_i^2 b_2}{\sigma_i^2 b_2 + (1 - \gamma_n - r_1 b_2)}$$

is a strictly increasing function on $[0, r_1^{-1}(1-\gamma_n)]$, and $f(0) = 0 < 1$, $f(r_1^{-1}(1-\gamma_n)) = 1+\gamma_n > 1$. Hence there exists a unique solution (b_2, b_3) to (C.19). Moreover, it is easy to check that $f'(a) = \mathrm{O}(1)$ for $a \in [0, r_1^{-1}(1-\gamma_n)]$, and $f(1) > 1$ if $1 \leqslant r_1^{-1}(1-\gamma_n)$. Hence there exists a constant $\tau > 0$, such that

$$r_1\tau \leqslant r_1 b_2 < \min\{(1-\gamma_n) - r_1\tau, r_1(1-\tau)\}, \quad \tau < r_3 b_3 \leqslant 1 - \gamma_n - r_1\tau. \tag{C.20}$$

For general z around $z = 0$, the existence and uniqueness of the solution $(m_{2c}(z), m_{3c}(z))$ is given by the following lemma. Moreover, we will also include some basic estimates on it. (say something about the previous work)

Lemma C.5. *There exist constants $c_0, C_0 > 0$ depending only on τ in* (C.3)*,* (C.5)*,* (C.7) *and* (C.20) *such that the following statements hold. There exists a unique solution to* (C.16) *under the conditions*

$$|z| \leqslant c_0, \quad |m_{2c}(z) - m_{2c}(0)| + |m_{3c}(z) - m_{3c}(0)| \leqslant c_0. \tag{C.21}$$

Moreover, the solution satisfies

$$\max_{al=2}^{3} |m_{\alpha c}(z) - m_{\alpha c}(0)| \leqslant C_0|z|. \tag{C.22}$$

The proof is a standard application of the contraction principle. For reader's convenience, we will gives its proof in Appendix D.4. As a byproduct of the contraction mapping argument there, we also obtain the following stability result that will be useful for our proof of Theorem C.7 below.

Lemma C.6. *There exist constants $c_0, C_0 > 0$ depending only on τ in* (C.3), (C.5), (C.7) *and* (C.20) *such that the self-consistent equations in* (C.16) *are stable in the following sense. Suppose $|z| \leqslant c_0$ and $m_\alpha : \mathbb{C}_+ \mapsto \mathbb{C}_+$, $\alpha = 2, 3$, are analytic functions of z such that*

$$|m_2(z) - m_{2c}(0)| + |m_3(z) - m_{3c}(0)| \leqslant c_0.$$

Suppose they satisfy the system of equations

$$\frac{1}{m_2} + 1 - \frac{1}{n}\sum_{i=1}^{p} \frac{\sigma_i^2}{z + \sigma_i^2 r_1 m_2 + r_2 m_3} = \mathcal{E}_2, \quad \frac{1}{m_3} + 1 - \frac{1}{n}\sum_{i=1}^{p} \frac{1}{z + \sigma_i^2 r_1 m_2 + r_2 m_3} = \mathcal{E}_3, \tag{C.23}$$

for some (random) errors satisfying

$$\max_{\alpha=2}^{3} |\mathcal{E}_\alpha| \leqslant \delta(z),$$

where $\delta(z)$ is any deterministic z-dependent function $\delta(z) \leqslant (\log n)^{-1}$. Then we have

$$\max_{\alpha=2}^{3} |m_\alpha(z) - m_{\alpha c}(z)| \leqslant C_0 \delta(z). \tag{C.24}$$

In the following proof, we choose a sufficiently small constants $c_0 > 0$ such that Lemma C.5 and Lemma C.6 hold. Then we define a domain of the spectral parameter z as

$$\mathbf{D} := \left\{ z = E + \mathrm{i}\eta \in \mathbb{C}_+ : |z| \leqslant (\log n)^{-1} \right\}. \tag{C.25}$$

The following theorem gives almost optimal estimates on the resolvent G, which are conventionally called local laws.

Theorem C.7. *Suppose Assumption C.1 holds, and Z_1, Z_2 satisfy the bounded support condition* (C.8) *for some deterministic parameter $q \equiv q(n)$ satisfying $n^{-1/2} \leqslant q \leqslant n^{-\phi}$ for some (small) constant $\phi > 0$. Then there exists a sufficiently small constant $c_0 > 0$ such that the following* **anisotropic local law** *holds uniformly for all $z \in \mathbf{D}$. For any deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{\mathcal{I}}$, we have*

$$\left|\mathbf{u}^\top (G(z) - \Pi(z))\mathbf{v}\right| \prec q. \tag{C.26}$$

The proof of this theorem will be given in Section D. Using a simple cutoff argument, it is easy to obtain the following corollary under certain moment assumptions.

Corollary C.8. *Suppose Assumption C.1 holds. Moreover, assume that the entries of Z_1 and Z_2 are i.i.d. random variables satisfying* (C.1) *and*

$$\max_{i,j} \mathbb{E}|\sqrt{n} z_{ij}^{(\alpha)}|^a = \mathrm{O}(1), \quad \alpha = 1, 2, \tag{C.27}$$

for some fixed $a > 4$. Then (C.26) *holds for $q = n^{2/a - 1/2}$ on an event with probability* $1 - \mathrm{o}(1)$.

Proof of Corollary C.8. Fix any sufficiently small constant $\varepsilon > 0$. We then choose $q = n^{-c_a + \varepsilon}$ with $c_a = 1/2 - 2/a$. Then we introduce the truncated matrices $\widetilde{Z}_1$ and $\widetilde{Z}_2$, with entries

$$\tilde{z}_{ij}^{(\alpha)} := \mathbf{1}\left\{ |\tilde{z}_{ij}^{(\alpha)}| \leqslant q \right\} \cdot z_{ij}^{(\alpha)}, \quad \alpha = 1, 2.$$

By the moment conditions (C.27) and a simple union bound, we have

$$\mathbb{P}(\widetilde{Z}_1 = Z_1, \widetilde{Z}_2 = Z_2) = 1 - \mathrm{O}(n^{-a\varepsilon}). \tag{C.28}$$

Using (C.27) and integration by parts, it is easy to verify that

$$\mathbb{E}|z_{ij}^{(\alpha)}| 1_{|z_{ij}^{(\alpha)}| > q} = \mathrm{O}(n^{-2-\varepsilon}), \quad \mathbb{E}|z_{ij}^{(\alpha)}|^2 1_{|z_{ij}^{(\alpha)}| > q} = \mathrm{O}(n^{-2-\varepsilon}), \quad \alpha = 1, 2,$$

which imply that

$$|\mathbb{E}\tilde{z}_{ij}^{(\alpha)}| = \mathrm{O}(n^{-2-\varepsilon}), \quad \mathbb{E}|\tilde{z}_{ij}^{(\alpha)}|^2 = n^{-1} + \mathrm{O}(n^{-2-\varepsilon}), \quad \alpha = 1, 2, . \tag{C.29}$$

Moreover, we trivially have

$$\mathbb{E}|\tilde{z}_{ij}^{(\alpha)}|^4 \leqslant \mathbb{E}|z_{ij}^{(\alpha)}|^4 = \mathrm{O}(n^{-2}), \quad \alpha = 1, 2.$$

Then we centralize and rescale $\widetilde{Z}_1$ and $\widetilde{Z}_2$ as

$$\widehat{Z}_\alpha := \frac{\widetilde{Z}_\alpha - \mathbb{E}\widetilde{Z}_\alpha}{(\mathbb{E}|\widetilde{z}_{11}^{(\alpha)}|^2)^{1/2}}, \quad \alpha = 1, 2.$$

Now $\widehat{Z}_1$ and $\widehat{Z}_2$ satisfy the assumptions in Theorem C.7 with $q = n^{-c_a+\varepsilon}$, and (C.26) gives that

$$\left|\mathbf{u}^\top (G(\widehat{Z}_1, \widehat{Z}_2, z) - \Pi(z))\mathbf{v}\right| \prec q.$$

Then using (C.29) and (D.4) below, we can easily get that

$$\left|\mathbf{u}^\top (G(\widehat{Z}_1, \widehat{Z}_2, z) - G(\widetilde{Z}_1, \widetilde{Z}_2, z))\mathbf{v}\right| \prec n^{-1-\varepsilon},$$

where we also used the bound $\|\mathbb{E}\widetilde{Z}_\alpha\| = \mathrm{O}(n^{-1-\varepsilon})$. This shows that (C.26) also holds for $G(\widetilde{Z}_1, \widetilde{Z}_2, z)$ with $q = n^{-c_a+\varepsilon}$, and hence concludes the proof by (C.28). $\square$

Using Corollary C.8, we can complete the proof of Lemma A.2 and Lemma B.2.

