Revisiting the Bias-Variance Tradeoff of Multi-Task and Transfer Learning in High Dimensions

Anonymous Author(s)
Affiliation
Address
email

Abstract

When does multi-task and transfer learning outperform single-task learning? In this work, we address this question by studying the test performance of predicting a particular task given multiple tasks using a commonly used hard parameter sharing architecture, in terms of bais-variance tradeoff. In the case of high-dimensional linear regression, we provide a sharp analysis of the bias-variance tradeoff for multi-task and transfer learning estimators. A key technical tool that we develop is the trace of $(X_1^\top X_1 + X_2^\top X_2)^{-1}$ for two random matrices X_1 and X_2 in the setting of two tasks. Based on the theory, we provide more precise interpretations of many empirical phenomena in multi-task and transfer learning. For example, we quantify the benefit of multi-task learning for reducing the amount of labeled data, which is a key finding in Taskonomy by Zamir et al.'18. Finally, we show practical implications of our theory for detecting and improving negative effects in image and text classification tasks.

1 Introduction

Multi-task learning has become a powerful paradigm to solve complex prediction tasks in computer vision [1, 2], natural language processing [3, 4] and numerous other areas [5]. By combining multiple information sources, it is possible to share new information among different sources in the same model [6]. Intuitively, the performance of multi-task learning depends on the relationship of the information sources. When the information sources are heterogeneous, negative transfer – where multi-task learning performs worse than single-task learning – has often been observed [7, 8]. While numerous studies have sought to alleviate negative transfer when it occurs [5], a rigorous understanding to the contributing causes of negative transfer has remained elusive in the literature [10]. In this work, we develop technical tools to better understand when and why negative transfer occurs for learning multiple linear regression tasks in high dimensions. We use the tools to show theoretical and practical implications for detecting and improving negative transfer.

Identifying negative transfer requires developing tight generalization bounds for both multi-task learning and single-task learning. In classical Rademacher or VC based theory of multi-task learning [11, 12, 13], the generalization bounds are usually presented in a way so that the error goes down as more labeled data is added. On the other hand, we have observed that adding more labeled data does not always improve performance in multi-task learning. More recent work has shown the benefit of learning multi-task representations for certain half-spaces [14] and multiple sparse regressions [15, 16]. In order to rigorously understand negative transfer, the technical challenge is to develop generalization bounds that scale tightly with the qualities of the data.

In this work, we consider a setting where multiple labeled linear regression tasks are available and focus on predicting a particular task whose amount of labeled data is limited. Following Hastie et al. [17] and Bartlett et al. [18], we assume that for every task $1 \leqslant i \leqslant t$, its features are random vectors

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

$x = \Sigma_i^{1/2} z$, where $z \in \mathbb{R}^p$ consists of i.i.d. entries with mean zero and unit variance, and $\Sigma_i \in \mathbb{R}^{p \times p}$ is a positive semidefinite matrix. Let n_i denote the data size and $X_i \in \mathbb{R}^{n_i \times p}$ denote the features of task i, for every $1 \leqslant i \leqslant t$. The label of task i is given by $Y_i = X_i \beta_i + \varepsilon_i$, where $\beta_i \in \mathbb{R}^p$ denotes the ground truth parameters for task i and ε_i denotes i.i.d. random noise with mean zero and variance σ^2. Without loss of generality, let the t-th task denote the target task. We assume that the data size of task t is a small constant $\rho_t > 1$ times p to capture the need for more labeled data.

We combine all the labeled data using a hard parameter sharing architecture that contains a shared body $B \in \mathbb{R}^{p \times r}$ for all tasks and a separate prediction head $\{W_i \in \mathbb{R}^r\}_{i=1}^t$ for each task [10, 19, 26]. This corresponds to minimizing the following objective.

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

For a target task t, let $\hat{\beta}_t^{\text{MTL}} = BW_t$ denote the optimal multi-task estimator by solving equation (1.1). Let $\hat{\beta}_t^{\text{STL}}$ denote the standard linear regression estimator using task t alone. We say there is negative transfer if the test error of $\hat{\beta}_t^{\text{MTL}}$ is smaller than that of $\hat{\beta}_t^{\text{STL}}$, or positive transfer otherwise.

