Tight Generalization Bounds and Their Insights for Multi-Task Learning in High-Dimensions

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

In multi-task learning, having related task data is fundamental to its performance. Multi-task learning is particularly powerful when there is limited labeled data for a task to be solved, meanwhile more labeled data from different but related tasks is available. For example, many applications in computer vision [34], natural language processing [], and many other areas have been achieved by learning from multiple tasks together. In spite of these promising empirical results, multi-task learning is not well-understood because of the prevalence of *negative transfer* – when the trained multi-task model performs worse than single-task learning for the target task. [Todo: transition] In this work, we present a theoretical study to better understand when multi-task learning works better than single-task learning. [Todo: practical insight]

The technical challenge to develop a theory for multi-task learning is how to capture generalization performance that scales tightly with the amount of labeled data, in particular when the size of the training set is small. Prior generalization theory using uniform stability [23], Rademacher complexity [5, 4] is unable to explain the phenomenon of negative transfer because there is no tight bound on test error that scales with the amount of labeled data. 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.

Setup. To gain insight into the working of multi-task learning, we consider a simplified setting for learning multiple high-dimensional linear regression tasks. We focus on a hard parameter sharing model proposed in [29, 30] 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), $\ldots$, (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.

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

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. [21, 30]. 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.

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

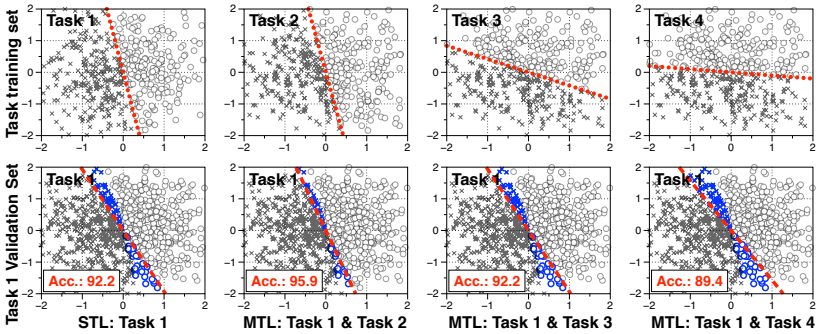


Figure 1: Positive to negative transfer as the source task rotates further away from the target task (left to right).

Quantifying model shift bias versus variance trade-off. 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 [15]. 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.

A crucial technical tool that we develop is the asymptotic limit of the trace of two random matrices. The result extends the well-known Marchenko–Pastur law in random matrix theory to the sum of two independent random matrices, which may be of independent interest.

How task data affects transfer. We identify three factors: task model similarities and their noise level, data size and covariate shift [27, 20]. These are achieved through tight generalization bounds established in the high-dimensional regression setting. Using these tools, we can explain several phenomena that are not explained by the techniques of [30].

- We provide a sharp tr[illegible]ion to show that task model similarities can determine whether there is positive transfer. For settings where tasks are similar, we further show that the transfer effect depends on the single-task accuracy of the source task.
- We show how source task data size can also determine transfer by providing a sharp transition. We use our tools to explain the result of taskonomy [34] [illegible]rding the data efficiency of multi-task learning.
- 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 [30].

Finally, we extend our reuslts to study the transfer functions u[illegible] taskonomy [34].

Experimental results. We provide practical implications to validate our theory.

- We validate on text and image classification tasks that comparing single-task accuracies can help determine whether multi-task learning performs better than single-task learning.
- We show that when the number of source task datapoints is large compared to the target task, then aligning task covariances always improves performance. On the other hand, if the number of source task datapoints is comparable to the target task, aligning task covariances may hurt performance.

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

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

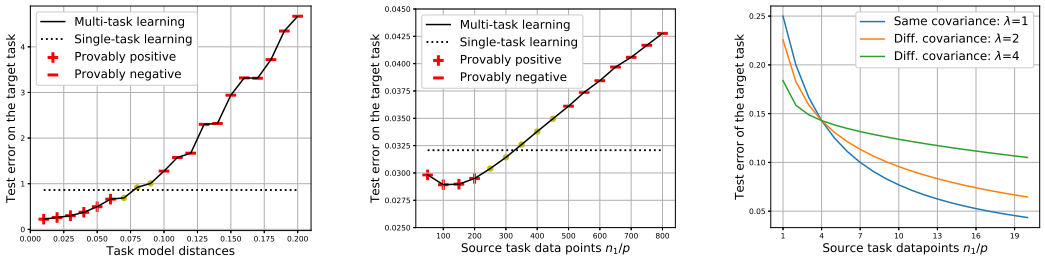
for some constants $c_1, c_2 \in (1,\infty)$. A crucial quantity is what we call the *covariate shift* matrix $M = \Sigma_1^{1/2}\Sigma_2^{-1/2}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ denote the singular values of M.

3 Dissecting the Effects of Different Task Data in Multi-Task learning

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 Task Model Similarity versus Noise

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.



(a) Task model similarity (b) Data size (c) Covariate shift

Figure 2: Comparing the test error of multi-task learning to single-task learning: we observe transitions from positive to negative transfer. (a) Section 3.1: Task model similarities; (b) Section 3.2: Source task noise level; (c) Section 3.3: [Todo: same vs. different covariance matrices].

Example 3.1. 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_2 has i.i.d. entries with mean zero and variance κ^2 and $\beta_1 - \beta_2$ also has i.i.d. entries with mean 0 and variance d^2. We have $n_i = c_i \cdot p$ data points from each task, for $i = 1, 2$.

We illustrate the example in a synthetic setting. We demonstrate our result with a simulation. ([Todo: uses the tighter bound Proposition B.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 C.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 3.2. *In the setting of Example 3.1 with $\sigma_1 = \sigma_2 = \sigma$, assume that $c_1 > 100$ is a fixed constant. 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 $d^2 < \frac{2\sigma^2}{3p} \frac{(c_1+c_2-1)^2}{c_1(c_1+c_2)(c_2-1)}$, then whp we have that $te(\hat{\beta}_t^{MTL}) < te(\hat{\beta}_t^{STL})$.*
- *If $d^2 \geqslant \frac{3\sigma^2}{2p} \frac{(c_1+c_2-1)^2}{c_1(c_1+c_2)(c_2-1)}$, then whp we have that $te(\hat{\beta}_t^{MTL}) \geqslant te(\hat{\beta}_t^{STL})$.*

The proof of Proposition 3.2 involves two parts. First, adding the source task has a positive effect of reducing the variance of the estimator, which scales with $n_1 = c_1 p$, the number of source task data points. Second, the difference between task models β_1 and β_2 introduces an additional bias term, which scales with pd^2, the distance between β_1 and β_2. Hence, the type of transfer is determined by the trade-off between model shift bias and the reduction of variance. The proof can be found in Appendix B.1, which is based on our main result described later in Theorem C.1.

Consider a more general setting where the noise level σ_1 of task 1 differs from the noise level σ_2 of task 2. We derive a sharp transition similar to Proposition 3.2.

Proposition 3.3. *In the setting of Example 3.1 with d being fixed but σ_1 varies, assume that $c_1 > 100$ is a fixed constant and $d^2 < \frac{2\sigma^2}{3p} \frac{(c_1+c_2-1)^2}{c_1(c_1+c_2)(c_2-1)}$. Then we derive the following transition as a parameter of σ_1:*

- *If $\sigma_1^2 \leqslant pd^2 \cdot c_1 + \left[1 + \frac{2}{3} \frac{(c_1+c_2-1)^2}{(c_1+c_2)(c_2-1)}\right] \cdot \sigma_2^2$, then whp $te(\hat{\beta}_t^{MTL}) < te(\hat{\beta}_t^{STL})$.*
- *If $\sigma_1^2 > pd^2 \cdot c_1 + \left[1 + \frac{3}{2} \frac{(c_1+c_2-1)^2}{(c_1+c_2)(c_2-1)}\right] \cdot \sigma_2^2$, then whp $te(\hat{\beta}_t^{MTL}) > te(\hat{\beta}_t^{STL})$.*

Implications. [Todo: add connection to tasksonomy]

3.2 Data Size and Efficiency

In classical Rademacher or VC based theory of multi-task learning, adding more labeled data improves the generalization performance of a model. On the other hand, we have observed that adding more

labeled data does not always improve performance in multi-task learning. Using Example 3.1, we analyze the effect of varying the source task data size.