Proof of Lemma A.2. In the setting of Lemma A.2, we can write

$$\mathcal{R} := (w^2 X_1^\top X_1 + X_2^\top X_2)^{-1} = n^{-1} \left(\widetilde{\Sigma}_1^{1/2} Z_1^\top Z_1 \widetilde{\Sigma}_1^{1/2} + \Sigma_2^{1/2} Z_2^\top Z_2 \Sigma_2^{1/2}\right)^{-1},$$

where $\widetilde{\Sigma}_1 := w^2\Sigma_1$, Σ_2, Z_1 and Z_2 satisfy Assumption C.1. Here the extra n^{-1} is due to the choice of the variances—in the setting of Lemma A.2 the variances of the entries of $Z_{1,2}$ are equal to 1, while in (C.1) they are taken to be n^{-1}. As in (C.6), we assume that $M := \widetilde{\Sigma}_1^{1/2}\Sigma_2^{-1/2}$ has singular value decomposition

$$M = U\Lambda V^\top, \quad \Lambda = \mathrm{diag}(\sigma, \dots, \sigma_p). \tag{C.30}$$

Then as in (C.9), we can write

$$\mathcal{R} = \Sigma_2^{-1/2} V \mathcal{G}(0) V^\top \Sigma_2^{-1/2}, \quad \mathcal{G}(0) = \left(\Lambda U^\top Z_1^\top Z_1 U \Lambda + V^\top Z_2^\top Z_2 V\right)^{-1}.$$

Now by Corollary C.8, we obtain that for any small constant $\varepsilon > 0$, with probability $1 - \mathrm{o}(1)$,

$$\max_{1\leqslant i\leqslant p} |(\Sigma_2 \mathcal{R} - \Sigma_2^{1/2} V \Pi(0) V^\top \Sigma_2^{-1/2})_{ii}| \leqslant n^\varepsilon q, \quad q = n^{2/a-1/2}, \tag{C.31}$$

where by (C.18),

$$\Pi(0) = -(r_1 m_{2c}(0)\Lambda^2 + r_2 m_{3c}(0))^{-1} = (r_1 b_2 V^\top M^\top M V + r_2 b_3)^{-1},$$

with (b_2, b_3) satisfying (C.19). Thus from (C.31) we get that

$$n^{-1}\operatorname{Tr}(\Sigma_2 \mathcal{R}) = n^{-1}\operatorname{Tr}(r_1 b_2 M^\top M + r_2 b_3)^{-1} + \mathrm{O}(n^\varepsilon q)$$

with probability $1 - \mathrm{o}(1)$. This concludes (A.4) if we rename $r_1 b_2 \to a_1$ and $r_2 b_3 \to a_2$. For (A.4), it is a well-known result for inverse Whishart matrices (add some references). In fact, if we set $n_1 = 0$ and $n_2 = n$, then it is easy to check that $a_1 = 0$ and $a_2 = (n_2 - p)/n_2$ is the solution to (4.2). This gives (**??**) by (A.4). $\square$

Proof of Lemma B.2. In the setting of Lemma B.2, we can write

$$\begin{aligned}\Delta &:= n^2 \left\|\Sigma_2^{1/2}(\hat{w}^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_1(\beta_s - w\beta_t)\right\|^2 \\ &= (\beta_s - w\beta_t)\Sigma_1^{1/2} M \left(M^\top Z_1^\top Z_1 M + Z_2^\top Z_2\right)^{-2} M^\top \Sigma_1^{1/2}(\beta_s - w\beta_t),\end{aligned}$$

where $\widetilde{\Sigma}_1 := w^2\Sigma_1$, Σ_2, Z_1 and Z_2 satisfy Assumption C.1 and $M := \widetilde{\Sigma}_1^{1/2}\Sigma_2^{-1/2}$. Here again the n^2 factor disappears due to the choice of scaling. Again we assume that M has the singular value decomposition (C.30). Then we can write

$$\Delta := \mathbf{v}^\top (\mathcal{G}^2)(0)\,\mathbf{v}, \quad \mathbf{v} := V^\top M^\top \Sigma_1^{1/2}(\beta_s - w\beta_t).$$

Note that $\mathcal{G}^2(0) = \partial_z \mathcal{G}|_{z=0}$. Now using Cauchy's integral formula and Corollary C.8, we get that with probability $1 - \mathrm{o}(1)$,

$$\mathbf{v}^\top \mathcal{G}^2(0)\,\mathbf{v} = \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{\mathbf{v}^\top \mathcal{G}(z)\,\mathbf{v}}{z^2} \mathrm{d}z = \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{\mathbf{v}^\top \Pi(z)\,\mathbf{v}}{z^2} \mathrm{d}z + \mathrm{O}_\prec(q) = \mathbf{v}^\top \Pi'(0)\,\mathbf{v} + \mathrm{O}_\prec(q), \tag{C.32}$$

where $\mathcal{C}$ is the contour $\{z \in \mathbb{C} : |z| \leqslant (\log n)^{-1}\}$ and we used (C.26) in the second step. Hence it remains to study the derivatives

$$\mathbf{v}^\top \Pi'(0)\,\mathbf{v} = \mathbf{v}\, \frac{1 + r_1 m_{2c}'(0)\Lambda^2 + r_2 m_{3c}'(0)}{(r_1 m_{2c}(0)\Lambda^2 + r_2 m_{3c}(0))^2}\,\mathbf{v}\,. \tag{C.33}$$

It remains to calculate the derivatives $m_{2c}'(0)$ and $m_{3c}'(0)$.

By the implicit differentiation of (C.16), we obtain that

$$\begin{aligned} \frac{1}{m_{2c}^2(0)} m_{2c}'(0) &= \frac{1}{n} \sum_{i=1}^p \frac{\sigma_i^2 \left(1 + \sigma_i^2 r_1 m_{2c}'(0) + r_2 m_{3c}'(0)\right)}{(\sigma_i^2 r_1 m_{2c}(0) + r_2 m_{3c}(0))^2}, \\ \frac{1}{m_{3c}^2(0)} m_{3c}'(0) &= \frac{1}{n} \sum_{i=1}^p \frac{1 + \sigma_i^2 r_1 m_{2c}'(0) + r_2 m_{3c}'(0)}{(\sigma_i^2 r_1 m_{2c}(0) + r_2 m_{3c}(0))^2}. \end{aligned}$$

If we rename $-r_1 m_{2c}(0) \to a_1$, $-r_2 m_{3c}(0) \to a_2$, $r_2 m_{3c}'(0) \to a_3$ and $r_1 m_{2c}'(0) \to a_4$, then this equation becomes

$$\begin{aligned} &\left(\frac{r_2}{a_2^2} - \frac{1}{n}\sum_{i=1}^p \frac{1}{(\sigma_i^2 a_1 + a_2)^2}\right) a_3 - \left(\frac{1}{n}\sum_{i=1}^p \frac{\sigma_i^2}{(\sigma_i^2 a_1 + a_2)^2}\right) a_4 = \frac{1}{n}\sum_{i=1}^p \frac{1}{(\sigma_i^2 a_1 + a_2)^2}, \\ &\left(\frac{r_1}{a_1^2} - \frac{1}{n}\sum_{i=1}^p \frac{\sigma_i^4}{(\sigma_i^2 a_1 + a_2)^2}\right) a_4 - \left(\frac{1}{n}\sum_{i=1}^p \frac{\sigma_i^2}{(\sigma_i^2 a_1 + a_2)^2}\right) a_3 = \frac{1}{n}\sum_{i=1}^p \frac{\sigma_i^2}{(\sigma_i^2 a_1 + a_2)^2}. \end{aligned} \tag{C.34}$$

Then by (C.32) and (C.33), we get

$$\begin{aligned} \Delta &= (\beta_s - w\beta_t)^\top \Sigma_1^{1/2} M V \frac{1 + a_3 + a_4 \Lambda^2}{(a_1 \Lambda^2 + a_2)} V^\top M^\top \Sigma_1^{1/2} (\beta_s - w\beta_t) \\ &= (\beta_s - w\beta_t)^\top \Sigma_1^{1/2} M \frac{1 + a_3 + a_4 M^\top M}{(a_1 M^\top M + a_2)} M^\top \Sigma_1^{1/2} (\beta_s - w\beta_t) \end{aligned}$$

where we used $M^\top M = V\Lambda^2 V^\top$ in the second step. This concludes Lemma B.2. $\square$

D Proof of Theorem C.7

The main difficulty for the proof of Theorem C.7 is due to the fact that the entries of $Y_1 = Z_1 U \Lambda$ and $Y_2 = Z_2 V$ are not independent. However, notice that if the entries of $Z_1 \equiv Z_1^{Gauss}$ and $Z_2 \equiv Z_2^{Gauss}$ are i.i.d. Gaussian, then by the rotational invariance of the multivariate Gaussian distribution, we have

$$Z_1^{Gauss} U \Lambda \overset{d}{=} Z_1^{Gauss} \Lambda, \quad Z_2^{Gauss} V \overset{d}{=} Z_2^{Gauss}.$$

In this case, the problem is reduced to proving the anisotropic local law for G with $U = \mathrm{Id}$ and $V = \mathrm{Id}$, such that the entries of Y_1 and Y_2 are independent. This can be handled using the standard resolvent methods as in e.g. [**?**, **?**, **?**]. To go from the Gaussian case to the general X case, we will adopt a continuous self-consistent comparison argument developed in [**?**].

For the case $U = \mathrm{Id}$ and $V = \mathrm{Id}$, we need to deal with the following resolvent:

$$G_0(z) := \begin{pmatrix} -z & \Lambda Z_1^\top & Z_2^\top \\ Z_1 \Lambda & -I_{n_1} & 0 \\ Z_2 & 0 & -I_{n_2} \end{pmatrix}^{-1}, \quad z \in \mathbb{C}_+, \tag{D.1}$$

and prove the following result.

Proposition D.1. *Suppose Assumption C.1 holds, and Z_1, Z_2 satisfy the bounded support condition* (C.8) *with $q = n^{-1/2}$. Suppose U and V are identity. Then the estimate* (C.26) *holds for $G_0(z)$.*

In Section D.1, we will collect some a priori estimates and resolvent identities that will be used in the proof of Theorem C.7 and Proposition D.1. Then in Section D.2 we give the proof of Proposition D.1, which, as discussed above, concludes Theorem C.7 for i.i.d. Gaussian Z_1 and Z_2. Finally, in Section D.3, we will describe how to extend the result in Theorem C.7 from the Gaussian case to the case with generally distributed entries of Z_1 and Z_2. In the proof, we always denote the spectral parameter by $z = E + \mathrm{i}\eta$.