Main results. We revisit the bias-variance tradeoff of the multi-task estimator. Interestingly, the variance of $\hat{\beta}_t^{\text{MTL}}$ is always smaller than that of $\hat{\beta}_t^{\text{STL}}$, hence resulting in a positive effect of variance reduction. The bias of $\hat{\beta}_t^{\text{MTL}}$, which we term as *model shift bias*, results in a negative effect caused by the difference between β_t and the rest $\{\beta_i\}_{i=1}^{t-1}$. Hence, the tradeoff between the variance reduction effect and model shift bias determines whether we observe positive or negative transfer.

To quantify the tradeoff more precisely, we develop a new technical result on the trace of $(X_1^\top X_1 + X_2^\top X_2)^{-1}$, for the setting of two tasks with general covariance matrices. Using the result, we provide a sharp bias-variance tradeoff of $\hat{\beta}_t^{\text{MTL}}$. When there are more than two tasks, we show a similar result if they all have the same covariates. This setting is prevalent in applications of multi-task learning to image classification, where there are multiple prediction labels/tasks for every image [1, 20]. Finally, we apply our technical tools to the related setting of transfer learning. We study the transfer function used by Taskonomy [2], which pools learnt representations from source tasks into a shared body similar to B in equation (1.1). We find that the model shift bias can be captured nicely by the orthogonal projection of β_t to the subspace spanned by $\{\beta_i\}_{i=1}^{t-1}$. These results are presented more precisely in Section 3.

The tradeoff that we developed in Section 3 depends on the qualities of each task dataset such as its data size and covariance matrix. In Section 4, we use the technical results to provide guidance on when positive transfer is likely to occur and identify causes of negative transfer.

- First, we provide a more precise interpretation to a folklore understanding in multi-task learning, which posits that negative transfer is caused by having dissimilar tasks. For an example of two tasks, we show a sharp transition from positive to negative transfer determined by the ratio of $\|\beta_1 - \beta_2\|^2$ and a certain function of data sizes (cf. Proposition 4.1). We further show that positive transfer is more likely to occur when transferring from a less noisy source task to a more noisy target task. In Section **??**, we validate the observation on text and image classification tasks.
- Second, we find that depending on how large $\|\beta_1 - \beta_2\|^2$ is, adding more labeled data from the source task does not always reduce the test error of $\hat{\beta}_t^{\text{MTL}}$ (cf. Proposition 4.2). We connect the observation to explain a key finding of Taskonomy [2]. We define the *data efficiency ratio* as the smallest $\alpha \in (0, 1)$ such that if we only use an α fraction of labeled data, then the test error of $\hat{\beta}_t^{\text{MTL}}$ matches that of the $\hat{\beta}_t^{\text{STL}}$ on the entire set. In Proposition 4.5, we show that the data efficiency ratio of an illustrative example is at most $\frac{1}{2\rho_t}$. In Section **??**, we validate that performing multi-task learning can reduce the need for labeled data on 6 sentiment analysis tasks.
- Finally, we find that covariate shift, i.e. having different covariance matrices, is another cause for suboptimal performance for $\hat{\beta}_t^{\text{MTL}}$. We show that as n_1/n_2 becomes large, the best performing source task has have the same covariance matrix as the target task (cf. Proposition 4.3).

The technical tool we developed extends a well-known result in the random matrix theory literature on the trace of $(X^\top X)^{-1}$ [21] for a single random matrix to two random matrices. Our error bound for the asymptotic limit is nearly optimal, which may be of independent interest.

We show practical implications of our theory on text and image classification tasks. First, we provide a metric to determine positive versus negative transfer by comparing the test accuracies of single-task models. Second, [Todo:] Finally, we show that as the data size between the source and target task becomes more imbalanced, aligning the covariances of the tasks becomes more beneficial.