Proposition 3.4. *In the setting of Example 3.1, assume that $c_2 \geqslant 3$ is fixed and $c_1 > a$ for some fixed integer a. We have the following conditions to determine whether $te(\hat{\beta}_t^{MTL})$ is lower than $te(\hat{\beta}_t^{STL})$:*

- *If $d^2 \leqslant (1+a^{-1/2})^{-4}\frac{\sigma^2}{p(c_2-1)}$, then whp $te(\hat{\beta}_t^{MTL}) < te(\hat{\beta}_t^{STL})$.*
- *If $d^2 > (1-a^{-1/2})^{-4}\frac{\sigma^2}{p(c_2-1)}$, then we have the following transition depending on c_1:*
  - *If $c_1 < \frac{(c_2-2)\sigma^2}{(1+a^{-1/2})^4(c_2-1)pd^2-\sigma^2}$, then whp $te(\hat{\beta}_t^{MTL}) < te(\hat{\beta}_t^{STL})$.*
  - *If $c_1 > \frac{(c_2-2)\sigma^2}{(1-a^{-1/2})^4(1-(a+c_2-2)^{-2})(c_2-1)pd^2-\sigma^2}$, then whp $te(\hat{\beta}_t^{MTL}) > te(\hat{\beta}_t^{STL})$.*

The proof of Proposition 3.4 is similar to Proposition 3.2. We compare the model shift bias and the amount of reduced variance of $\hat{\beta}_t^{\text{MTL}}$. An intuitive interpretation of Proposition 3.4 is that: i) If the two task models are sufficiently similar (as specified under the first bullet), adding the source task always provides positive transfer; ii) Otherwise, as we increase the number of source task data points, the transfer is positive initially, but transitions to negative eventually. We leave the proof of Proposition 3.4 to Appendix B.2.

Implications. We use our tools to explain a key result of task [illegible]my [34], which shows that by learning from multiple related tasks, one can reduce the amount of labeled data from each task. This is formalized by a metric called the data efficiency ratio as follows. Given several tasks, let $\alpha^\star$ be the largest factors such that the total number of labeled datapoints needed for solving all the tasks can be reduced by an $\alpha^\star$ factor (compared to training independently) while keeping the performance nearly the same. More precisely, suppose we have n_i datapoints for each task, for $i = 1, 2$. If we only use αn_i datapoints from every task to train the multi-task learning estimator $\hat{\beta}(\alpha)$, then $\alpha \in (0, 1)$ will be the smallest number such that

$$\alpha^\star := \arg\min_{\alpha\in(0,1)} \; te_1(\hat{\beta}(\alpha)) + te_2(\hat{\beta}(\alpha)) \leqslant te_1(\hat{\beta}_t^{\text{STL}}) + te_2(\hat{\beta}_t^{\text{STL}}).$$

We quantify the data efficiency ratio of $\hat{\beta}_t^{\text{MTL}}$ for Example 3.1 as follows.

Proposition 3.5. *In the setting of Example 3.1, assume that $c_1 = c_2 = c \geqslant 200$ and $d^2 < 8\sigma^2/(3pc)$. Then the data efficiency ratio is at most $\frac{1}{2c} + \frac{\sigma^2}{2\sigma^2-3pd^2c/4}$.*

Note that we have stated the result assuming that $c_1 = c_2$. Similar results can also be obtained when they are different. We omit the details. The proof of Proposition 3.5 can be found in Appendix B.2.

3.3 Covariate Shift

So far we have considered settings where $\Sigma_1 = \Sigma_2$. This setting is relevant for multi-class image classification settings, where different tasks share the same input features. In general, the covariance matrices of the two tasks may be different, e.g. in text classification. In this part, we use our tools to provide a case study on the effect of applying multi-task learning for two tasks when $\Sigma_1 \neq \Sigma_2$.

For this setting, covariate shift is captured by the matrix $M = \Sigma_1^{1/2}\Sigma_2^{-1/2}$. We ask: is it better to have M as being close to identity, or should M involve varying levels of singular values? Understanding this question has implications for applying normalization methods in multi-task learning [22, 9, 33]. Our result shows that if n_1 is much larger than n_2, then the optimal M matrix is equal to identity, under certain assumptions on its range of singular values (to be formulated in Proposition 3.7). On the other hand, if n_1 is comparable or even smaller than n_2, we show an example where having "complementary" covariance matrices is better performing than having the same covariance matrices.

Example 3.6. To compare different choices of M on the performance of $\hat{\beta}_t^{\text{MTL}}$, we assume an upper bound on the scale of M. Consider the following family of matrices

$$\mathcal{S}_\mu := \left\{ M \mid \det\left(M^\top M\right) \leqslant \mu^p, \lambda(M) \in [\mu_{\min}, \mu_{\max}] \right\},$$

where $\mu, \mu_{\min}, \mu_{\max}$ are fixed values that do not grow with p. For the task models, we assume that β_2 has i.i.d. entries with mean zero and variance κ^2 and $\beta_1 - \beta_2$ has i.i.d. entries with mean 0

and variance d^2. The following proposition shows that when n_1 is large enough compared to n_2, $te(\hat{\beta}^{\text{MTL}})$ is minimized approximately within the family of $\mathcal{S}_\mu$ when $M = \mu\,\text{Id}$.

Proposition 3.7. *In the setting of Example 3.1, assume that $c_1 > 3$ and $c_2 > 3$, and $\|\Sigma_1\| \leqslant C_1$ for some constant $C_1 > 0$. As a parameter of $M \in \mathcal{S}_\mu$, we have that $(te(\hat{\beta}_t^{MTL}))(M)$ as a function of M satisfies that*

$$(te(\hat{\beta}_t^{MTL}))(\mu\,\text{Id}) \leqslant \left[1 + C\left(c_2 c_1^{-1} + c_1^{-1/2}\right)\right] \cdot \min_{M \in \mathcal{S}_\mu} (te(\hat{\beta}_t^{MTL}))(M), \tag{3.1}$$

where the constant $C > 0$ depends only on $\mu_{\max}$, $\mu_{\min}$ and C_1, but otherwise does not depend on c_1 and c_2.

Proposition 3.7 shows that when $n_1 \gg n_2$, $te(\hat{\beta}_t^{\text{MTL}})$ is minimized when Σ_1 and Σ_2 are proportional to each other. In other words, there is no covariate shift between the source task data and target task data. This provides evidence that covariate shift is unfavorable when there are many source task datapoints, To complement the result, we show an example when the statement is not true if $n_1 \leqslant n_2$.

Example 3.8. Within the setting of Example 3.6, we compare two cases: (i) when $M = \text{Id}$; (ii) when M has $p/2$ singular values that are equal to λ and $p/2$ singular values that are equal to $1/\lambda$. For simplicity, we assume that $d = 0$. Hence the two tasks have the same model parameters.

In Figure **??**, we plot the test error of the target task for $n_2 = 4p$ and n_1 ranging from p to $20p$. 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 $\Sigma_1 = \Sigma_2$.

A theoretical justification of Example 3.8 can be found in Appendix B.3.

Implications.

4 Technical Tools: Quantifying Model Shift Bias versus Variance Trade-off

We begin by observing that the test error of $\hat{\beta}_t^{\text{MTL}}$ consists of two parts. One part captures how similar the task models are and the other part captures the variance of $\hat{\beta}_t^{\text{MTL}}$. Compared with $\hat{\beta}_t^{\text{STL}}$, we observe that the variance part of $\hat{\beta}_t^{\text{MTL}}$ gets reduced, since more data is added from source tasks. The bias part of $\hat{\beta}_t^{\text{MTL}}$, which we term as *model shift bias*, affects performance negatively. We derive the asympotic limit of $te(\hat{\beta}_t^{\text{MTL}})$ as p approaches infinity. We compare it with the asympotic limit of $te(\hat{\beta}_t\text{STL})$, for settings where the target data size is limited. We show sharp generalization bounds for two settings: i) two tasks with general covaraites; ii) many tasks with the same covariates.

4.1 Two Tasks with General Covariance Matrices

For the case of two tasks, we decompose the test error of $\hat{\beta}_t^{\text{MTL}}$ on the targe task into two parts

$$\begin{aligned} te(\hat{\beta}_t^{\text{MTL}}) =& \hat{v}^2 \left\| \Sigma_2^{1/2} (\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 (\beta_s - \hat{v}\beta_t) \right\|^2 \\ &+ \sigma^2 \cdot \text{Tr}\left[(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_2 \right], \end{aligned} \tag{4.1}$$

where $\hat{v}$ denotes the ratio of the output layer weights (to be defined more precisely in Appendix B). It is not hard to show that the variance of $\hat{\beta}_t^{\text{MTL}}$ is reduced compared to $\hat{\beta}_t^{\text{STL}}$, i.e.

$$\text{Tr}\left[(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_2 \right] \leqslant \text{Tr}\left[(X_2^\top X_2)^{-1} \Sigma_2 \right].$$

Because of model shift bias, we can no longer guarantee that $te(\hat{\beta}_t^{\text{MTL}}) \leqslant te(\hat{\beta}_t^{\text{STL}})$. The technical crux of our approach is to derive the asymptotic limit of $te(\hat{\beta}_t^{\text{MTL}})$ in the high-dimensional setting, when p approaches infinity. A key highlight of our approach implies a precise limit on $\text{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1} \right]$, which only depends on Σ_1, Σ_2 and n_1, n_2 (see Lemma B.2 in Appendix B for the result).

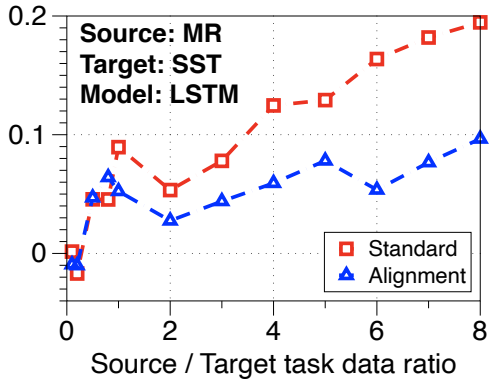


Figure 3: Covariate shift experiment.