D.1 Basic estimates

The estimates in this section work for general G, that is, we do not require U and V to be identity.

First, note that $Z_1^\top Z_1$ (resp. $Z_2^\top Z_2$) is a standard sample covariance matrix, and it is well-known that its nonzero eigenvalues are all inside the support of the Marchenko-Pastur law $[(1-\sqrt{d_1})^2, (1+\sqrt{d_1})^2]$ (resp. $[(1-\sqrt{d_2})^2, (1+\sqrt{d_2})^2]$) with probability $1-\mathrm{o}(1)$ [**?**]. In our proof, we shall need a slightly stronger probability bound, which is given by the following lemma. Denote the nonzero eigenvalues of $Z_1^\top Z_1$ and $Z_2^\top Z_2$ by $\lambda_1(Z_1^\top Z_1) \geqslant \cdots \geqslant \lambda_{p\wedge n_1}(Z_1^\top Z_1)$ and $\lambda_1(Z_2^\top Z_2) \geqslant \cdots \geqslant \lambda_p(Z_2^\top Z_2)$.

Lemma D.2. *Suppose Assumption C.1 holds, and Z_1, Z_2 satisfy the bounded support condition* (C.8) *for some deterministic parameter $q \equiv q(n)$ satisfying $n^{-1/2} \leqslant q \leqslant n^{-\phi}$ for some (small) constant $\phi > 0$. Then for any constant $\varepsilon > 0$, we have with high probability,*

$$\lambda_1(Z_1^\top Z_1) \leqslant (1+\sqrt{d_1})^2 + \varepsilon, \tag{D.2}$$

and

$$(1-\sqrt{d_2})^2 - \varepsilon \leqslant \lambda_p(Z_2^\top Z_2) \leqslant \lambda_1(Z_2^\top Z_2) \leqslant (1+\sqrt{d_2})^2 + \varepsilon. \tag{D.3}$$

Proof. This lemma essentially follows from [**?**, Theorem 2.10], although the authors considered the case with $q \prec n^{-1/2}$ only. The results for larger q follows from [**?**, Lemma 3.12], but only the bounds for the largest eigenvalues are given there in order to avoid the issue with the smallest eigenvalue when d_2 is close to 1. However, under the assumption (C.3), the lower bound for the smallest eigenvalue follows from the exactly the same arguments as in [**?**]. Hence we omit the details. □

With this lemma, we can easily obtain the following a priori estimate on the resolvent $G(z)$ for $z \in \mathbf{D}$.

Lemma D.3. *Suppose the assumptions of Lemma D.2 holds. Then there exists a constant $C > 0$ such that the following estimates hold uniformly in $z, z' \in \mathbf{D}$ with high probability:*

$$\|G(z)\| \leqslant C, \tag{D.4}$$

and for any deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{\mathcal{I}}$,

$$\left|\mathbf{u}^\top \left[G(z) - G(z')\right] \mathbf{v}\right| \leqslant C|z - z'|. \tag{D.5}$$

Proof. As in (C.15), we let $\{\lambda_k\}_{1\leqslant k\leqslant p}$ be the eigenvalues of $WW^\top$. By Lemma D.2 and the assumption (C.3), we obtain that

$$\lambda_p \geqslant \lambda_p(Z_2^\top Z_2) \geqslant \varepsilon > 0 \tag{D.6}$$

for some constant $\varepsilon > 0$. In particular, it implies that

$$\inf_{z\in\mathbf{D}} \min_{1\leqslant k\leqslant p} |\lambda_k - z| \gtrsim 1.$$

Together with (C.15), it implies the estimates (D.4) and (D.5). □

Now we introduce the concept of minors, which are defined by removing certain rows and columns of the matrix H.

Definition D.4 (Minors). *For any $(p+n)\times(p+n)$ matrix $\mathcal{A}$ and $\mathbb{T}\subseteq\mathcal{I}$, we define the minor $\mathcal{A}^{(\mathbb{T})} := (\mathcal{A}_{\mathfrak{ab}} : \mathfrak{a},\mathfrak{b}\in\mathcal{I}\setminus\mathbb{T})$ as the $(p+n-|\mathbb{T}|)\times(p+n-|\mathbb{T}|)$ matrix obtained by removing all rows and columns indexed by $\mathbb{T}$. Note that we keep the names of indices when defining $\mathcal{A}^{(\mathbb{T})}$, i.e. $(\mathcal{A}^{(\mathbb{T})})_{ab} = \mathcal{A}_{ab}$ for $a,b\notin\mathbb{T}$. Correspondingly, we define the resolvent minor as (recall* (C.13)*)*

$$G^{(\mathbb{T})} := \left[\left(H - \begin{pmatrix} zI_p & 0 \\ 0 & I_n \end{pmatrix}\right)^{(\mathbb{T})}\right]^{-1} = \begin{pmatrix} \mathcal{G}^{(\mathbb{T})} & \mathcal{G}^{(\mathbb{T})}W^{(\mathbb{T})} \\ \left(W^{(\mathbb{T})}\right)^\top \mathcal{G}^{(\mathbb{T})} & \mathcal{G}_R^{(\mathbb{T})} \end{pmatrix},$$

and the partial traces (recall (C.14)*)*

$$\begin{aligned} m^{(\mathbb{T})} &:= \frac{1}{p}\sum_{i\in\mathcal{I}_1} G_{ii}^{(\mathbb{T})}(z), \quad m_1^{(\mathbb{T})} := \frac{1}{p}\sum_{i\in\mathcal{I}_1}^{(\mathbb{T})} \sigma_i^2 G_{ii}^{(\mathbb{T})}(z), \\ m_2^{(\mathbb{T})}(z) &:= \frac{1}{n_1}\sum_{\mu\in\mathcal{I}_2} G_{\mu\mu}^{(\mathbb{T})}(z), \quad m_3^{(\mathbb{T})}(z) := \frac{1}{n_2}\sum_{\mu\in\mathcal{I}_3} G_{\mu\mu}^{(\mathbb{T})}(z), \end{aligned} \tag{D.7}$$

where we abbreviated that $\sum_a^{(\mathbb{T})} := \sum_{a\notin\mathbb{T}}$. For convenience, we will adopt the convention that for any minor $\mathcal{A}^{(T)}$ defined as above, $\mathcal{A}_{ab}^{(T)} = 0$ if $a\in\mathbb{T}$ or $b\in\mathbb{T}$. Moreover, we will abbreviate $(\{a\})\equiv(a)$ and $(\{a,b\})\equiv(ab)$.

Then we record the following resolvent identities.

Lemma D.5. *(Resolvent identities).*

(i) For $i\in\mathcal{I}_1$ and $\mu\in\mathcal{I}_2\cup\mathcal{I}_3$, we have

$$\frac{1}{G_{ii}} = -z - \left(WG^{(i)}W^\top\right)_{ii}, \quad \frac{1}{G_{\mu\mu}} = -1 - \left(W^\top G^{(\mu)}W\right)_{\mu\mu}. \tag{D.8}$$

(ii) For $i\neq j\in\mathcal{I}_1$ and $\mu\neq\nu\in\mathcal{I}_2\cup\mathcal{I}_3$, we have

$$G_{ij} = -G_{ii}\left(WG^{(i)}\right)_{ij}, \quad G_{\mu\nu} = -G_{\mu\mu}\left(W^\top G^{(\mu)}\right)_{\mu\nu}. \tag{D.9}$$

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

$$G_{i\mu} = -G_{ii}\left(WG^{(i)}\right)_{i\mu}, \quad G_{\mu i} = -G_{\mu\mu}\left(W^\top G^{(\mu)}\right)_{\mu i}. \tag{D.10}$$

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

$$G_{bc}^{(a)} = G_{bc} - \frac{G_{ba}G_{ac}}{G_{aa}}, \quad \frac{1}{G_{bb}} = \frac{1}{G_{bb}^{(a)}} - \frac{G_{ba}G_{ab}}{G_{bb}G_{bb}^{(a)}G_{aa}}. \tag{D.11}$$

Proof. All these identities can be proved directly using Schur's complement formula. The reader can also refer to, for example, [**?**, Lemma 4.4]. ☐

The following lemma gives large deviation bounds for bounded supported random variables.

Lemma D.6 (Lemma 3.8 of [**?**]). *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, y_j have variance at most n^{-1} and satisfy the bounded support condition (C.8) with $q\leqslant n^{-\phi}$ for some constant $\phi>0$. Then we have the following bound:*

$$\Big|\sum_i A_i x_i\Big| \prec q\max_i|A_i| + \frac{1}{\sqrt{n}}\Big(\sum_i |A_i|^2\Big)^{1/2}, \quad \Big|\sum_{i,j} x_i B_{ij} y_j\Big| \prec q^2 B_d + qB_o + \frac{1}{n}\Big(\sum_{i\neq j}|B_{ij}|^2\Big)^{1/2}, \tag{D.12}$$

$$\Big|\sum_i \bar{x}_i B_{ii} x_i - \sum_i (\mathbb{E}|x_i|^2)B_{ii}\Big| \prec qB_d, \quad \Big|\sum_{i\neq j}\bar{x}_i B_{ij} x_j\Big| \prec qB_o + \frac{1}{n}\left(\sum_{i\neq j}|B_{ij}|^2\right)^{1/2}, \tag{D.13}$$

where $B_d := \max_i|B_{ii}|$ and $B_o := \max_{i\neq j}|B_{ij}|$.