2 Preliminaries

We describe the bias-variance tradeoff of our setting as a warmup. Recall that we have t labeled tasks available, denoted by $(X_1, Y_1), (X_2, Y_2), \ldots, (X_t, Y_t)$, where $X_i \in \mathbb{R}^{n_i \times p}$ and $Y_i \in \mathbb{R}^{n_i}$ for $1 \leqslant i \leqslant t$. Without loss of generality, let the t-th task denote the target task. For an estimator $\hat{\beta} \in \mathbb{R}^p$, the test error of the target task is defined as

$$te_t(\hat{\beta}) := \mathop{\mathbb{E}}_{z}\left[\mathop{\mathbb{E}}_{\varepsilon_t}\left[\left((\Sigma_t^{1/2} z)^\top \hat{\beta} - (\Sigma_t^{1/2})^\top \beta_t\right)^2\right]\right] = \mathop{\mathbb{E}}_{\varepsilon_t}\left[(\hat{\beta} - \beta_t)^\top \Sigma_t (\hat{\beta} - \beta_t)\right].$$

The single-task estimator $\hat{\beta}_t^{\text{STL}}$ is given by $(X_t^\top X_t)^{-1} X_t^\top Y_t$. The bias-variance trade-off [22] says

$$te_t(\hat{\beta}) = \left\| \mathop{\mathbb{E}}_{\varepsilon_t}\left[\hat{\beta}\right] - \beta_t \right\|^2 + \mathop{\mathbb{E}}_{\varepsilon_t}\left[\left\| \hat{\beta} - \mathop{\mathbb{E}}_{\varepsilon_t}\left[\hat{\beta}\right] \right\|^2\right].$$

In order to study the trade-off between model-shift bias and variance reduction, we need tight concentration bounds to quantify both effects. For this purpose, we consider the high-dimensional regime where n_i is a fixed constant $\rho_i > 1$ times p for every $1 \leqslant i \leqslant t$, and p is large.

We focus on a setting where ρ_t is a small constant. This setting captures the need for adding more labeled data to reduce the test error of the target task. A well-known result for this setting states that $te_t(\hat{\beta}_t^{\text{STL}}) = \sigma^2 \cdot \text{Tr}[(X_t^\top X_t)^{-1}\Sigma_t]$ is concentrated around $\frac{\sigma^2}{\rho_t - 1}$ (e.g. Chapter 6 of [21]), which scales with the data size and noise level of the target task. However, this result only applies to the single-task setting. Therefore, our goal is to extend this result to the multi-task setting.

To illustrate our intuition, we begin by considering the setting of two tasks with general covariance matrices. Recall that $\hat{\beta}_t^{\text{MTL}}$ is defined as BW_t after solving equation (1.1). We decompose the test error of $\hat{\beta}_t^{\text{MTL}}$ on the target task into two parts (to be derived in Appendix A) as follows

$$te_t(\hat{\beta}_t^{\text{MTL}}) = \hat{v}^2 \left\| \Sigma_2^{1/2} (\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 (\beta_1 - \hat{v}\beta_2) \right\|^2 \tag{2.1}$$

$$+ \sigma^2 \cdot \text{Tr}\left[(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_2\right], \tag{2.2}$$

where $\hat{v} = W_2/W_1$ denotes the ratio of the output layer weights.

Notations. When there is no ambiguity, we drop the subscript t from $te_t(\hat{\beta}_t^{\text{MTL}})$ to $te(\hat{\beta}_t^{\text{MTL}})$ for simplicity. We refer to the first task as the source task when there are only two tasks. We call $M = \Sigma_1^{1/2}\Sigma_2^{-1/2}$ the covariate shift matrix.

3 A Technical Tool to Quantify the Bias-Variance Trade-off

We develop a technical tool to derive $te(\hat{\beta}_t^{\text{MTL}})$ that only depends on the qualities of task data such as data sizes and covariance matrices, for two tasks with general covariances. Then, we extend the result to more than two tasks that share the same features but have different labels. Finally, we show a sharp bias-variance tradeoff for transfer learning settings.