Theorem 4.1. *Let $X_i \in \mathbb{R}^{n_i \times p}$ and $Y_i = X_i\beta_i + \varepsilon_i$, for $i = 1, 2$. Suppose that $n_1 = c_1 p$ and $n_2 = c_2 p$, where $c_1 >??$ and $c_2 > 3$ are fixed constants. There exists two deterministic functions Δ_β and Δ_{var} that only depends on $\{\hat{v}, M, n_1, n_2, \beta_1, \beta_2\}$ such that*

- *If $\Delta_{var} - \Delta_\beta \geqslant \delta$, then whp $te(\hat{\beta}_t^{MTL}) < te(\hat{\beta}_t^{STL})$.*
- *If $\Delta_{var} - \Delta_\beta \leqslant \delta$, then whp $te(\hat{\beta}_t^{MTL}) > te(\hat{\beta}_t^{STL})$.*

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.

4.2 Many Tasks with the Same Covariates

We extend the above result to any number of tasks that have the same covariates. Since the tasks all have the same number of datapoints and covariance matrix, the trade-off between model shift bias and variance will be captured by their task models $\{\beta_i\}_{i=1}^k$. Let $B^\star = [\beta_1, \beta_2, \ldots, \beta_k] \in \mathbb{R}^{p \times k}$ denote the underlying task model parameters. We derive the model shift bias and variance in the following result.

Theorem 4.2. *Let $n = c \cdot p$. Let $X \in \mathbb{R}^{n \times p}$ and $Y_i = X\beta_i + \varepsilon_i$, for $i = 1, \ldots, k$. Let $U_r U_r^\top$ denote the best rank-r approximation subspace of $B^\star \Sigma B^\star$, where $U_r \in \mathbb{R}^{k \times r}$. Let $U_r(i)$ denote the i-th row vector of U_r. We have the following*

- *If $(1 - \|U_r(i)\|^2) \cdot \frac{\sigma^2}{c-1} \geqslant \|\Sigma(B^\star U_r U_r(i) - \beta_i)\|^2$, then whp $te(\hat{\beta}_t^{MTL}) < te(\hat{\beta}_t^{STL})$.*
- *If $(1 - \|U_r(i)\|^2) \cdot \frac{\sigma^2}{c-1} < \|\Sigma(B^\star U_r U_r(i) - \beta_i)\|^2$, then whp $te(\hat{\beta}_t^{MTL}) > te(\hat{\beta}_t^{STL})$.*

As a remark, since the spectral norm of U_r is less than 1, we have that $\|U_r(i)\| < 1$, for any $1 \leqslant i \leqslant k$. Compared to Theorem 4.1, we can get a simple expression for the two functions Δ_{vari} and Δ_β. The proof of Theorem 4.2 can be found in Appendix C.2.

5 Experiments

A metric to determine when MTL performs better STL.

When should we align task covariances in MTL.

6 Related Work

Adding a regularization over B, e.g. [25, 26]. Moreover, [21] observed that controlling the capacity can outperform the implicit capacity control of adding regularization over B.

References

[1] Johannes Alt. Singularities of the density of states of random Gram matrices. *Electron. Commun. Probab.*, 22:13 pp., 2017.

[2] Johannes Alt, L. Erdős, and Torben Krüger. Local law for random Gram matrices. *Electron. J. Probab.*, 22:41 pp., 2017.

[3] Z. D. Bai and Jack W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.*, 26(1):316–345, 1998.

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[5] Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003.

[6] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.*, 19(33):1–53, 2014.

[7] A. Bloemendal, A. Knowles, H.-T. Yau, and J. Yin. On the principal components of sample covariance matrices. *Prob. Theor. Rel. Fields*, 164(1):459–552, 2016.

[8] P. Bourgade, H.-T. Yau, and J. Yin. Local circular law for random matrices. *Probab. Theory Relat. Fields*, 159:545–595, 2014.

[9] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pages 794–803, 2018.

[10] Xiucai Ding and Fan Yang. A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *Ann. Appl. Probab.*, 28(3):1679–1738, 2018.

[11] L. Erdős, A. Knowles, and H.-T. Yau. Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré*, 14:1837–1926, 2013.

[12] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Delocalization and diffusion profile for random band matrices. *Commun. Math. Phys.*, 323:367–416, 2013.

[13] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. The local semicircle law for a general class of random matrices. *Electron. J. Probab.*, 18:1–58, 2013.

[14] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Spectral statistics of Erdős-Rényi graphs I: Local semicircle law. *Ann. Probab.*, 41(3B):2279–2375, 2013.

[15] Sabri Eyuboglu, Geoffrey Angus, Bhavik N. Patel, Anuj Pareek, Guido Davidzon, JaredDunnmon, and Matthew P. Lungren. Multi-task weak supervision enables automatedabnormality localization in whole-body fdg-pet/ct. 2020.

[16] Viacheslav Leonidovich Girko. *Theory of random determinants*, volume 45. Springer Science & Business Media, 2012.

[17] VL Girko. Random matrices. *Handbook of Algebra, ed. Hazewinkel*, 1:27–78, 1975.

[18] Vyacheslav L Girko. Spectral theory of random matrices. *Russian Mathematical Surveys*, 40(1):77, 1985.

[19] Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, pages 1–96, 2016.

[20] Wouter M Kouw. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.

[21] Abhishek Kumar and Hal Daumé III. Learning task grouping and overlap in multi-task learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012.

[22] Yunsheng Li and Nuno Vasconcelos. Efficient multi-domain learning by covariance normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5424–5433, 2019.

[23] Tongliang Liu, Dacheng Tao, Mingli Song, and Stephen J Maybank. Algorithm-dependent generalization bounds for multi-task learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):227–241, 2016.

[24] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019.

[25] Karim Lounici, Massimiliano Pontil, Alexandre B Tsybakov, and Sara Van De Geer. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, 2009.

[26] Karim Lounici, Massimiliano Pontil, Sara Van De Geer, Alexandre B Tsybakov, et al. Oracle inequalities and optimal inference under group sparsity. *The annals of statistics*, 39(4):2164–2204, 2011.

[27] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[28] Natesh S. Pillai and Jun Yin. Universality of covariance matrices. *Ann. Appl. Probab.*, 24:935–1001, 2014.

[29] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[30] Sen Wu, Hongyang R. Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*, 2020.

[31] Haokai Xi, Fan Yang, and Jun Yin. Local circular law for the product of a deterministic matrix with a random matrix. *Electron. J. Probab.*, 22:77 pp., 2017.

[32] Fan Yang. Edge universality of separable covariance matrices. *Electron. J. Probab.*, 24:57 pp., 2019.

[33] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020.

[34] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018.

A Extension to Transfer Learning

We study the transfer function of taskonomy [34]. The algorithm is as follows. First, we obtain the single-task estimator $\hat{\beta}_i$ from every task, for $1 \leqslant i \leqslant k$. This forms the shared representation B in Algorithm **??**. Then, we learn the output layer on the target task. We use our tools to analyze this setting as follows.

B Supplementary Materials for Section 3

From [30], 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

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

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 (B.1) [21].

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

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

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

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

The intuition for deriving $\hat{\beta}_t^{\text{MTL}}$ 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 $\hat{\beta}_t^{\text{MTL}}$ for the target task is

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

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

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

We state several helper lemmas to get a bound on the variance of $\hat{\beta}_t^{\text{STL}}$ and $\hat{\beta}_t^{\text{MTL}}$. The first lemma, which is a folklore result in random matrix theory, helps determine the asymptotic limit of $te(\hat{\beta}_t^{\text{STL}})$, as p goes to infinity.

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

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

The second lemma, which deals the inverse of the sum of two random matrices, can be viewed as a special case of Theorem C.1.

Lemma B.2. *Let $X_i \in \mathbb{R}^{n_i\times p}$ be a random matrix that contains i.i.d. row vectors with mean 0 and variance $\Sigma_i \in \mathbb{R}^{p\times p}$, for $i = 1, 2$. Denote by $M = \Sigma_1^{1/2}\Sigma_2^{-1/2}$ and let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the singular values of $M^\top M$ in decreasing order. When $n_1 = c_1 p$ and $n_2 = c_2 p$, we have that with high probability over the randomness of X_1 and X_2, the following equation holds*

$$\operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_2\right] = \frac{1}{c_1+c_2}\cdot \operatorname{Tr}\left[\frac{1}{p}\cdot(a_1 M^\top M + a_2)^{-1}\right] + \mathrm{O}\left(n^{-1/2+\varepsilon}\right), \quad \text{(B.4)}$$

for any constant $\varepsilon > 0$, where a_1, a_2 are solutions to the following deterministic equations:

$$a_1 + a_2 = 1 - \frac{1}{c_1+c_2},\ a_1 + \frac{1}{p}\cdot\sum_{i=1}^{p}\frac{a_1}{(c_1+c_2)(a_1+a_2/\lambda_i^2)} = \frac{c_1}{c_1+c_2}. \quad \text{(B.5)}$$

We will give the proof of Lemma B.2 in Section D.