D.2 Entrywise local law

The main goal of this subsection is to prove the following entrywise local law. The anisotropic local law (C.26) then follows from the entrywise local law combined with a polynomialization method as we will explain in next subsection. Recall that in the setting of Proposition D.1, we have $q = n^{-1/2}$ and

$$W = (\Lambda Z_1^\top, Z_2^\top). \tag{D.14}$$

Lemma D.7. *Suppose the assumptions in Proposition D.1 hold. Then the following estimate holds uniformly for $z \in \mathbf{D}$:*

$$\max_{\mathfrak{a},\mathfrak{b}} |(G_0)_{\mathfrak{a}\mathfrak{b}}(z) - \Pi_{\mathfrak{a}\mathfrak{b}}(z)| \prec n^{-1/2}. \tag{D.15}$$

Proof. The proof of Lemma D.7 is divided into three steps. For simplicity, we will still denote $G \equiv G_0$ in the following proof, while keeping in mind that W takes the form in (D.14).

Step 1: Large deviations estimates. In this step, we prove some (almost) optimal large deviations estimates on the off-diagonal entries of G, and on the following Z variables. In analogy to [**?**, Section 3] and [**?**, Section 5], we introduce the Z variables

$$Z_{\mathfrak{a}}^{(\mathbb{T})} := (1 - \mathbb{E}_{\mathfrak{a}})\big(G_{\mathfrak{a}\mathfrak{a}}^{(\mathbb{T})}\big)^{-1}, \quad \mathfrak{a} \notin \mathbb{T},$$

where $\mathbb{E}_{\mathfrak{a}}[\cdot] := \mathbb{E}[\cdot \mid H^{(\mathfrak{a})}]$, i.e. it is the partial expectation over the randomness of the $\mathfrak{a}$-th row and column of H. By (D.8), we have

$$\begin{aligned} Z_i = (\mathbb{E}_i - 1)\left(WG^{(i)}W^\top\right)_{ii} &= \sigma_i^2 \sum_{\mu,\nu\in\mathcal{I}_2} G_{\mu\nu}^{(i)}\left(\frac{1}{n}\delta_{\mu\nu} - z_{\mu i}z_{\nu i}\right) \\ &\quad + \sum_{\mu,\nu\in\mathcal{I}_3} G_{\mu\nu}^{(i)}\left(\frac{1}{n}\delta_{\mu\nu} - z_{\mu i}z_{\nu i}\right), \quad i \in \mathcal{I}_1, \end{aligned} \tag{D.16}$$

and

$$\begin{aligned} Z_\mu &= (\mathbb{E}_\mu - 1)\left(W^\top G^{(\mu)}W\right)_{\mu\mu} = \sum_{i,j\in\mathcal{I}_1} \sigma_i\sigma_j G_{ij}^{(\mu)}\left(\frac{1}{n}\delta_{ij} - z_{\mu i}z_{\mu j}\right), \quad \mu \in \mathcal{I}_2, \\ Z_\mu &= (\mathbb{E}_\mu - 1)\left(W^\top G^{(\mu)}W\right)_{\mu\mu} = \sum_{i,j\in\mathcal{I}_1} G_{ij}^{(\mu)}\left(\frac{1}{n}\delta_{ij} - z_{\mu i}z_{\mu j}\right), \quad \mu \in \mathcal{I}_3. \end{aligned} \tag{D.17}$$

For simplicity, we introduce the following random error

$$\Lambda_o := \max_{\mathfrak{a}\neq\mathfrak{b}} \left|G_{\mathfrak{a}\mathfrak{a}}^{-1}G_{\mathfrak{a}\mathfrak{b}}\right|. \tag{D.18}$$

The following lemma gives the desired large deviations estimates on the Λ_o and the Z variables.

Lemma D.8. *Suppose the assumptions in Proposition D.1 hold. Then the following estimates hold uniformly for all $z \in \mathbf{D}$:*

$$\Lambda_o + \max_{\mathfrak{a}\in\mathcal{I}} |Z_{\mathfrak{a}}| \prec n^{-1/2}. \tag{D.19}$$

Proof. Note that for any $\mathfrak{a} \in \mathcal{I}$, $H^{(\mathfrak{a})}$ and $G^{(\mathfrak{a})}$ also satisfies the assumptions for Lemma D.3. Hence (D.4) and (D.5) also hold for $G^{(\mathfrak{a})}$. Now applying Lemma D.6 to (D.16) and (D.17), and using the a priori bound (D.4), we get that for any $i \in \mathcal{I}_1$,

$$|Z_i| \lesssim \sum_{\alpha=2}^{3}\left|\sum_{\mu,\nu\in\mathcal{I}_\alpha} G_{\mu\nu}^{(i)}\left(\frac{1}{n}\delta_{\mu\nu} - z_{\mu i}z_{\nu i}\right)\right| \prec n^{-1/2} + \frac{1}{n}\left(\sum_{\mu,\nu\in\mathcal{I}_2\cup\mathcal{I}_3}\left|G_{\mu\nu}^{(i)}\right|^2\right)^{1/2} \prec n^{-1/2},$$

where in the last step we used that for any μ,

$$\sum_{\nu\in\mathcal{I}_2\cup\mathcal{I}_3}\left|G_{\mu\nu}^{(i)}\right|^2 \leqslant \sum_{\mathfrak{a}\in\mathcal{I}}\left|G_{\mu\mathfrak{a}}^{(i)}\right|^2 = \left[G^{(i)}(G^{(i)})^*\right]_{\mu\mu} = \mathrm{O}(1) \tag{D.20}$$

by (D.4). Similarly, applying Lemma D.6 to Z_μ in (D.17) and using (D.4), we obtain the same bound.

Then we prove the off-diagonal estimates. For $i \neq j \in \mathcal{I}_1$ and $\mu \neq \nu \in \mathcal{I}_2 \cup \mathcal{I}_3$, using (D.9), Lemma D.6 and (D.4), we obtain that

$$\left|G_{ii}^{-1} G_{ij}\right| \prec n^{-1/2} + \frac{1}{\sqrt{n}} \left(\sum_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} \left| G_{\mu j}^{(i)} \right|^2 \right)^{1/2} \prec n^{-1/2},$$

and

$$\left|G_{\mu\mu}^{-1} G_{\mu\nu}\right| \prec n^{-1/2} + \frac{1}{\sqrt{n}} \left(\sum_{i \in \mathcal{I}_1} \left| G_{i\nu}^{(\mu)} \right|^2 \right)^{1/2} \prec n^{-1/2}.$$

For $i \in \mathcal{I}_1 \cup \mathcal{I}_2$ and $\mu \in \mathcal{I}_3$, using (D.10), Lemma D.6 and (D.4), we obtain that

$$\left|G_{ii}^{-1} G_{i\mu}\right| + \left|G_{\mu\mu}^{-1} G_{\mu i}\right| \prec n^{-1/2} + \frac{1}{\sqrt{n}} \left(\sum_{\nu \in \mathcal{I}_2 \cup \mathcal{I}_3} \left| G_{\nu\mu}^{(i)} \right|^2 \right)^{1/2} + \frac{1}{\sqrt{n}} \left(\sum_{j \in \mathcal{I}_1} \left| G_{ji}^{(\mu)} \right|^2 \right)^{1/2} \prec n^{-1/2}.$$

Thus we obtain that $\Lambda_o \prec n^{-1/2}$, which concludes (D.19). □

Note that comibining (D.4) and (D.19), we immediately conclude (D.15) for $\mathfrak{a} \neq \mathfrak{b}$.

Step 2: Self-consistent equations. This is the key step of the proof for Proposition D.7, which derives approximate self-consistent equations safisfised by $m_2(z)$ and $m_3(z)$. More precisely, we will show that $(m_2(z), m_3(z))$ satisfies (C.23) up to some small error $|\mathcal{E}_{2,3}| \prec n^{-1/2}$. Then applying Lemma C.6 shows that $(m_2(z), m_3(z))$ is close to $(m_{2c}(z), m_{3c}(z))$—this will discussed in Step 3.

We define the following z-dependent event

$$\Xi(z) := \left\{ |m_2(z) - m_{2c}(z)| + |m_3(z) - m_{3c}(z)| \leqslant (\log n)^{-1/2} \right\}. \tag{D.21}$$

Note that by (C.22), we have $|m_{2c} + b_2| \lesssim (\log n)^{-1}$ and $|m_{3c} + b_3| \lesssim (\log n)^{-1}$. Together with (C.16), (C.20) and (C.7), we obtain the following basic estimates

$$|m_{2c}| \sim 1, \quad |m_{3c}| \sim 1, \quad |z + \sigma_i^2 r_1 m_{2c} + r_2 m_{3c}| \sim 1, \quad |1 + \gamma_n m_c| \sim 1, \quad |1 + \gamma_n m_{1c}| \sim 1, \tag{D.22}$$

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

$$m_c(z) := -\frac{1}{n} \sum_{i \in \mathcal{I}_1} \frac{1}{z + \sigma_i^2 r_1 m_{2c} + r_2 m_{3c}}, \quad m_{1c}(z) := -\frac{1}{n} \sum_{i \in \mathcal{I}_1} \frac{\sigma_i^2}{z + \sigma_i^2 r_1 m_{2c} + r_2 m_{3c}}.$$

Plugging (D.22) into (C.18), we get

$$|\Pi_{\mathfrak{a}\mathfrak{a}}(z)| \sim 1 \text{ uniformly in } z \in \mathbf{D},\ \mathfrak{a} \in \mathcal{I}. \tag{D.23}$$

Then we claim the following result.