Multi-task learning. applies to the setting of two tasks where their covariance matrices may be arbitrarily different. We derive a precise limit of $\text{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_2\right]$, which is a deterministic function that only depends on Σ_1, Σ_2 and $n_1/p, n_2/p$ (see Lemma A.3 in Appendix A.1 for the result). Based on the result, we show how to determine positive versus negative transfer as follows.

Theorem 3.1 (Informal). *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 $n_1 = \rho_1 p$ and $n_2 = \rho_2 p$, where $\rho_1 > 1$ and $\rho_2 > 1$ are fixed constants. There exists two deterministic functions Δ_β and Δ_{var} and a small deterministic error δ that only depend on $\{\hat{v}, \Sigma_1, n_1, n_2, \beta_1, \beta_2\}$ 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 3.1 shows nearly tight bounds on the trade-off between model-shift bias and variance reduction. The bounds get tighter and tighter as ρ_1 increases. While the general form of Δ_{var} and Δ_β can be quite complex, we will show that they provide nice interpretation for simplified settings later on in Section 4.

We describe an overview of the proof of Theorem 3.1. By comparing the bias and variance of $te_t(\hat{\beta}_t^{\text{MTL}})$ to $te_t(\hat{\beta}_t^{\text{STL}})$, we observe the following two effects, respectively. We term equation (2.1) as *model-shift bias*, which captures how similar β_1 and β_2 are. This part introduces a negative effect to multi-task learning. The scaling term $\hat{v}$ corresponds to the intuition that the shared module B learns a subspace for both tasks. The output layer of each task scales the direction of B suitably to fit the task.

Next, it is not hard to verify that equation (2.2), the variance of $\hat{\beta}_t^{\text{MTL}}$, is always smaller than the variance of $\hat{\beta}_t^{\text{STL}}$. This part introduces a positive variance reduction effect to performing multi-task learning. Hence, whether $te(\hat{\beta}_t^{\text{MTL}})$ is lower than $te(\hat{\beta}_t^{\text{STL}})$ is determined by the trade-off between two effects: (i) the positive effect from variance reduction; (ii) the negative effect from model shift bias. A formal version of Theorem 3.1 is presented in Theorem A.5 and its proof is presented in Appendix A.1.

Extension. The above result also applies to the setting of more than two tasks with 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 $\{\beta_i\}_{i=1}^k$. We derive a similar trade-off between model shift bias and variance reduction for this setting as well. The formal statement is stated in Theorem A.8 and its proof can be found in Appendix A.2.

Transfer learning. We extend the intuition behind Theorem 3.1 to transfer learning settings. We provide an analysis of the transfer function of Taskonomy [2] using our setup. Specifically, the source task encoder consists of the representations learnt from one or more source tasks. The transfer function then tries to fit the target task data to the source task encoder. For more details, we refer the reader to Figure 4 in Taskonomy.

We map the procedure to our setup as follows. First, we obtain the single-task estimator $\hat{\beta}_i$ from the source tasks, for $1 \leqslant i \leqslant t-1$. This forms the shared representation $B_{t-1}^\star = [\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{t-1}]$. Then, we learn the output layer W_t on the target task by minimizing the following objective