Finally, we can get a bound tighter than Theorem C.1 as follows. [Todo: restate this]

Lemma B.3. *In the setting of Theorem C.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 $\Delta_\beta = \left((1-\hat{w})^2\kappa^2 + d^2\right)\operatorname{Tr}[Z]$ and we have*

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

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

B.1 Proof of Proposition 3.2 and Proposition 3.3

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 $te(\hat{\beta}_t^{\text{MTL}})$ to show the result.

We denote

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

Under the setting of Proposition B.3, using concentration of random vector with i.i.d. Gaussian entries, Lemma E.6, we have that

$$val(\hat{B}; w_1, w_2) = val(w)\left(1 + \mathrm{O}(p^{-1/2})\right) \quad \text{with probability } 1 - \mathrm{o}(1).$$

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 **??**.

Lemma B.4. *Suppose the assumptions of Proposition B.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*

$$|\hat{w} - 1| = \mathrm{O}(p^{-1}). \tag{B.6}$$

Proof. The proof is also similar to the one for Proposition **??**. 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

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

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

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

where the matrix $\mathcal{A}$ is

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

Using the estimate (C.8), we get that $\mathcal{A}$ is lower bounded as

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

as long as

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

Hence $val'(w) > 0$ on (w_1, ∞), i.e. $val(w)$ is strictly increasing for $w > w_1$. Hence we must have $\hat{w} \leqslant w_1$. 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

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

where the matrix $\mathcal{B}$ is

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

Using the estimate (C.8), we get that $\mathcal{B}$ is upper bounded as

$$\mathcal{B} \preceq -2(1-w)\kappa^2(\sqrt{n_2}-\sqrt{p})^4 + 2d^2 w\frac{n_2}{n_1}(\sqrt{n_1}+\sqrt{p})^4 \prec 0,$$

as long as

$$w < w_2 := 1 - \frac{d^2}{\kappa^2}\frac{n_2(\sqrt{n_1}+\sqrt{p})^4}{n_1(\sqrt{n_2}-\sqrt{p})^4}.$$

Hence $val'(w) < 0$ on $[0, w_2)$, i.e. $val(w)$ is strictly increasing for $w < w_2$. Hence we must have $\hat{w} \leqslant w_2$. Note that under our assumptions, we have $w_2 = 1 - \mathrm{O}(p^{-1})$. □

Based on Lemma B.4, we can prove Proposition 3.2.

Proof of Proposition 3.2. Since $\Sigma_1 = \Sigma_2$, we know that $M = \Sigma_1^{1/2}\Sigma_2^{-1/2} = \hat{v}\,\mathrm{Id}$. Using Lemma B.1 and B.2, we can track the reduction of variance from $\hat{\beta}_t^{\text{MTL}}$ to $\hat{\beta}_t^{\text{STL}}$ as whp

$$\begin{aligned}\Delta_{\text{var}} &:= \sigma^2\left(\frac{1}{c_2-1} - \frac{1}{c_1+c_2}\cdot\frac{1}{a_1\hat{v}^2+a_2} + \mathrm{O}\left(p^{-1/2+\varepsilon}\right)\right)\\ &= \sigma^2\left(\frac{c_1}{(c_2-1)(c_1+c_2-1)} + \mathrm{O}\left(p^{-1/2+\varepsilon}\right)\right),\end{aligned} \tag{B.7}$$

where we use equation (B.5) and Lemma B.4. Next we consider the model shift bias

$$\begin{aligned}\Delta_\beta &:= \hat{v}^2\left\|(\hat{v}^2X_1^\top X_1 + X_2^\top X_2)^{-1}X_1^\top X_1(\beta_s - \hat{v}\beta_t)\right\|^2\\ &= d^2\cdot\left\|(\hat{v}^2X_1^\top X_1 + X_2^\top X_2)^{-1}X_1^\top X_1\right\|_F^2 + \mathrm{O}\left(p^{-1/2+\varepsilon}d^2\right)\end{aligned}$$

Using Lemma B.3, we get an upper and lower bound on Δ_β as

$$\left(1-\sqrt{\frac{1}{c_1}}\right)^4 \leqslant \Delta_\beta/\left(p\cdot d^2\cdot\frac{c_1^2}{(c_1+c_2)^2}\cdot\frac{1+a_3+a_4}{(a_1+a_2)^2} + \mathrm{O}\left(p^{1/2+\varepsilon}d^2\right)\right) \leqslant \left(1+\sqrt{\frac{1}{c_1}}\right)^4. \tag{B.8}$$

By solving equations (B.5), (C.2) and (C.3), we get

$$\begin{aligned}a_1 &= \frac{c_1(c_1+c_2-1)}{(c_1+c_2)^2} + O(p^{-1/2+\varepsilon}), a_2 = \frac{c_2(c_1+c_2-1)}{(c_1+c_2)^2} + O(p^{-1/2+\varepsilon}),\\ a_3 &= \frac{c_2}{(c_1+c_2)(c_1+c_2-1)} + O(p^{-1/2+\varepsilon}), a_4 = \frac{c_1}{(c_1+c_2)(c_1+c_2-1)} + O(p^{-1/2+\varepsilon}).\end{aligned}$$

Hence we obtain that

$$\frac{1+a_3+a_4}{(a_1+a_2)^2} = \frac{(c_1+c_2)^3}{(c_1+c_2-1)^3} + \mathrm{O}\left(p^{-1/2+\varepsilon}\right).$$

To sum up, we have shown that when

$$\frac{p\cdot d^2}{\sigma^2} < \frac{(c_1+c_2-1)^2}{c_1(c_1+c_2)(c_2-1)}\cdot\left(1+\sqrt{\frac{1}{c_1}}\right)^{-4},$$

we have that $\Delta_{\text{var}} > \Delta_\beta$, which implies that $te(\hat{\beta}_t^{\text{MTL}})$ is lower than $te(\hat{\beta}_t^{\text{STL}})$. When

$$\frac{p\cdot d^2}{\sigma^2} > \frac{(c_1+c_2-1)^2}{c_1(c_1+c_2)(c_2-1)}\cdot\left(1-\sqrt{\frac{1}{c_1}}\right)^{-4},$$

we have that $\Delta_{\text{var}} < \Delta_\beta$, which implies that $te(\hat{\beta}_t^{\text{MTL}})$ is higher than $te(\hat{\beta}_t^{\text{STL}})$. The result follows by taking $c_1 = c_2 = c$ for the above equation. □

Proof of Proposition 3.3. Under the setting of Proposition 3.3, we have

$$\begin{aligned}te(\hat{\beta}_t^{\text{MTL}}) = {}& \hat{v}^2\left\|(\hat{v}^2X_1^\top X_1 + X_2^\top X_2)^{-1}X_1^\top X_1(\beta_s - \hat{v}\beta_t)\right\|^2\\ &+ \sigma_2^2\cdot\mathrm{Tr}\left[(\hat{v}^2X_1^\top X_1 + X_2^\top X_2)^{-1}\right] + (\sigma_1^2-\sigma_2^2)\hat{v}^2\cdot\mathrm{Tr}\left[(\hat{v}^2X_1^\top X_1 + X_2^\top X_2)^{-2}X_1^\top X_1\right].\end{aligned}$$

Using concentration of random vector with i.i.d. Gaussian entries, Lemma E.6, we obtain that

$$\begin{aligned}te(\hat{\beta}_t^{\text{MTL}}) = {}& \hat{v}^2\left[d^2 + (w-1)^2\kappa^2\right]\mathrm{Tr}\left[(\hat{v}^2X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right]\cdot\left(1+\mathrm{O}(p^{-1/2})\right)\\ &+ (\sigma_1^2-\sigma_2^2)\hat{v}^2\cdot\mathrm{Tr}\left[(\hat{v}^2X_1^\top X_1 + X_2^\top X_2)^{-2}X_1^\top X_1\right] + \sigma_2^2\cdot\mathrm{Tr}\left[(\hat{v}^2X_1^\top X_1 + X_2^\top X_2)^{-1}\right].\end{aligned}$$

In the setting of Proposition 3.3, we can also obtain that

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

as in (B.6). We omit the details of the proof, since it is almost the same as the one in the proof of Lemma B.4. Thus $te(\hat{\beta}_t^{\text{MTL}})$ can be reduced to

$$te(\hat{\beta}_t^{\text{MTL}}) = \left[\Delta_\beta + \Delta_{diff} + \sigma_2^2 \cdot \text{Tr}(X_1^\top X_1 + X_2^\top X_2)^{-1}\right] \cdot \left(1 + \text{O}(p^{-1/2})\right),$$

where $\Delta_\beta := \sigma_2^2 \cdot \text{Tr}\left[(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\right]$ has been estimated in (B.8) and Δ_{diff} is

$$\Delta_{diff} := (\sigma_1^2 - \sigma_2^2) \cdot \text{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2} X_1^\top X_1\right].$$