Lemma D.9. *Suppose the assumptions in Proposition D.1 hold. Then the following estimates hold uniformly in $z \in \mathbf{D}$:*

$$\begin{aligned} \mathbf{1}(\Xi) \left| \frac{1}{m_2} + 1 - \frac{1}{n} \sum_{i=1}^p \frac{\sigma_i^2}{z + \sigma_i^2 r_1 m_2 + r_2 m_3} \right| &\prec n^{-1/2}, \\ \mathbf{1}(\Xi) \left| \frac{1}{m_3} + 1 - \frac{1}{n} \sum_{i=1}^p \frac{1}{z + \sigma_i^2 r_2 m_2 + r_2 m_3} \right| &\prec n^{-1/2}. \end{aligned} \tag{D.24}$$

Proof. By (D.8), (D.16) and (D.17), we obtain that

$$\frac{1}{G_{ii}} = -z - \frac{\sigma_i^2}{n}\sum_{\mu\in\mathcal{I}_2} G^{(i)}_{\mu\mu} - \frac{1}{n}\sum_{\mu\in\mathcal{I}_3} G^{(i)}_{\mu\mu} + Z_i = -z - \sigma_i^2 r_1 m_2 - r_2 m_3 + \varepsilon_i, \quad i\in\mathcal{I}_1, \tag{D.25}$$

$$\frac{1}{G_{\mu\mu}} = -1 - \frac{1}{n}\sum_{i\in\mathcal{I}_1} \sigma_i^2 G^{(\mu)}_{ii} + Z_\mu = -1 - \gamma_n m_1 + \varepsilon_\mu, \quad \mu\in\mathcal{I}_2, \tag{D.26}$$

$$\frac{1}{G_{\nu\nu}} = -1 - \frac{1}{n}\sum_{i\in\mathcal{I}_1} G^{(\nu)}_{ii} + Z_\nu = -1 - \gamma_n m + \varepsilon_\nu, \quad \nu\in\mathcal{I}_3, \tag{D.27}$$

where we recall (D.7), and

$$\varepsilon_i := Z_i + \sigma_i r_1\left(m_2 - m_2^{(i)}\right) + r_2(m_3 - m_3^{(i)}), \quad \varepsilon_\mu := \begin{cases} Z_\mu + \gamma_n(m_1 - m_1^{(\mu)}), & \text{if } \mu\in\mathcal{I}_2 \\ Z_\mu + \gamma_n(m - m^{(\mu)}), & \text{if } \mu\in\mathcal{I}_3 \end{cases}.$$

By (D.11) we can bound that

$$|m_2 - m_2^{(i)}| \leqslant \frac{1}{n_1}\sum_{\mu\in\mathcal{I}_2}\left|\frac{G_{\mu i}G_{i\mu}}{G_{ii}}\right| \prec n^{-1},$$

where we used (D.19) in the second step. Similarly, we can get that

$$|m - m^{(\mu)}| + |m_1 - m_1^{(\mu)}| + |m_2 - m_2^{(i)}| + |m_3 - m_3^{(i)}| \prec n^{-1} \tag{D.28}$$

for any $i\in\mathcal{I}_1$ and $\mu\in\mathcal{I}_2\cup\mathcal{I}_3$. Together with (D.19), we obtain that for all i and μ,

$$|\varepsilon_i| + |\varepsilon_\mu| \prec n^{-1/2}. \tag{D.29}$$

With (D.22) and the definition of Ξ, we get that $\mathbf{1}(\Xi)|z+\sigma_i^2 r_1 m_2 + r_2 m_3| \sim 1$. Hence using (D.25), (D.29) and (D.19), we obtain that

$$\mathbf{1}(\Xi)G_{ii} = \mathbf{1}(\Xi)\left[-\frac{1}{z+\sigma_i^2 r_1 m_2 + r_2 m_3} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right]. \tag{D.30}$$

Plugging it into the definitions of m and m_1 in (D.7), we get

$$\mathbf{1}(\Xi)m = \mathbf{1}(\Xi)\left[-\frac{1}{p}\sum_{i\in\mathcal{I}_1}\frac{1}{z+\sigma_i^2 r_1 m_2 + r_2 m_3} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right], \tag{D.31}$$

$$\mathbf{1}(\Xi)m_1 = \mathbf{1}(\Xi)\left[-\frac{1}{p}\sum_{i\in\mathcal{I}_1}\frac{\sigma_i^2}{z+\sigma_i^2 r_1 m_2 + r_2 m_3} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right]. \tag{D.32}$$

As a byproduct, we obtain from the two estimates that

$$\mathbf{1}(\Xi)\left(|m-m_c| + |m_1 - m_{1c}|\right) \lesssim (\log n)^{-1/2}, \quad \text{with high probability.} \tag{D.33}$$

Together with (D.22), we get that

$$|1+\gamma_n m_1| \sim 1, \quad |1+\gamma_n m| \sim 1, \quad \text{with high probability on } \Xi. \tag{D.34}$$

Now using (D.26), (D.27), (D.29), (D.19) and (D.34), we can obtain that with high probability,

$$\mathbf{1}(\Xi)G_{\mu\mu} = \mathbf{1}(\Xi)\left[-\frac{1}{1+\gamma_n m_1} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right], \quad \mu\in\mathcal{I}_2, \tag{D.35}$$

$$\mathbf{1}(\Xi)G_{\nu\nu} = \mathbf{1}(\Xi)\left[-\frac{1}{1+\gamma_n m} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right], \quad \nu\in\mathcal{I}_3. \tag{D.36}$$

Taking average over $\mu\in\mathcal{I}_2$ and $\nu\in\mathcal{I}_3$, we get that with high probability,

$$\mathbf{1}(\Xi)m_2 = \mathbf{1}(\Xi)\left[-\frac{1}{1+\gamma_n m_1} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right], \quad \mathbf{1}(\Xi)m_3 = \mathbf{1}(\Xi)\left[-\frac{1}{1+\gamma_n m} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right], \tag{D.37}$$

which further implies

$$\mathbf{1}(\Xi)\left(\frac{1}{m_2}+1+\gamma_n m_1\right)\prec n^{-1/2},\quad \mathbf{1}(\Xi)\left(\frac{1}{m_3}+1+\gamma_n m\right)\prec n^{-1/2}. \tag{D.38}$$

Finally, plugging (D.31) and (D.32) into (D.38), we conclude (D.24). □

Step 3: Ξ holds with high probability. In this step, we show that the event $\Xi(z)$ in fact holds with high probability for all $z\in\mathbf{D}$. Once we have proved this fact, then applying Lemma C.6 to (D.24) immediately shows that $(m_2(z),m_3(z))$ is equal to $(m_{2c}(z),m_{3c}(z))$ up to an error of order $n^{-1/2}$.

First we claim that it suffices to show that

$$|m_2(0)-m_{2c}(0)|+|m_3(0)-m_{3c}(0)|\prec n^{-1/2}. \tag{D.39}$$

Once we know (D.39), then by (C.22) and (D.5), we know $\max_{\alpha=2}^3|m_{\alpha c}(z)-m_{\alpha c}(0)| = \mathrm{O}((\log n)^{-1})$ and $\max_{\alpha=2}^3|m_{\alpha}(z)-m_{\alpha}(0)| = \mathrm{O}((\log n)^{-1})$ with high probability for $z\in\mathbf{D}$. Together with (D.39), we obtain that

$$|m_2(z)-m_{2c}(z)|+|m_3(z)-m_{3c}(z)|\lesssim(\log n)^{-1}\quad\text{with high probability.} \tag{D.40}$$

and

$$\sup_{z\in\mathbf{D}}\left(|m_2(z)-m_{2c}(0)|+|m_3(z)-m_{3c}(0)|\right)\lesssim(\log n)^{-1}\quad\text{with high probability.} \tag{D.41}$$

The condition (D.40) shows that Ξ holds with high probability, and the condition (D.41) verifies the condition (C.21) of Lemma C.6. Hence applying Lemma C.6 to (D.24), we obtain that

$$|m_2(z)-m_{2c}(z)|+|m_3(z)-m_{3c}(z)|\prec n^{-1/2} \tag{D.42}$$

for all $z\in\mathbf{D}$. Plugging (D.42) into (D.25)-(D.27), we get the diagonal estimate

$$\max_{\mathfrak{a}\in\mathcal{I}}|G_{\mathfrak{aa}}(z)-\Pi_{\mathfrak{aa}}(z)|\prec n^{-1/2}. \tag{D.43}$$

Together with the off-diagonal estimate in (D.19), we conclude (D.15).