$$g(W_t) = \|X_t B W_t - Y_t\|^2 . \tag{3.1}$$

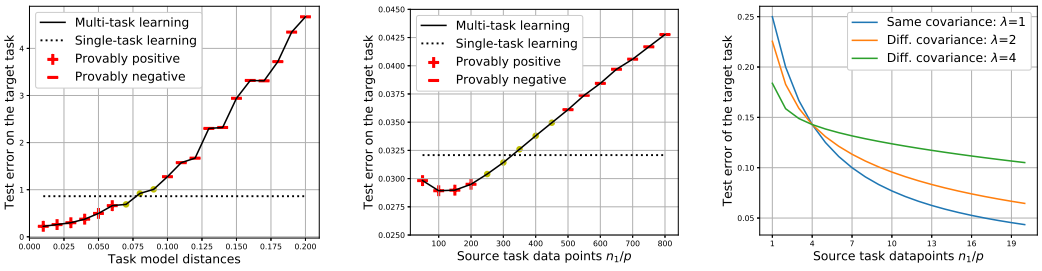
After solving W_t, we use $\hat{\beta}_t^{\text{TL}} = BW_t$ as the estimator for the target task. By comparing $te(\hat{\beta}_t^{\text{TL}})$ to $te(\hat{\beta}_t^{\text{STL}})$, we observe a similar trade-off between model-shift bias and variance reduction for this setting. The formal statement is presented in Theorem A.9 and its proof in Appendix A.3.

4 Theoretical Implications

Based on Theorem 3.1, we provide precise interpretations of many empirical phenomena in multi-task learning. In Section 4.1, we explain three contributing causes of negative transfer, including model dissimilarity, large source data size and covariate shift. In Section 4.2, we describe two implications for positive transfer, including labeled data efficiency and de-noising.

4.1 Contributing Causes of Negative Transfer

It is well-known since the seminal work of Caruana [6] that how well multi-task learning performs depends on task relatedness. As a warmup, we first describe an example that rigorously formalizes the connection using our setup.



(a) Model dissimilarity (b) Large source task data size (c) Covariate shift

Figure 1: Three contributing causes of negative transfer: (a) As model dissimilarity increases, we observe a transition from positive to negative transfer (Proposition 4.1). (b) As source task data size increases, we observe a transition positive to negative transfer (Proposition 4.2). (c) As covariate shift becomes more severe, test performance gets worse (Proposition 4.3).

The isotropic model. Consider two tasks with isotropic covariances $\Sigma_1 = \Sigma_2 = \mathrm{Id}$. Recall that each task has data size $n_i = \rho_i \cdot p$, for $i = 1, 2$. And $X_1 \in \mathbb{R}^{n_1 \times p}$, $X_2 \in \mathbb{R}^{n_2 \times p}$ denotes the covariates of the two tasks, respectively. We assume that for the target task, β_2 has i.i.d. entries with mean zero and variance κ^2. For the source task, β_1 equals β_2 plus i.i.d. entries with mean 0 and variance d^2. The labels are given by $Y_i = X_i\beta_i + \varepsilon_i$, where ε_i consists of i.i.d. entries with mean zero and variance σ_i^2, for $i = 1, 2$. [1]

Model dissimilarity. We measure model dissimilarity as $\|\beta_1 - \beta_2\|^2$, which is the distance between source and target in the isotropic model. We derive a sharp threshold when positive transfer transitions to negative transfer, as model dissimilarity increases. Define the following function

$$\Phi(\rho_1, \rho_2) = \frac{(\rho_1 + \rho_2 - 1)^2}{\rho_1(\rho_1 + \rho_2)(\rho_2 - 1)}.$$

Proposition 4.1. *In the isotropic model, assume that $\rho_1 > 40$ is a fixed constant. Let $\sigma_1 = \sigma_2 = \sigma$. Whether $te(\hat{\beta}_t^{MTL})$ is lower than $te(\hat{\beta}_t^{STL})$ is determined by the ratio between model dissimilarity and $\Phi(\rho_1, \rho_2)$:*

- *If $pd^2 < \frac{1}{2}\sigma^2 \cdot \Phi(\rho_1, \rho_2)$, then whp we have that $te(\hat{\beta}_t^{MTL}) < te(\hat{\beta}_t^{STL})$.*
- *If $pd^2 \geqslant 2\sigma^2 \cdot \Phi(\rho_1, \rho_2)$, then whp we have that $te(\hat{\beta}_t^{MTL}) \geqslant te(\hat{\beta}_t^{STL})$.*

Proposition 4.1 simplifies Theorem 3.1 in the isotropic model, allowing for a more explicit statement of the bias-variance tradeoff. We remark that the constants $1/2$ and 2 in the statement above are chosen for ease of presentation. In general we can close their gap by increasing ρ_1.