Using (C.8) and Lemma C.3, a similar bound holds for Δ_{diff} as in (B.8):

$$\left(1 - \sqrt{\frac{1}{c_1}}\right)^2 \leqslant \frac{\Delta_{diff}}{\sigma_1^2 - \sigma_2^2} \cdot \left(\frac{c_1(c_1+c_2)}{(c_1+c_2-1)^3} + \text{O}(p^{-1/2+\varepsilon})\right)^{-1} \leqslant \left(1 + \sqrt{\frac{1}{c_1}}\right)^2.$$

Next we compare Δ_{Var} with $\Delta_\beta + \Delta_{diff}$. Under the condition $d^2 < \frac{2\sigma^2}{3p} \frac{(c_1+c_2-1)^2}{c_1(c_1+c_2)(c_2-1)}$, if $\sigma_1^2 \leqslant \sigma_2^2$ then we always have that $\Delta_{\text{var}} > \Delta_\beta$ and hence $te(\hat{\beta}_t^{\text{MTL}})$ is smaller than $te(\hat{\beta}_t^{\text{STL}})$. It remains to consider the case $\sigma_1^2 \geqslant \sigma_2^2$. If

$$\begin{aligned} &\sigma_2^2 \frac{c_1}{(c_2-1)(c_1+c_2-1)} \left(1 + \text{O}\left(p^{-1/2+\varepsilon}\right)\right) \\ &> \left[\frac{c_1^2(c_1+c_2)}{(c_1+c_2-1)^3} + \text{O}\left(p^{-1/2+\varepsilon}\right)\right] \left[pd^2 \left(1 + \sqrt{\frac{1}{c_1}}\right)^4 + \frac{\sigma_1^2 - \sigma_2^2}{c_1} \left(1 + \sqrt{\frac{1}{c_1}}\right)^2\right], \end{aligned} \tag{B.9}$$

then we have $\Delta_{\text{var}} > \Delta_\beta$ and we have $te(\hat{\beta}_t^{\text{MTL}})$ is smaller than $te(\hat{\beta}_t^{\text{STL}})$. For $c_1 \geqslant 100$, we see that (B.9) is implied by the following condition

$$\sigma_1^2 \leqslant pd^2 \cdot c_1 + \left[1 + \frac{2}{3} \frac{(c_1+c_2-1)^2}{(c_1+c_2)(c_2-1)}\right] \sigma_2^2$$

as long as p is large enough. On the other hand, if

$$\begin{aligned} &\sigma_2^2 \frac{c_1}{(c_2-1)(c_1+c_2-1)} \left(1 + \text{O}\left(p^{-1/2+\varepsilon}\right)\right) \\ &< \left[\frac{c_1^2(c_1+c_2)}{(c_1+c_2-1)^3} + \text{O}\left(p^{-1/2+\varepsilon}\right)\right] \left[pd^2 \left(1 - \sqrt{\frac{1}{c_1}}\right)^4 + \frac{\sigma_1^2 - \sigma_2^2}{c_1} \left(1 - \sqrt{\frac{1}{c_1}}\right)^2\right], \end{aligned} \tag{B.10}$$

then we have $\Delta_{\text{var}} < \Delta_\beta$ and we have $te(\hat{\beta}_t^{\text{MTL}})$ is larger than $te(\hat{\beta}_t^{\text{STL}})$. For $c_1 \geqslant 100$, we see that (B.10) is implied by the following condition

$$\sigma_1^2 \geqslant pd^2 \cdot c_1 + \left[1 + \frac{3}{2} \frac{(c_1+c_2-1)^2}{(c_1+c_2)(c_2-1)}\right] \sigma_2^2.$$

□

B.2 Proof of Proposition 3.4 and Proposition 3.5

Proof of Proposition 3.4. Recall that under the setting of Example 3.1, (B.7) and (B.8) hold. Then we get that $te(\hat{\beta}_t^{\text{MTL}}) < te(\hat{\beta}_t^{\text{STL}})$ whp if

$$\left(pd^2 \frac{c_1(c_1+c_2)(c_2-1)}{(c_1+c_2-1)^2} + \text{O}\left(p^{1/2+\varepsilon} d^2\right)\right) \left(1 + \sqrt{\frac{1}{c_1}}\right)^4 < \sigma^2 \left(1 + \text{O}\left(p^{-1/2+\varepsilon}\right)\right); \tag{B.11}$$

otherwise if

$$\left(pd^2 \frac{c_1(c_1+c_2)(c_2-1)}{(c_1+c_2-1)^2} + \text{O}\left(p^{1/2+\varepsilon} d^2\right)\right) \left(1 - \sqrt{\frac{1}{c_1}}\right)^4 > \sigma^2 \left(1 + \text{O}\left(p^{-1/2+\varepsilon}\right)\right), \tag{B.12}$$

then we have $te(\hat{\beta}_t^{\text{MTL}}) > te(\hat{\beta}_t^{\text{STL}})$ whp.

Now we prove the first statement of Proposition 3.4. Notice that the function

$$\frac{c_1(c_1+c_2)(c_2-1)}{(c_1+c_2-1)^2} = (c_2-1)\left(1+\frac{c_2-2}{c_1}+\frac{1}{c_1(c_1+c_2)}\right)^{-1}$$

is strictly increasing with respect to c_1 as long as $c_2 \geqslant 3$. In particular, by taking $c_1 \to \infty$, we get the bound:

$$\frac{c_1(c_1+c_2)(c_2-1)}{(c_1+c_2-1)^2}\cdot\left(1+\sqrt{\frac{1}{c_1}}\right)^4 < (c_2-1)\left(1+\sqrt{\frac{1}{c_1}}\right)^4.$$

Hence we conclude that $te(\hat{\beta}_t^{\mathrm{MTL}}) < te(\hat{\beta}_t^{\mathrm{STL}})$ whp as long as

$$pd^2+\mathrm{o}(1) \leqslant \left(1+\sqrt{\frac{1}{c_1}}\right)^{-4}\frac{\sigma^2}{c_2-1}.$$

This gives the first statement by taking $c_1 > a$.

The second statement can be proved in a similar way. Suppose $c_1 > a$ and $pd^2 > (1-a^{-1/2})^{-4}\frac{\sigma^2}{c_2-1}$. If $c_1 > \frac{(c_2-2)\sigma^2}{(1-a^{-1/2})^4(1-(a+c_2-2)^{-2})(c_2-1)pd^2-\sigma^2}$, we have

$$\begin{aligned}
&pd^2\cdot\frac{c_1(c_1+c_2)(c_2-1)}{(c_1+c_2-1)^2}\cdot\left(1-\sqrt{\frac{1}{c_1}}\right)^4\\
&>\left(1-\sqrt{\frac{1}{c_1}}\right)^4\left(1-\frac{1}{(c_1+c_2-2)^2}\right)\cdot\frac{pd^2(c_2-1)c_1}{c_1+(c_2-2)} > \sigma^2\cdot(1+\mathrm{o}(1)),
\end{aligned}$$

where in the first step we used that

$$\frac{(c_1+c_2)(c_1+c_2-2)}{(c_1+c_2-1)^2} > 1-\frac{1}{(c_1+c_2-2)^2}.$$

This shows that (B.12) holds as long as p is large enough, and hence $te(\hat{\beta}_t^{\mathrm{MTL}}) > te(\hat{\beta}_t^{\mathrm{STL}})$ holds. On the other hand, if $c_1 < \frac{(c_2-2)\sigma^2}{(1+a^{-1/2})^4(c_2-1)pd^2-\sigma^2}$, then we have

$$pd^2\cdot\frac{c_1(c_1+c_2)(c_2-1)}{(c_1+c_2-1)^2}\cdot\left(1+\sqrt{\frac{1}{c_1}}\right)^4 < \left(1+\sqrt{\frac{1}{c_1}}\right)^4\cdot\frac{pd^2(c_2-1)c_1}{c_1+(c_2-2)} < \sigma^2\cdot(1-\mathrm{o}(1)).$$

This shows that (B.11) holds as long as p is large enough, and hence $te(\hat{\beta}_t^{\mathrm{MTL}}) < te(\hat{\beta}_t^{\mathrm{STL}})$ holds. □

Remark. Furthermore, as a function of c_1 over the range $[??,\infty]$, the maximum of $te(\hat{\beta}_t^{\mathrm{STL}}) - te(\hat{\beta}_t^{\mathrm{MTL}})$ is attained when $c_1 = c_2\sigma^2/\max(2(c_2-1)pd^2-\sigma^2,0)$. (cannot get this because we only have some bounds. If we let $c_1\to\infty$, then the curve of $te(\hat{\beta}_t^{\mathrm{STL}}) - te(\hat{\beta}_t^{\mathrm{MTL}})$ already becomes flat, and it is meaningless to discuss the minimum of this function at this point?)