Lemma D.10. *Under the assumptions in Proposition D.1, the estimate* (D.39) *holds.*

Proof. By (C.15), we get

$$m(0)=\frac{1}{p}\sum_{i\in\mathcal{I}_1}G_{ii}(0)=\frac{1}{p}\sum_{k=1}^{p}\frac{|\xi_k(i)|^2}{\lambda_k}\geqslant\lambda_1^{-1}\gtrsim 1.$$

Similarly, we can also get that $m_1(0)$ is positive and has size $m_1(0)\sim 1$. Hence we have

$$1+\gamma_n m_1(0)\sim 1,\quad 1+\gamma_n m_1(0)\sim 1.$$

Together with (D.26), (D.27) and (D.29), we obtain that (D.37) and (D.38) hold at $z=0$ even without the indicator function $\mathbf{1}(\Xi)$. Furthermore, it gives that

$$\left|\sigma_i^2 r_1 m_2(0)+r_2 m_3(0)\right|=\left|\frac{\sigma_i^2 r_1}{1+\gamma_n m_1(0)}+\frac{r_2}{1+\gamma_n m(0)}+\mathrm{O}_\prec(n^{-1/2})\right|\sim 1$$

with high probability. Then using (D.25) and (D.29), we obtain that (D.31) and (D.32) hold at $z=0$ even without the indicator function $\mathbf{1}(\Xi)$. Finally, plugging (D.31) and (D.32) into (D.38), we conclude (D.24) holds at $z=0$, that is,

$$\begin{aligned}&\left|\frac{1}{m_2(0)}+1-\frac{1}{n}\sum_{i=1}^{p}\frac{\sigma_i^2}{\sigma_i^2 r_1 m_2(0)+r_2 m_3(0)}\right|\prec n^{-1/2},\\&\left|\frac{1}{m_3(0)}+1-\frac{1}{n}\sum_{i=1}^{p}\frac{1}{\sigma_i^2 r_2 m_2(0)+r_2 m_3(0)}\right|\prec n^{-1/2}.\end{aligned} \tag{D.44}$$

Denoting $\omega_2 = -m_{2c}(0)$ and $\omega_3 = -m_{2c}(0)$. By (D.38), we have

$$\omega_2 = \frac{1}{1+\gamma_n m_1(0)} + \mathrm{O}_\prec(n^{-1/2}), \quad \omega_3 = \frac{1}{1+\gamma_n m(0)} + \mathrm{O}_\prec(n^{-1/2}).$$

Hence there exists a sufficiently small constant $c > 0$ such that

$$c \leqslant \omega_2 \leqslant 1, \quad c \leqslant \omega_3 \leqslant 1, \quad \text{with high probability.} \tag{D.45}$$

Moreover, one can verify from (D.44) that (ω_2, ω_3) satisfy approximately the same equations as in (C.19):

$$r_1\omega_2 + r_2\omega_3 = 1 - \gamma_n + \mathrm{O}_\prec(n^{-1/2}), \quad f(\omega_2) = 1 + \mathrm{O}_\prec(n^{-1/2}). \tag{D.46}$$

The first equation and (D.45) together implies that $\omega_2 \in [0, r_1^{-1}(1-\gamma_n)]$ with high probability. Since f is strictly increasing and has bounded derivatives on $[0, r_1^{-1}(1-\gamma_n)]$, by basic calculus the second equation in (D.46) gives that $|\omega_2 - b_2| \prec n^{-1/2}$. Together with the first equation in (D.46), we get $|\omega_3 - b_3| \prec n^{-1/2}$. This concludes (D.39). $\square$

This lemma concludes (D.39), and as explained above, concludes the proof of Lemma D.7. $\square$

With Lemma D.7, we can conclude the proof of Proposition D.1.

Proof of Proposition D.1. With (D.15), one can repeat the polynomialization method in [**?**, Section 5] to get the anisotropic local law (C.26) for G_0. The proof is exactly the same, except for some minor notation difference, so we omit the details. $\square$

D.3 Anisotropic local law

In this section, we finish the proof of Theorem C.7 for a general X satisfying the bounded support condition (C.8) with $q \leqslant n^{-\phi}$ for some constant $\phi > 0$. The proposition D.1 implies that (C.26) holds for Gaussian Z_1^{Gauss} and Z_2^{Gauss}. Thus the basic idea is to prove that for Z_1 and Z_2 satisfying the assumptions in Theorem C.7,

$$\mathbf{u}^\top \left(G(Z,z) - G(Z^{Gauss}, z)\right)\mathbf{v} \prec q$$

for any deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}}$ and $z \in \mathbf{D}$. Here we abbreviated $Z := \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ and $Z^{Gauss} := \begin{pmatrix} Z_1^{Gauss} \\ Z_2^{Gauss} \end{pmatrix}$. We prove the above statement using a continuous comparison argument introduced in [**?**]. The proof is similar to the ones in Sections 7-8 of [**?**], so we only give an outline without writing down all the details.

Definition D.11 (Interpolating matrices)**.** *We denote Introduce the notations $Z^0 := Z^{Gauss}$ and $Z^1 := Z$. Let $\rho_{\mu i}^0$ and $\rho_{\mu i}^1$ be the laws of $Z_{\mu i}^0$ and $Z_{\mu i}^1$, respectively. For $\theta \in [0,1]$, we define the interpolated law*

$$\rho_{\mu i}^\theta := (1-\theta)\rho_{\mu i}^0 + \theta\rho_{\mu i}^1.$$

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 $Z^\theta = (Z_{\mu i}^\theta)$ has law

$$\prod_{i\in\mathcal{I}_1}\prod_{\mu\in\mathcal{I}_2\cup\mathcal{I}_3} \rho_{\mu i}^\theta(\mathrm{d}Z_{\mu i}^\theta). \tag{D.47}$$

For $\lambda \in \mathbb{R}$, $i \in \mathcal{I}_1$ and $\mu \in \mathcal{I}_2 \cup \mathcal{I}_3$, we define the matrix $Z_{(\mu i)}^{\theta,\lambda}$ through

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

We also introduce the matrices

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

We shall prove (C.26) through interpolation matrices Z^θ between Z^0 and Z^1. We have see that (C.26) holds for Z^0 by Proposition D.1. Using (D.47) and fundamental calculus, we get the following basic interpolation formula: for $F:\mathbb{R}^{n\times p}\to\mathbb{C}$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}F(Z^\theta)=\sum_{i\in\mathcal{I}_1}\sum_{\mu\in\mathcal{I}_2\cup\mathcal{I}_3}\left[\mathbb{E}F\left(Z^{\theta,Z^1_{\mu i}}_{(\mu i)}\right)-\mathbb{E}F\left(Z^{\theta,Z^0_{\mu i}}_{(\mu i)}\right)\right] \tag{D.48}$$

provided all the expectations exist.

We shall apply (D.48) to $F(Z):=F^p_{\mathbf{u}\,\mathbf{v}}(Z,z)$ for (large) $p\in 2\mathbb{N}$ and $F_{\mathbf{v}}(Z,z)$ defined as

$$F_{\mathbf{u}\,\mathbf{v}}(Z,z):=\left|G_{\mathbf{uv}}(Z,z)-\Pi_{\mathbf{uv}}(z)\right|. \tag{D.49}$$

Here for simplicity of notations, we introduce the following notation of generalized entries: for $\mathbf{u},\mathbf{v}\in\mathbb{R}^{\mathcal{I}}$, we shall denote $G_{\mathbf{uv}}:=\mathbf{u}^\top G\mathbf{v}$. Moreover, we shall abbreviate $G_{\mathbf{u}\mathfrak{a}}:=G_{\mathbf{u}\,\mathbf{e}_\mathfrak{a}}$ for $\mathfrak{a}\in\mathcal{I}$, where $\mathbf{e}_\mathfrak{a}$ is the standard unit vector along $\mathfrak{a}$-th axis. Given any vector $\mathbf{u}\in\mathbb{R}^{\mathcal{I}_{1,2,3}}$, we always identify it with its natural embedding in $\mathbb{R}^{\mathcal{I}}$. The exact meanings will be clear from the context. The main work is to show the following self-consistent estimate for the right-hand side of (D.48) for any fixed $p\in 2\mathbb{N}$ and constant $\varepsilon>0$:

$$\sum_{i\in\mathcal{I}_1}\sum_{\mu\in\mathcal{I}_2\cup\mathcal{I}_3}\left[\mathbb{E}F^p_{\mathbf{u}\,\mathbf{v}}\left(Z^{\theta,Z^1_{\mu i}}_{(\mu i)},z\right)-\mathbb{E}F^p_{\mathbf{u}\,\mathbf{v}}\left(Z^{\theta,Z^0_{\mu i}}_{(\mu i)},z\right)\right]=\mathrm{O}\left((n^\varepsilon q)^p+\mathbb{E}F^p_{\mathbf{u}\,\mathbf{v}}\left(Z^\theta,z\right)\right) \tag{D.50}$$

for all $\theta\in[0,1]$. If (D.50) holds, then combining (D.48) with a Grönwall's argument we obtain that for any fixed $p\in 2\mathbb{N}$ and constant $\varepsilon>0$:

$$\mathbb{E}\left|G_{\mathbf{uv}}(Z^1,z)-\Pi_{\mathbf{uv}}(z)\right|^p\leqslant(n^\varepsilon q)^p$$

Together with Markov's inequality, we conclude (C.26). In order to prove (D.50), 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

$$\sum_{i\in\mathcal{I}_1}\sum_{\mu\in\mathcal{I}_2\cup\mathcal{I}_3}\left[\mathbb{E}F^p_{\mathbf{u}\,\mathbf{v}}\left(Z^{\theta,Z^a_{\mu i}}_{(\mu i)},z\right)-\mathbb{E}F^p_{\mathbf{v}}\left(Z^{\theta,0}_{(\mu i)},z\right)\right]=\mathrm{O}\left((n^\varepsilon q)^p+\mathbb{E}F^p_{\mathbf{u}\,\mathbf{v}}\left(Z^\theta,z\right)\right) \tag{D.51}$$

for all $a\in\{0,1\}$ and $\theta\in[0,1]$. Underlying the proof of (D.51) is an expansion approach. We define the $\mathcal{I}\times\mathcal{I}$ matrix $\Delta^\lambda_{(\mu i)}$ as

$$\Delta^\lambda_{(\mu i)}:=\lambda\begin{pmatrix}0 & \mathbf{u}_i^{(\mu)}\mathbf{e}_\mu^\top\\ \mathbf{e}_\mu(\mathbf{u}_i^{(\mu)})^\top & 0\end{pmatrix}, \tag{D.52}$$

where we denote $\mathbf{u}_i^{(\mu)}:=\Lambda U\mathbf{e}_i$ if $\mu\in\mathcal{I}_2$ and $\mathbf{u}_i^{(\mu)}:=V\mathbf{e}_i$ if $\mu\in\mathcal{I}_3$. Then by the definition of H in (C.11)), we have for any $\lambda,\lambda'\in\mathbb{R}$ and $K\in\mathbb{N}$,

$$G^{\theta,\lambda'}_{(i\mu)}=G^{\theta,\lambda}_{(\mu i)}+\sum_{k=1}^K G^{\theta,\lambda}_{(\mu i)}\left(\Delta^{\lambda-\lambda'}_{(\mu i)}G^{\theta,\lambda}_{(\mu i)}\right)^k+G^{\theta,\lambda'}_{(\mu i)}\left(\Delta^{\lambda-\lambda'}_{(\mu i)}G^{\theta,\lambda}_{(\mu i)}\right)^{K+1}. \tag{D.53}$$

Using this expansion and the a priori bound (D.4), it is easy to prove the following estimate: if y is a random variable satisfying $|y|\prec q$, then

$$G^{\theta,y}_{(\mu i)}=\mathrm{O}(1),\quad i\in\mathcal{I}_1,\ \mu\in\mathcal{I}_2\cup\mathcal{I}_3, \tag{D.54}$$

with high probability.