[Todo:] We illustrate the example with a simulation. We consider a setting where $p = 200$, $n_1 = 90p$, $n_2 = 30p$. We fix the target task and vary the source task, in particular the parameter d which determines $\|\beta_1 - \beta_2\|$. Figure 1a shows the result. We observe that Proposition 4.1 explains most of the observations in Figure 1a.

The proof of Proposition 4.1 involves two parts. First, in equation (2.2), the positive variance reduction effect scales with $n_1 = \rho_1 p$, the number of source task data points. Second, we show that the negative effect of model-shift bias scales with pd^2, which is the expectation of $\|\beta_1 - \beta_2\|^2$. The proof, which is based on Theorem 3.1, can be found in Appendix B.1.

Data size. In classical Rademacher or VC based theory of multi-task learning, the generalization bounds are usually presented for settings where the data sizes are equal for all tasks [11, 13, 14]. More generally, such results are still applicable when all task data are being added simultaneously. On the other hand, for many applications of multi-task learning, the data sources are usually heterogeneous. For such settings, we have observed that adding more labeled data from one task does not always help. Using the isotropic model, we consider what happens if we vary the source task data size.

[1] For simplicity, we assume that all the random variables have subexponential decay, while keeping in mind that our results can be applied under weaker moments assumptions as shown in the supplementary material.

Proposition 4.2. *In the isotropic model, assume that $\rho_1 > 40$ and $\rho_2 > 500$ are fixed constants. We have the following conditions to determine whether $te(\hat{\beta}_t^{MTL})$ is lower than $te(\hat{\beta}_t^{STL})$:*

a) If $pd^2 \leqslant \frac{1}{2} \cdot \frac{\sigma^2}{\rho_2 - 1}$, then whp $te(\hat{\beta}_t^{MTL}) < te(\hat{\beta}_t^{STL})$ for any $\rho_1 > 40$.

b) If $pd^2 > 2 \cdot \frac{\sigma^2}{\rho_2 - 1}$, then we have the following transition depending on ρ_1:

- *If $\rho_1 \cdot p < \frac{(\rho_2-2)\sigma^2}{(1+\rho_1{}^{-0.5})^4(\rho_2-1)d^2-\sigma^2/p}$, then whp $te(\hat{\beta}_t^{MTL}) < te(\hat{\beta}_t^{STL})$.*
- *If $\rho_1 \cdot p > \frac{(\rho_2-2)\sigma^2}{(1-\rho_1{}^{-0.5})^4(\rho_2-3)d^2-\sigma^2/p}$, then whp $te(\hat{\beta}_t^{MTL}) > te(\hat{\beta}_t^{STL})$.*

The intuition behind Proposition 4.2 is as follows. a) If $\|\beta_1 - \beta_2\|^2 \approx pd^2$ is sufficiently small, then adding any amount of labeled data from the source task always provides positive transfer; ii) Otherwise, as source data size increases, we observe a transition from positive to negative transfer. The proof of Proposition 4.2 is similar to Proposition 4.1, and can be found in Appendix B.1.

Covariate shift. So far we have considered the isotropic model where $\Sigma_1 = \Sigma_2$. This setting is relevant for settings where different tasks share the same input features such as multi-class image classification. In general, the covariance matrices of the two tasks may be different such as in text classification. In this part, we consider what happens when $\Sigma_1 \neq \Sigma_2$.

We measure covariate shift 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 [23, 24, 9]. We show that if n_1 is much larger than n_2, then the optimal M matrix should be proportional to identity, under certain assumptions on its range of singular values (to be formulated in Proposition 4.3). 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.

To compare different choices of M on the performance of $\hat{\beta}_t^{\text{MTL}}$, 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. Similar to the isotropic model, we assume that β_2 has i.i.d. entries with mean zero and variance κ^2 and β_1 equals β_2 plus i.i.d. entries with mean 0 and variance d^2. The following proposition shows that when n_1 is much larger than n_2, $te(\hat{\beta}^{\text{MTL}})$ is minimized at $M = \mu\,\text{Id}$ among all matrices in $\mathcal{S}_\mu$.