Proof of Proposition 3.5. Suppose we have reduced number of datapoints—αn_1 for task 1 and αn_2 for task 2 with $n_1 = n_2$. In the setting of Proposition 3.5, we still have (B.6), and by symmetry it suffices to focus on one of tasks, say task 2. Using Lemmas B.2 and B.3, we have that whp,

$$te_2(\hat{\beta}(\alpha)) = \sigma^2\left(\frac{1}{2\alpha c-1}+\mathrm{O}\left(p^{-1/2+\varepsilon}\right)\right)+\Delta_\beta^{(2)}, \tag{B.13}$$

where $\Delta_\beta^{(2)}$ satisfies

$$\left(1-\sqrt{\frac{1}{\alpha c}}\right)^4 \leqslant \Delta_\beta^{(2)}\Big/\left(pd^2\cdot\frac{2\alpha^3c^3}{(2\alpha c-1)^3}+\mathrm{O}\left(p^{1/2+\varepsilon}d^2\right)\right) \leqslant \left(1+\sqrt{\frac{1}{\alpha c}}\right)^4.$$

On the other hand, using Lemma B.1, we have whp

$$te_2(\hat{\beta}_t^{\mathrm{STL}}) = \frac{\sigma^2}{c-1}\left(1+\mathrm{O}\left(p^{-1/2+\varepsilon}\right)\right). \tag{B.14}$$

Comparing (B.13) and (B.14), we observe that $te_2(\hat{\beta}(\alpha)) \geqslant te_2(\hat{\beta}_t^{\mathrm{STL}})$ for $\alpha \leqslant 1/2 - \mathrm{o}(1)$, which gives $\alpha^* \geqslant 1/2 - \mathrm{o}(1)$.

Now for $c \geqslant 200$ and $\alpha \geqslant 1/2$, we have whp

$$te_2(\hat{\beta}(\alpha)) < \sigma^2 \frac{1}{2\alpha c - 1} + \frac{3}{8} p d^2 \tag{B.15}$$

for large enough p. Moreover, if

$$\alpha \geqslant \alpha_0 := \frac{1}{2c} + \left(2 - \frac{3}{4} c \cdot \frac{p d^2}{\sigma^2}\right)^{-1},$$

we can check from (B.15) that $te_2(\hat{\beta}(\alpha)) < te_2(\hat{\beta}_t^{\mathrm{STL}})$. By symmetry, we also have $te_1(\hat{\beta}(\alpha)) < te_1(\hat{\beta}_t^{\mathrm{STL}})$. Thus we conclude that

$$te_1(\hat{\beta}(\alpha)) + te_2(\hat{\beta}(\alpha)) \leqslant te_1(\hat{\beta}_t^{\mathrm{STL}}) + te_2(\hat{\beta}_t^{\mathrm{STL}})$$

for $\alpha \geqslant \alpha_0$, which shows that $\alpha^* \leqslant \alpha_0$. □

B.3 Proofs of Proposition 3.7 and Theoretical justification of Example 3.8

Proof of Proposition 3.7. Let

$$M_0 := \operatorname*{arg\,min}_{M \in \mathcal{S}_\mu} (te(\hat{\beta}_t^{\mathrm{MTL}}))(M).$$

We now calculate $(te(\hat{\beta}_t^{\mathrm{MTL}}))(M_0)$. In this case, as in Lemma B.4 we also have that

$$|\hat{v} - 1| = \mathrm{O}(p^{-1}) \tag{B.16}$$

in the setting Proposition 3.7. In fact, Lemma B.4 was proved assuming that $M = \mathrm{Id}$, but its proof can be easily extended to the case with general $M \in \mathcal{S}_\mu$ by using that $\lambda(M) \in [\mu_{\min}, \mu_{\max}]$. We omit the details here.

Now using (B.16), Lemma B.2 and Lemma B.3, we get that whp,

$$(te(\hat{\beta}_t^{\mathrm{MTL}}))(M_0) = \frac{\sigma^2}{c_1 + c_2} \cdot \frac{1}{p} \operatorname{Tr}\left(\frac{1}{a_1 M_0^\top M_0 + a_2}\right) + \Delta_\beta(M_0),$$

where $\Delta_\beta(M_0)$ satisfies

$$\begin{aligned}
&\left|\Delta_\beta(M_0) - \frac{d^2 \cdot c_1^2}{(c_1 + c_2)^2} \operatorname{Tr}\left[M_0 \frac{(1 + a_3)\,\mathrm{Id} + a_4 {M_0}^\top M_0}{(a_2 + a_1 {M_0}^\top M_0)^2} {M_0}^\top\right]\right| \\
&\leqslant \left(\left(1 + \sqrt{\frac{1}{c_1}}\right)^4 - 1\right)\left(\frac{c_1 \mu_{\max}}{(\sqrt{c_1} - 1)^2 \mu_{\min} + (\sqrt{c_2} - 1)^2}\right)^2 \cdot d^2 \operatorname{Tr}(\Sigma_1).
\end{aligned}$$

From equation (C.1), we get

$$a_1 \geqslant \frac{c_1 - 1}{c_1 + c_2}, \quad a_2 \leqslant \frac{c_2}{c_1 + c_2}.$$

Then solving equations (C.2) and (C.3), we get that

$$\begin{aligned}
0 \leqslant a_3 &= \frac{\frac{1}{n_2}\sum_{i=1}^p \frac{a_2^2}{(a_2 + \lambda_i^2 a_1)^2}\left(1 - \frac{1}{n_1}\sum_{i=1}^p \frac{\lambda_i^4 a_1^2}{(a_2 + \lambda_i^2 a_1)^2}\right) + \left(\frac{1}{n_1}\sum_{i=1}^p \frac{\lambda_i^2 a_1^2}{(a_2 + \lambda_i^2 a_1)^2}\right)\left(\frac{1}{n_2}\sum_{i=1}^p \frac{\lambda_i^2 a_2^2}{(a_2 + \lambda_i^2 a_1)^2}\right)}{\left(1 - \frac{1}{n_2}\sum_{i=1}^p \frac{a_2^2}{(a_2 + \lambda_i^2 a_1)^2}\right)\left(1 - \frac{1}{n_1}\sum_{i=1}^p \frac{\lambda_i^4 a_1^2}{(a_2 + \lambda_i^2 a_1)^2}\right) - \left(\frac{1}{n_2}\sum_{i=1}^p \frac{\lambda_i^2 a_2^2}{(a_2 + \lambda_i^2 a_1)^2}\right)\left(\frac{1}{n_1}\sum_{i=1}^p \frac{\lambda_i^2 a_1^2}{(a_2 + \lambda_i^2 a_1)^2}\right)} \\
&\leqslant \frac{c_2^{-1}}{1 - c_1^{-1} - c_2^{-1}}
\end{aligned}$$

and

$$\begin{aligned}
0 \leqslant a_4 &= \frac{\frac{1}{n_1}\sum_{i=1}^p \frac{\lambda_i^2 a_1^2}{(a_2 + \lambda_i^2 a_1)^2}}{\left(1 - \frac{1}{n_2}\sum_{i=1}^p \frac{a_2^2}{(a_2 + \lambda_i^2 a_1)^2}\right)\left(1 - \frac{1}{n_1}\sum_{i=1}^p \frac{\lambda_i^4 a_1^2}{(a_2 + \lambda_i^2 a_1)^2}\right) - \left(\frac{1}{n_2}\sum_{i=1}^p \frac{\lambda_i^2 a_2^2}{(a_2 + \lambda_i^2 a_1)^2}\right)\left(\frac{1}{n_1}\sum_{i=1}^p \frac{\lambda_i^2 a_1^2}{(a_2 + \lambda_i^2 a_1)^2}\right)} \\
&\leqslant \frac{c_1^{-1} \cdot \mu_{\min}^{-2}}{1 - c_1^{-1} - c_2^{-1}},
\end{aligned}$$

where we also used that

$$\left(\sum_{i=1}^{p} \frac{\lambda_i^2}{(a_2+\lambda_i^2 a_1)^2}\right)^2 \leqslant \sum_{i=1}^{p} \frac{\lambda_i^4}{(a_2+\lambda_i^2 a_1)^2} \cdot \sum_{i=1}^{p} \frac{1}{(a_2+\lambda_i^2 a_1)^2}$$

by Cauchy-Schwarz inequality.

Combining the above estimates, we get that

$$(te(\hat{\beta}_t^{\mathrm{MTL}}))(M_0) = te(M_0) + \mathcal{E}, \tag{B.17}$$

where

$$te(M_0) := \frac{\sigma^2}{c_1+c_2} \cdot \frac{1}{p} \operatorname{Tr}\left(\frac{1}{a_1 M_0^\top M_0}\right) + \frac{d^2 \cdot c_1^2}{(c_1+c_2)^2} \operatorname{Tr}\left(\frac{1+a_3}{a_1^2 M_0^\top M_0}\right),$$

and the error satisfies

$$|\mathcal{E}| \leqslant C\left(\frac{c_2}{c_1} + c_1^{-1/2}\right) te(M_0),$$

Here the constant $C > 0$ depends only on $\mu_{\max}$, $\mu_{\min}$ and $\|\Sigma_1\|$, but otherwise does not depend on c_1 and c_2.