In the following proof, for simplicity of notations, we introduce $f_{(\mu i)}(\lambda):=F^p_{\mathbf{v}}(Z^{\theta,\lambda}_{(\mu i)})$. We use $f^{(r)}_{(\mu i)}$ to denote the r-th derivative of $f_{(\mu i)}$. By (D.54), it is easy to see that for any fixed $r\in\mathbb{N}$, $f^{(r)}_{(\mu i)}(y)=\mathrm{O}(1)$ with high probability for any random variable y satisfying $|y|\prec q$. Then the Taylor expansion of $f_{(\mu i)}$ gives

$$f_{(\mu i)}(y)=\sum_{r=0}^{p+4}\frac{y^r}{r!}f^{(r)}_{(\mu i)}(0)+\mathrm{O}_\prec\left(q^{p+4}\right), \tag{D.55}$$

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

$$\mathbb{E}F_{\mathbf{v}}^{p}\left(Z_{(\mu i)}^{\theta,Z_{\mu i}^{a}}\right)-\mathbb{E}F_{\mathbf{v}}^{p}\left(Z_{(\mu i)}^{\theta,0}\right)=\mathbb{E}\left[f_{(\mu i)}\left(Z_{i\mu}^{a}\right)-f_{(\mu i)}(0)\right]$$
$$=\mathbb{E}\,f_{(\mu i)}(0)+\frac{1}{2n}\,\mathbb{E}\,f_{(\mu i)}^{(2)}(0)+\sum_{r=4}^{p+4}\frac{1}{r!}\,\mathbb{E}\,f_{(\mu i)}^{(r)}(0)\,\mathbb{E}\left(Z_{i\mu}^{a}\right)^{r}+\mathrm{O}_{\prec}(q^{p+4}). \tag{D.56}$$

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

$$\mathbb{E}(Z_{\mu i}^{1})^{3}=0,\quad 1\leqslant\mu\leqslant n,\ \ 1\leqslant i\leqslant p. \tag{D.57}$$

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

By (C.2) and the bounded support condition, we have

$$\left|\mathbb{E}\left(Z_{i\mu}^{a}\right)^{r}\right|\prec n^{-2}q^{r-4},\quad r\geqslant 4. \tag{D.58}$$

Thus to show (D.51), we only need to prove for $r=4,5,...,p+4$,

$$n^{-2}q^{r-4}\sum_{i\in\mathcal{I}_1}\sum_{\mu\in\mathcal{I}_2\cup\mathcal{I}_3}\left|\mathbb{E}\,f_{(\mu i)}^{(r)}(0)\right|=\mathrm{O}\left((n^{\varepsilon}q)^{p}+\mathbb{E}F_{\mathbf{u}\,\mathbf{v}}^{p}(Z^{\theta},z)\right). \tag{D.59}$$

In order to get a self-consistent estimate in terms of the matrix Z^{θ} on the right-hand side of (D.59), we want to replace $Z_{(\mu i)}^{\theta,0}$ in $f_{(\mu i)}(0)=F_{\mathbf{u}\,\mathbf{v}}^{p}(Z_{(\mu i)}^{\theta,0})$ with $Z^{\theta}=Z_{(\mu i)}^{\theta,Z_{\mu i}^{\theta}}$.

Lemma D.12. *Suppose that*

$$n^{-2}q^{r-4}\sum_{i\in\mathcal{I}_1}\sum_{\mu\in\mathcal{I}_2\cup\mathcal{I}_3}\left|\mathbb{E}\,f_{(\mu i)}^{(r)}(Z_{i\mu}^{\theta})\right|=\mathrm{O}\left((n^{\varepsilon}q)^{p}+\mathbb{E}F_{\mathbf{v}}^{p}(X^{\theta},z)\right) \tag{D.60}$$

holds for $r=4,...,4p+4$. Then (D.59) holds for $r=4,...,4p+4$.

Proof. The proof is the same as the one for [**?**, Lemma 7.16]. □

What remains now is to prove (D.60). For simplicity of notations, we shall abbreviate $Z^{\theta}\equiv Z$. For any $k\in\mathbb{N}$, we denote

$$A_{\mu i}(k):=\left(\frac{\partial}{\partial Z_{\mu i}}\right)^{k}\left(G_{\mathbf{uv}}-\Pi_{\mathbf{uv}}\right).$$

The derivative on the right-hand side can be calculated using the expansion (D.53). In particular, it is easy to verify that it satisfies the following bound

$$|A_{\mu i}(k)|\prec\begin{cases}(\mathcal{R}_i^{(\mu)})^2+\mathcal{R}_{\mu}^2, & \text{if } k\geqslant 2\\ \mathcal{R}_i^{(\mu)}\mathcal{R}_{\mu}, & \text{if } k=1\end{cases}, \tag{D.61}$$

where for $i\in\mathcal{I}_1$ and $\mu\in\mathcal{I}_2\cup\mathcal{I}_3$, we denote

$$\mathcal{R}_i^{(\mu)}:=|G_{\mathbf{u}\,\mathbf{u}_i^{(\mu)}}|+|G_{\mathbf{v}\,\mathbf{u}_i^{(\mu)}}|,\quad \mathcal{R}_{\mu}:=|G_{\mathbf{u}\mu}|+|G_{\mathbf{v}\mu}|. \tag{D.62}$$

Then we can calculate the derivative

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

Then to prove (D.60), it suffices to show that

$$n^{-2}q^{r-4}\sum_{i\in\mathcal{I}_1}\sum_{\mu\in\mathcal{I}_2\cup\mathcal{I}_3}\left|\mathbb{E}\prod_{t=1}^{p/2}A_{\mu i}(k_t)\overline{A_{\mu i}(k_{t+p/2})}\right|=\mathrm{O}\left((n^{\varepsilon}q)^{p}+\mathbb{E}F_{\mathbf{u}\,\mathbf{v}}^{p}(Z,z)\right) \tag{D.63}$$

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

$$n^{-2} q^{r-4} \sum_{i \in \mathcal{I}_1} \sum_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} \mathbb{E}\, |A_{\mu i}(0)|^{p-l} \prod_{t=1}^{l} |A_{\mu i}(k_t)| = \mathrm{O}\left((n^{\varepsilon} q)^p + \mathbb{E} F^p_{\mathbf{u}\,\mathbf{v}}(Z, z)\right) \tag{D.64}$$

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

Now we first consider the case $r \leqslant 2l-2$. Then by pigeonhole principle, there exist at least two k_t's with $k_t = 1$. Therefore by (D.61) we have

$$\prod_{t=1}^{l} |A_{\mu i}(k_t)| \prec \mathbf{1}(r \geqslant 2l-1) \left[(\mathcal{R}_i^{(\mu)})^2 + \mathcal{R}_\mu^2\right] + \mathbf{1}(r \leqslant 2l-2)(\mathcal{R}_i^{(\mu)})^2 \mathcal{R}_\mu^2. \tag{D.65}$$

Using (D.4) and a similar argument as in (D.20), we get that

$$\sum_{i \in \mathcal{I}_1} (\mathcal{R}_i^{(\mu)})^2 = \mathrm{O}(1), \qquad \sum_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} \mathcal{R}_\mu^2 = \mathrm{O}(1), \quad \text{with high probability.} \tag{D.66}$$

Using (D.66) and $n^{-1/2} \leqslant q$, we get that

$$\begin{aligned} n^{-2} q^{r-4} \sum_{i \in \mathcal{I}_1} \sum_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} |A_{\mu i}(0)|^{p-l} \prod_{t=1}^{l} |A_{\mu i}(k_t)| &\prec q^{r-4} F^{p-l}_{\mathbf{u}\,\mathbf{v}}(Z) \left[\mathbf{1}(r \geqslant 2l-1) n^{-1} + \mathbf{1}(r \leqslant 2l-2) n^{-2}\right] \\ &\leqslant F^{p-l}_{\mathbf{u}\,\mathbf{v}}(Z) \left[\mathbf{1}(r \geqslant 2l-1) q^{r-2} + \mathbf{1}(r \leqslant 2l-2) q^{r}\right]. \end{aligned}$$

If $r \leqslant 2l-2$, then we get $q^r \leqslant q^l$ using 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 $q^r \leqslant q^{l+2}$. Therefore we conclude that

$$n^{-2} q^{r-4} \sum_{i \in \mathcal{I}_1} \sum_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} |A_{\mu i}(0)|^{p-l} \prod_{t=1}^{l} |A_{\mu i}(k_t)| \prec F^{p-l}_{\mathbf{u}\,\mathbf{v}}(Z) q^l.$$

Now (D.64) follows from Hölder's inequality. This concludes the proof of (D.60), and hence of (D.51), and hence of (C.26).