Proposition 4.3. *In the setting described above, assume that $\rho_1 > 3$ and $\rho_2 > 3$, and $\|\Sigma_1\| \leqslant C_1$ for some constant $C_1 > 0$. Let $g(M)$ denote the test error of $\hat{\beta}_t^{MTL}$ when the covariance shift matrix is equal to $M \in \mathcal{S}_\mu$. We have that*

$$g(\mu\,\text{Id}) \leqslant \left(1 + \text{O}\left(\frac{\rho_2}{\rho_1} + \frac{1}{\sqrt{\rho_1}}\right)\right) \min_{M \in \mathcal{S}_\mu} g(M).$$

Proposition 4.3 implies that when $\rho_1 \gg \rho_2$, having no covariate shift is the optimal choice for choosing the source task.

[Todo:] To complement the result, we show an example when the statement is not true if $n_1 \leqslant n_2$.

Example 4.4. In the setting of Proposition 4.3, 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 2, 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 can be found in Appendix B.1.

4.2 Implications for Positive Transfer

On the positive side, we describe two implications for positive transfer, assuming that $\|\beta_1 - \beta_2\|^2$ is small in the isotropic model.

Labeled data efficiency. We use our tools to explain a key result of taskonomy [2], 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 the simplified model as follows.

Proposition 4.5. *In the simplified model, assume that $\rho_1 = \rho_2 = \rho \geqslant 200$ and $d^2 < 8\sigma^2/(3p\rho)$. Then the data efficiency ratio is at most $\frac{1}{2\rho} + \frac{\sigma^2}{2\sigma^2 - 3pd^2\rho/4}$.*

Note that we have stated the result assuming that $\rho_1 = \rho_2$. Similar results can also be obtained when they are different. We omit the details. The proof of Proposition 4.5 can be found in Appendix B.2.

Labeled data de-noising. We further show that multi-task learning is particular powerful when the labeled data of the source task is less noisy compared to the target task. Consider a more general setting of Proposition 4.1 where the noise level σ_1 of task 1 differs from the noise level σ_2 of task 2. We derive a sharp transition from positive to negative transfer as a parameter of σ_1^2.

Proposition 4.6. *In the simplified model with d being fixed but σ_1 varies, assume that $\rho_1 > 50$ is a fixed constant and $d^2 < \frac{\sigma_2^2}{2p} \cdot \Phi(\rho_1, \rho_2)$. We derive the following transition as a parameter of σ_1^2:*

- *If $\sigma_1^2 \leqslant -\frac{3}{2}\rho_1 \cdot pd^2 + \big(1 + \frac{3}{4}\rho_1\Phi(\rho_1, \rho_2)\big) \cdot \sigma_2^2$, then whp $te(\hat{\beta}_t^{MTL}) < te(\hat{\beta}_t^{STL})$.*
- *If $\sigma_1^2 > -\frac{1}{2}\rho_1 \cdot pd^2 + \big(1 + \frac{3}{2}\rho_1\Phi(\rho_1, \rho_2)\big) \cdot \sigma_2^2$, then whp $te(\hat{\beta}_t^{MTL}) > te(\hat{\beta}_t^{STL})$.*

As a corollary, if $\sigma_1^2 \leqslant \sigma_2^2$, then we always get positive transfer. The proof of Proposition 4.6 is similar to Proposition 4.1. The details can be found in Appendix B.2.

5 Practical Implications for Multi-Task Learning

5.1 Experimental Setup

Datasets and models. We describe the datasets and models we use in the experiments.

Sentiment Analysis: This dataset includes six tasks: movie review sentiment (MR), sentence subjectivity (SUBJ), customer reviews polarity (CR), question type (TREC), opinion polarity (MPQA), and the Stanford sentiment treebank (SST) tasks.

For each task, the goal is to categorize sentiment opinions expressed in the text. We use an embedding layer (with GloVe embeddings[2]) followed by an LSTM layer proposed by [25].