Finally using AM-GM inequality, we observe that

$$\operatorname{Tr}\left(\frac{1}{M^\top M}\right) = \sum_{i=1}^{p} \frac{1}{\lambda_i}$$

is minimized when $\lambda_1 = \cdots \lambda_p = \mu$ under the restriction $\prod_{i=1}^{p} \lambda_i \leqslant \mu^p$. Hence we get that

$$te(M_0) \leqslant te(\mu \operatorname{Id}).$$

Together with (B.17), we conclude (3.1). □

Theoretical justification of Example 3.8. In the setting of Example 3.8, equations in (C.1) become

$$a_1 + a_2 = 1 - \frac{p}{n_1+n_2}, \;\; a_1 + \frac{p}{2(n_1+n_2)} \cdot \left(\frac{a_1}{a_1+\lambda^2 a_2} + \frac{a_1}{a_1+\frac{a_2}{\lambda^2}}\right) = \frac{n_1}{n_1+n_2}. \tag{B.18}$$

It's not hard to verify that there is only one valid solution (a_1, a_2) to (B.18). After solving these, we get the test error for the target task as follows.

$$te(\lambda) = \frac{p}{2(n_1+n_2)} \cdot \left(\frac{1}{\frac{a_1}{\lambda^2}+a_2} + \frac{1}{a_1\lambda^2+a_2}\right). \tag{B.19}$$

First we notice that the curves in Figure **??** 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 (B.18), where we denote $\gamma = p/(n_1+n_2)$. Then for any λ, the test error in (B.19) takes the value

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

This phenomenon can be also explained using our theory. With (B.18), 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 (B.18)), and hence $te(\lambda) > te(1)$. Otherwise if $n_1 < n_2$, we have $a_1 < (1-\gamma)/2$, and hence $te(\lambda) < te(1)$. □

C Supplementary Materials for Section 4

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

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

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

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

$$\left(\frac{n_1}{a_1^2} - \sum_{i=1}^{p} \frac{\lambda_i^4}{(a_2 + \lambda_i^2 a_1)^2}\right) a_4 - \left(\sum_{i=1}^{p} \frac{\lambda_i^2}{(a_2 + \lambda_i^2 a_1)^2}\right) a_3 = \sum_{i=1}^{p} \frac{\lambda_i^2}{(a_2 + \lambda_i^2 a_1)^2}. \tag{C.3}$$

Then we introduce the following matrix

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

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 C.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* (B.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*

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

$$te(\hat{\beta}_t^{MTL}) \geqslant te(\hat{\beta}_t^{STL}) \text{ when: } \Delta_{var} - \Delta_\beta \leqslant -2\left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1}\right)\delta, \tag{C.5}$$

where

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

C.1 Proof of Theorem C.1

[Todo: A proof outline; including the following key lemma.] To prove Theorem **??**, 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

$$\mathbb{E} z_{ij} = 0, \qquad \mathbb{E}|z_{ij}|^2 = 1. \tag{C.6}$$

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

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

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 (C.8). 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

$$\delta_1 = \delta_2 + 2\hat{w}^2 n_1 (\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t) + \hat{w}^2 \left\| \Sigma_2^{1/2} K \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} (\beta_s - \hat{w}\beta_t) \right\|^2 \tag{C.7}$$

Here we use the following on the second term in equation (C.7)

$$\begin{aligned}
&\left| (\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t) \right| \\
=&\left| \mathrm{Tr} \left[\mathcal{E} \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t)(\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} \right] \right| \\
\leqslant&\|\mathcal{E}\| \cdot \left\| \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t)(\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} \right\|_\star \\
\leqslant&n_1 \left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1} \right) \cdot \left\| \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t)(\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} \right\|_\star \qquad \text{(by equation (C.8))} \\
\leqslant&n_1 \left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1} \right) \left\| \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1^{1/2} \right\| \cdot \left\| \Sigma_1^{1/2} (\beta_s - \hat{w}\beta_t) \right\|^2 \\
&\qquad \text{(since the matrix inside is rank 1)}
\end{aligned}$$

The third term in equation (C.7) 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 (C.7) 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 (C.8). □

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 (C.6). 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 (C.6), If we assume the variances of the entries of Z_1 are 1, then we have

$$-n_1 \left(2\sqrt{\frac{p}{n_1}} - \frac{p}{n_1} \right) \leqslant Z_1^T Z_1 - n_1 \,\mathrm{Id} \leqslant n_1 \left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1} \right). \tag{C.8}$$

Lemma C.3. *In the setting of Theorem C.1, we have with high probability* $1 - \mathrm{o}(1)$*,*

$$\begin{aligned}&\widehat{w}^2(n_1+n_2)^2 \left\|\Sigma_2^{1/2}(\hat{w}X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_1(\beta_s - w\beta_t)\right\|^2 \\ &= (\beta_s - w\beta_t)^\top \Sigma_1^{1/2} M \frac{(1+a_3)\,\mathrm{Id} + a_4 M^\top M}{(a_2 + a_1 M^\top M)^2} M^\top \Sigma_1^{1/2}(\beta_s - \hat{w}\beta_t) + \mathrm{O}(n^{-1/2+\varepsilon}),\end{aligned} \tag{C.9}$$

for any constant $\varepsilon > 0$*.*

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

Proof of Proposition B.3. the proof for tighter bound ..... □

add some arguments with ε-net.

C.2 Proof of Theorem 4.2

For this setting, the problem reduces to the following.

$$f(B; W_1, \ldots, W_k) = \sum_{i=1}^k \|XBW_i - Y_i\|^2 . \tag{C.10}$$

In order to prove Theorem 4.2, we will derive a closed form solution for equation (C.10).

Proof of Theorem 4.2. By fixing $W_1, W_2, \ldots, W_k$, we can derive a closed form solution for B as

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

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

Denote by $\varepsilon(W) = \sum_{i=1}^k \varepsilon_i W_i^\top$. We shall decompose the validation loss $val(\hat{B}; W_1, \ldots, W_k)$ 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

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

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

$$val(\hat{B}; W_1, \ldots, W_k) = \left\|\Sigma^{1/2}\left(B^\star Q - B^\star\right)\right\|_F^2 + \sigma^2 \cdot \mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right]$$

From the above we can solve for Q optimally as $U_r U_r^\top$. Furthermore, we can solve $\hat{\beta}_i^{\mathrm{MTL}}$ as $B^\star U_r U_r(i)$. Now we get that

$$\begin{aligned}te(\hat{\beta}_t^{\mathrm{MTL}}) &= \left\|\Sigma^{1/2}\left(\sum_{i=1}^k W_i^\top(ZZ^\top)^{-1}W_j\beta_i - \beta_j\right)\right\|^2 + \sigma^2 W_j^\top(ZZ^\top)^{-1}W_j \cdot \mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right] \\ &= \left\|\Sigma^{1/2}\left(B^\star U_r U_r(i)\right)\right\|^2 + \sigma^2\|U_r(i)\|^2 \cdot \mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right].\end{aligned}$$

By using Lemma B.1, we conclude the proof. □

D Proof of Lemma B.2 and Lemma C.3

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

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

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:

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

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

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

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 E.2 below.

We assume that Σ_1 and Σ_2 have eigendecompositions

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

where the eigenvalues satisfy that

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

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

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

where the singular values satisfy that

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

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

We summarize our basic assumptions here for future reference.

Assumption D.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 (D.1) and* (D.2)*, Σ_1 and Σ_2 are deterministic non-negative definite symmetric matrices satisfying* (D.4)-(D.7)*, and $d_{1,2}$ satisfy* (D.3)*.*

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

D.1 Notations

We will use the following notion of stochastic domination, which was first introduced in [11] and subsequently used in many works on random matrix theory, such as [6, 7, 8, 12, 13, 19]. 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 D.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 D.3 (Lemma 3.2 in [6])**.** *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 D.4 (Bounded support condition)**.** *We say a random matrix Z satisfies the bounded support condition with q, if*

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

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 (D.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 (D.6), we can rewrite it as

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

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

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

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 [16, 17, 18] and references therein. It has been proved to be useful in studying the local laws of random matrices of the Gram type [1, 2, 19, 31]. We define the following $(p + n) \times (p + n)$ self-adjoint block matrix, which is a linear function of X:

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

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

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

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 (D.10))

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

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):

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

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 (D.13), we can get that for $i, j \in \mathcal{I}_1$ and $\mu, \nu \in \mathcal{I}_2$,

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

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

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

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

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

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

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

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 (D.16) is reduced to

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

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 (D.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

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

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

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

Moreover, the solution satisfies

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

The proof is a standard application of the contraction principle. For reader's convenience, we will gives its proof in Appendix E.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 D.7 below.

Lemma D.6. *There exist constants $c_0, C_0 > 0$ depending only on τ in* (D.3)*,* (D.5)*,* (D.7) *and* (D.20) *such that the self-consistent equations in* (D.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

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

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

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

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

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

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

Theorem D.7. *Suppose Assumption D.1 holds, and Z_1, Z_2 satisfy the bounded support condition* (D.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*

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

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

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

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

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

Proof of Corollary D.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

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

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

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

Using (D.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

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

Moreover, we trivially have

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

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 D.7 with $q = n^{-c_a + \varepsilon}$, and (D.26) gives that

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

Then using (D.29) and (E.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 (D.26) also holds for $G(\widetilde{Z}_1, \widetilde{Z}_2, z)$ with $q = n^{-c_a + \varepsilon}$, and hence concludes the proof by (D.28). □

Using Corollary D.8, we can complete the proof of Lemma B.2 and Lemma C.3.