Finally, if the condition (D.57) does not hold, then there is also an $r = 3$ term in the Taylor expansion (D.56):

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

Note that $\mathbb{E}\left(Z^a_{i\mu}\right)^3$ is of order $n^{-3/2}$, while the sum over i and μ in (D.51) provides a factor n^2. In fact, $\mathbb{E}\, f^{(3)}_{(\mu i)}(0)$ will provide an extra $n^{-1/2}$ to compensate the remaining $n^{1/2}$ factor. This follows from an improved self-consistent comparison argument for sample covariance matrices in [**?**, Section 8]. The argument for our case is almost the same except for some notational differences, so we omit the details.

D.4 Proof of Lemma C.5 and Lemma C.6

Finally, we give the proof of Lemma C.5 and Lemma C.6 using the contraction principle.

Proof of Lemma C.5. One can check that the equations in (C.16) are equivalent to the following ones:

$$r_1 m_{2c} = -(1-\gamma_n) - r_2 m_{3c} - z\left(\frac{1}{m_{3c}} + 1\right), \quad g_z(m_{3c}(z)) = 1, \tag{D.67}$$

where

$$g_z(m_{3c}) := -m_{3c} + \frac{1}{n} \sum_{i=1}^{p} \frac{m_{3c}}{z - \sigma_i^2(1-\gamma_n) + (1-\sigma_i^2) r_2 m_{3c} - \sigma_i^2 z \left(m_{3c}^{-1} + 1\right)}.$$

We first show that there exists a unique solution $m_{3c}(z)$ to the equation $g_z(m_{3c}(z)) = r_2$ under the conditions in (C.21), and the solution satisfies (C.22). Now we abbreviate $\varepsilon(z) := m_{3c}(z) - m_{3c}(0)$, and from (D.67) we can obtain that

$$0 = [g_z(m_{3c}(z)) - g_0(m_{3c}(0)) - g_z'(m_{3c}(0))\varepsilon(z)] + g_z'(m_{3c}(0))\varepsilon(z),$$

which implies

$$g_z'(m_{3c}(0))\varepsilon(z) = -\left[g_z(m_{3c}(0)) - g_0(m_{3c}(0))\right] - \left[g_z(m_{3c}(0) + \varepsilon(z)) - g_z(m_{3c}(0)) - g_z'(m_{3c}(0))\varepsilon(z)\right]. \tag{D.68}$$

Inspired by the above equation, we define iteratively a sequence of vectors $\varepsilon^{(k)} \in \mathbb{C}$ such that $\varepsilon^{(0)} = 0$, and

$$\varepsilon^{(k+1)} = -\frac{g_z(m_{3c}(0)) - g_0(m_{3c}(0))}{g_z'(m_{3c}(0))} - \frac{g_z(m_{3c}(0) + \varepsilon^{(k)}) - g_z(m_{3c}(0)) - g_z'(m_{3c}(0))\varepsilon^{(k)}}{g_z'(m_{3c}(0))}. \tag{D.69}$$

In other words, the above equation defines a mapping $h : \mathbb{C} \to \mathbb{C}$, which maps $\varepsilon^{(k)}$ to $\varepsilon^{(k+1)} = h(\varepsilon^{(k)})$.

With direct calculation, one can get the derivative

$$g_z'(m_{3c}(0)) = -1 - \frac{1}{n}\sum_{i=1}^{p} \frac{\sigma_i^2(1-\gamma_n) - z\left[1 - \sigma_i^2\left(2r_2 m_{3c}^{-1} + 1\right)\right]}{\left[z - \sigma_i^2(1-\gamma_n) + (1-\sigma_i^2)m_{3c} - \sigma_i^2 z\left(r_2 m_{3c}^{-1} + 1\right)\right]^2}.$$

Using (C.20), it is easy to check that there exist constants $\widetilde{c}, \widetilde{C} > 0$ depending only on τ in (C.7) and (C.20) such that

$$|[g_z'(m_{3c}(0))]^{-1}| \leqslant \widetilde{C}, \quad \left|\frac{g_z(m_{3c}(0) + \varepsilon_1) - g_z(m_{3c}(0) + \varepsilon_2) - g_z'(m_{3c}(0))(\varepsilon_1 - \varepsilon_2)}{g_z'(m_{3c}(0))}\right| \leqslant \widetilde{C}|\varepsilon_1 - \varepsilon_2|^2, \tag{D.70}$$

and

$$\left|\frac{g_z(m_{3c}(0)) - g_0(m_{3c}(0))}{g_z'(m_{3c}(0))}\right| \leqslant \widetilde{C}|z|, \tag{D.71}$$

for all $|z| \leqslant \widetilde{c}$ and $|\varepsilon_1| \leqslant \widetilde{c}$, $|\varepsilon_2| \leqslant \widetilde{c}$. Then with (D.70) and (D.71), it is easy to see that there exists a sufficiently small constant $\delta > 0$ depending only on $\widetilde{C}$, such that h is a self-mapping

$$h : B_r \to B_r, \quad B_r := \{\varepsilon \in \mathbb{C} : |\varepsilon| \leqslant r\},$$

as long as $r \leqslant \delta$ and $|z| \leqslant c_\delta$ for some constant $c_\delta > 0$ depending only on $\widetilde{C}$ and δ. Now it suffices to prove that h restricted to B_r is a contraction, which then implies that $\varepsilon := \lim_{k\to\infty} \varepsilon^{(k)}$ exists and $m_{3c}(0) + \varepsilon$ is a unique solution to the second equation of (D.67) subject to the condition $\|\varepsilon\|_\infty \leqslant r$.

From the iteration relation (D.69), using (D.70) one can readily check that

$$\varepsilon^{(k+1)} - \varepsilon^{(k)} = h(\varepsilon^{(k)}) - h(\varepsilon^{(k-1)}) \leqslant \widetilde{C}|\varepsilon^{(k)} - \varepsilon^{(k-1)}|^2. \tag{D.72}$$

Hence as long as r is chosen to be sufficiently small such that $2r\widetilde{C} \leqslant 1/2$, then h is indeed a contraction mapping on B_r, which proves both the existence and uniqueness of the solution $m_{3c}(z) = m_{3c}(0) + \varepsilon$, if we choose c_0 in (C.21) as $c_0 = \min\{c_\delta, r\}$. After obtaining $m_{3c}(z)$, we can then find $m_{2c}(z)$ using the first equation in (D.67).

Note that with (D.71) and $\varepsilon^{(0)} = \mathbf{0}$, we get from (D.69) that

$$|\varepsilon^{(1)}| \leqslant \widetilde{C}|z|.$$

With the contraction mapping, we have the bound

$$|\varepsilon| \leqslant \sum_{k=0}^{\infty} \|\varepsilon^{(k+1)} - \varepsilon^{(k)}\|_\infty \leqslant 2\widetilde{C}|z|. \tag{D.73}$$

This gives the bound (C.22) for $m_{3c}(z)$. Using the first equation in (D.67), we immediately obtain the bound

$$r_1|m_{2c}(z) - m_{2c}(0)| \leqslant C|z|.$$

This gives (C.22) for $m_{2c}(z)$ as long as if $r_1 \gtrsim 1$. To deal with the small r_1 case, we go back to the first equation in (C.16) and treat $m_{2c}(z)$ as the solution to the following equation:

$$\widetilde{g}_z(m_{2c}(z)) = 1, \quad \widetilde{g}_z(x) := -x + \frac{\gamma_n}{p}\sum_{i=1}^p \frac{\sigma_i^2 x}{z + \sigma_i^2 r_1 x + r_2 m_{3c}(z)}.$$

We can calculate that

$$g_z'(m_{2c}(0)) = -1 + \frac{\gamma_n}{p}\sum_{i=1}^p \frac{\sigma_i^2(z + r_2 m_{3c}(z))}{(z + \sigma_i^2 r_1 m_{2c}(0) + r_2 m_{3c}(z))^2}.$$

At $z = 0$, we have

$$|g_0'(m_{2c}(0))| = \left| 1 + \frac{\gamma_n}{p}\sum_{i=1}^p \frac{\sigma_i^2 r_2 b_3}{(\sigma_i^2 r_1 b_2 + r_2 b_3)^2} \right| \geqslant 1,$$

where b_2 and b_3 satisfy (C.20). Thus under (C.21) we have $|g_z'(m_{2c}(0))| \sim 1$ as long as c_0 is taken sufficiently small. Then with the above arguments for $m_{3c}(z)$ between (D.67) and (D.73), we can conclude (C.22) for $m_{2c}(z)$. This concludes the proof of Lemma C.5. □

Proof of Lemma C.6. Under (C.21), we can obtain equation (D.67) approximately up to some small error

$$r_1 m_{2c} = -(1-\gamma_n) - r_2 m_{3c} - z\left(\frac{1}{m_{3c}} + 1\right) + \mathcal{E}_2'(z), \quad g_z(m_{3c}(z)) = 1 + \mathcal{E}_3'(z), \tag{D.74}$$

with $|\mathcal{E}_2'(z)| + |\mathcal{E}_3'(z)| = \mathrm{O}(\delta(z))$. Then we subtract the equations (D.67) from (D.74), and consider the contraction principle for the functions $\varepsilon(z) := m_3(z) - m_{3c}(z)$. The rest of the proof is exactly the one for Lemma C.5, so we omit the details. □