Proof of Lemma B.2. In the setting of Lemma B.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 D.1. Here the extra n^{-1} is due to the choice of the variances—in the setting of Lemma B.2 the variances of the entries of $Z_{1,2}$ are equal to 1, while in (D.1) they are taken to be n^{-1}. As in (D.6), we assume that $M := \widetilde{\Sigma}_1^{1/2} \Sigma_2^{-1/2}$ has singular value decomposition

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

Then as in (D.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 D.8, we obtain that for any small constant $\varepsilon > 0$, with probability $1 - \mathrm{o}(1)$,

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

where by (D.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 (D.19). Thus from (D.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 (**??**) if we rename $r_1 b_2 \to a_1$ and $r_2 b_3 \to a_2$. For (**??**), 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 (C.1). This gives (**??**) by (**??**). □

Proof of Lemma C.3. In the setting of Lemma C.3, 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 D.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 (D.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 D.8, we get that with probability $1 - \mathrm{o}(1)$,

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

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

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

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

By the implicit differentiation of (D.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

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

Then by (D.32) and (D.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 C.3. □

E Proof of Theorem D.7

The main difficulty for the proof of Theorem D.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

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

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. [6, 28, 32]. To go from the Gaussian case to the general X case, we will adopt a continuous self-consistent comparison argument developed in [19].

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

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

and prove the following result.

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

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

E.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 $1 - \mathrm{o}(1)$ [3]. 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 E.2. *Suppose Assumption D.1 holds, and Z_1, Z_2 satisfy the bounded support condition* (D.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,*

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

and

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

Proof. This lemma essentially follows from [6, Theorem 2.10], although the authors considered the case with $q \prec n^{-1/2}$ only. The results for larger q follows from [10, 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 (D.3), the lower bound for the smallest eigenvalue follows from the exactly the same arguments as in [10]. 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 E.3. *Suppose the assumptions of Lemma E.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:*

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

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

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

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

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

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 (D.15), it implies the estimates (E.4) and (E.5). □

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

Definition E.4 (Minors)**.** *For any $(p+n) \times (p+n)$ matrix $\mathcal{A}$ and $\mathbb{T} \subseteq \mathcal{I}$, we define the minor $\mathcal{A}^{(\mathbb{T})} := (\mathcal{A}_{\mathfrak{a}\mathfrak{b}} : \mathfrak{a}, \mathfrak{b} \in \mathcal{I} \setminus \mathbb{T})$ 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* (D.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 (D.14)*)*

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

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 E.5. *(Resolvent identities).*

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

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

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

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

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

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

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

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

Proof. All these identities can be proved directly using Schur's complement formula. The reader can also refer to, for example, [19, Lemma 4.4]. □

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

Lemma E.6 (Lemma 3.8 of [14])**.** *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 (D.8) with $q \leqslant n^{-\phi}$ for some constant $\phi > 0$. Then we have the following bound:*

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

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

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

E.2 Entrywise local law

The main goal of this subsection is to prove the following entrywise local law. The anisotropic local law (D.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 E.1, we have $q = n^{-1/2}$ and

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

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

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

Proof. The proof of Lemma E.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 (E.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 [14, Section 3] and [19, Section 5], we introduce the Z variables

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

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 (E.8), we have

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

and

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

For simplicity, we introduce the following random error

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

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

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

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

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

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

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

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

by (E.4). Similarly, applying Lemma E.6 to Z_μ in (E.17) and using (E.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 (E.9), Lemma E.6 and (E.4), we obtain that

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

and

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

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

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

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

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

Step 2: Self-consistent equations. This is the key step of the proof for Proposition E.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 (D.23) up to some small error $|\mathcal{E}_{2,3}|\prec n^{-1/2}$. Then applying Lemma D.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

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

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

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

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 (E.22) into (D.18), we get

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

Then we claim the following result.

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

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

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

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

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

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

where we recall (E.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 (E.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 (E.19) in the second step. Similarly, we can get that

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

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

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

With (E.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 (E.25), (E.29) and (E.19), we obtain that

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

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

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

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

As a byproduct, we obtain from the two estimates that

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

Together with (E.22), we get that

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

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

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

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

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

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

which further implies

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

Finally, plugging (E.31) and (E.32) into (E.38), we conclude (E.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 D.6 to (E.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

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

Once we know (E.39), then by (D.22) and (E.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 (E.39), we obtain that

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

and

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

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

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

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

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

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

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

Proof. By (D.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 (E.26), (E.27) and (E.29), we obtain that (E.37) and (E.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 (E.25) and (E.29), we obtain that (E.31) and (E.32) hold at $z=0$ even without the indicator function $\mathbf{1}(\Xi)$. Finally, plugging (E.31) and (E.32) into (E.38), we conclude (E.24) holds at $z=0$, that is,

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

Denoting $\omega_2 = -m_{2c}(0)$ and $\omega_3 = -m_{2c}(0)$. By (E.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

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

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

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

The first equation and (E.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 (E.46) gives that $|\omega_2 - b_2| \prec n^{-1/2}$. Together with the first equation in (E.46), we get $|\omega_3 - b_3| \prec n^{-1/2}$. This concludes (E.39). □

This lemma concludes (E.39), and as explained above, concludes the proof of Lemma E.7. □

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

Proof of Proposition E.1. With (E.15), one can repeat the polynomialization method in [6, Section 5] to get the anisotropic local law (D.26) for G_0. The proof is exactly the same, except for some minor notation difference, so we omit the details. □

E.3 Anisotropic local law

In this section, we finish the proof of Theorem D.7 for a general X satisfying the bounded support condition (D.8) with $q \leqslant n^{-\phi}$ for some constant $\phi > 0$. The proposition E.1 implies that (D.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 D.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 [19]. The proof is similar to the ones in Sections 7-8 of [19], so we only give an outline without writing down all the details.

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

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

We shall work on the probability space consisting of triples (Z^0, Z^θ, Z^1) of independent $n \times p$ random matrices, where the matrix $Z^\theta = (Z^\theta_{\mu i})$ has law

$$\prod_{i \in \mathcal{I}_1} \prod_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} \rho^\theta_{\mu i}(\mathrm{d}Z^\theta_{\mu i}). \tag{E.47}$$

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

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

We also introduce the matrices

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

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

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

provided all the expectations exist.

We shall apply (E.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

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

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 $G_{\mathbf{u}\mathbf{v}} := \mathbf{u}^\top G \mathbf{v}$. 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 (E.48) for any fixed $p \in 2\mathbb{N}$ and constant $\varepsilon > 0$:

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

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

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

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

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

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

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

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 (D.11)), we have for any $\lambda, \lambda' \in \mathbb{R}$ and $K \in \mathbb{N}$,

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

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

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

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 (E.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

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

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

$$\begin{aligned} &\mathbb{E} F^p_{\mathbf{v}}\left(Z^{\theta,Z^a_{\mu i}}_{(\mu i)}\right) - \mathbb{E} F^p_{\mathbf{v}}\left(Z^{\theta,0}_{(\mu i)}\right) = \mathbb{E}\left[f_{(\mu i)}\left(Z^a_{i\mu}\right) - f_{(\mu i)}(0) \right] \\ &= \mathbb{E}\, f_{(\mu i)}(0) + \frac{1}{2n} \mathbb{E}\, f^{(2)}_{(\mu i)}(0) + \sum_{r=4}^{p+4} \frac{1}{r!} \mathbb{E}\, f^{(r)}_{(\mu i)}(0)\, \mathbb{E}\left(Z^a_{i\mu}\right)^r + \mathrm{O}_\prec(q^{p+4}). \end{aligned} \tag{E.56}$$

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

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

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

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

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

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

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

Lemma E.12. *Suppose that*

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

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

Proof. The proof is the same as the one for [19, Lemma 7.16]. □

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

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

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

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

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

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

Then we can calculate the derivative

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

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

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

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

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

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 (E.61) we have

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

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

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

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

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

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

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

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

Finally, if the condition (E.57) does not hold, then there is also an $r = 3$ term in the Taylor expansion (E.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 (E.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 [19, Section 8]. The argument for our case is almost the same except for some notational differences, so we omit the details.

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

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

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

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

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 (D.21), and the solution satisfies (D.22). Now we abbreviate $\varepsilon(z) := m_{3c}(z) - m_{3c}(0)$, and from (E.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

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

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

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

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

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

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

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

and

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

for all $|z| \leqslant \widetilde{c}$ and $|\varepsilon_1| \leqslant \widetilde{c}$, $|\varepsilon_2| \leqslant \widetilde{c}$. Then with (E.70) and (E.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 (E.67) subject to the condition $\|\varepsilon\|_\infty \leqslant r$.

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

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

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 (D.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 (E.67).

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

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

With the contraction mapping, we have the bound

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

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

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

This gives (D.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 (D.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 (D.20). Thus under (D.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 (E.67) and (E.73), we can conclude (D.22) for $m_{2c}(z)$. This concludes the proof of Lemma D.5. □

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

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

with $|\mathcal{E}_2'(z)| + |\mathcal{E}_3'(z)| = \mathrm{O}(\delta(z))$. Then we subtract the equations (E.67) from (E.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 D.5, so we omit the details. □