ChestX-ray14: This dataset contains 112,120 frontal-view X-ray images and each image has up to 14 diseases. This is a 14-task multi-label image classification problem.

For all models, we share the main module across all tasks and assign a separate regression or classification layer on top of the shared module for each tasks.

5.2 Detecting and Improving Negative Effects

A Metric to Determine Positive or Negative Transfer We propose a simple metric to determine whether multi-task learning performs better than single-task learning. The metric is as follows. First, we train 2 single-task models on 2 separate tasks. We take the prediction accuracy of the two tasks. Let τ be a fixed threshold. If the accuracy of the source task is higher than the accuracy of the target

[2]http://nlp.stanford.edu/data/wordvecs/glove.6B.zip

Threshold	Text classification		ChestX-ray14	
	Precision	Recall	Precision	Recall
0.0	0.596	1.000	0.593	1.000
0.1	0.756	0.388	0.738	0.462
0.2	0.919	0.065	0.875	0.044
0.3	1.000	0.004	-	-

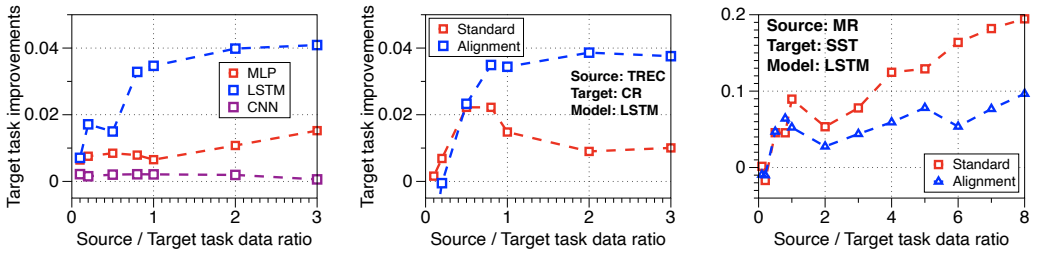
Table 1: Ablation study on when should use MTL via different source/target task accuracy. Note: For text classification tasks, the source task training data size ranges from 500 to 1,500 and target task training data size is 1000; For ChestX-ray14, the training data size is 10,000.

task by τ, then we predict positive transfer. On the other hand, if the accuracy of the source task is lower than the accuracy of the target task by τ, then we predict negative transfer.

Table 1 shows the results on a sentiment analysis and an image classification task.

Covariance Alignment for Imbalanced Tasks with Covariate Shift. We validate that the performance of aligning task covariances depends on the size of source task data. Recall that the covariance alignment procedure in [26] adds an additional module between the word embedding representation and the shared module.

In Figure 2, we observe that the benefit from aligning task covariances becomes more significant as we increase the size of the source task dataset.



(a) Averaged (b) Task pair TREC and CR (c) Task pair MR and SST

Figure 2: The performance of aligning task covariances depends on data size. As the ratio between source task data size and target task data size increases, the performance improvement from aligning task covariances increases.

5.3 Validating Theoretical Implications

Causes of negative transfer.

Benefits of multi-task learning.

Improving labeled data efficiency. We measure the data efficiency ratio on text classification tasks. We find that by performing multi-task learning, only 40% of the data is needed to achieve comparable performance to single-task learning over all six tasks.

Transferring from a high accuracy to low accuracy.

6 Related Work

Adding a regularization over B, e.g. [15, 16]. Moreover, [27] observed that controlling the capacity can outperform the implicit capacity control of adding regularization over B.

References

[1] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[2] 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.

[3] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

[4] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275, 2019.

[5] Yu Zhang and Qiang Yang. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*, 2017.

[6] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[7] Héctor Martínez Alonso and Barbara Plank. When is multitask learning effective? semantic sequence prediction under varying data conditions. *arXiv preprint arXiv:1612.02251*, 2016.

[8] Joachim Bingel and Anders Søgaard. Identifying beneficial task relations for multi-task learning in deep neural networks. *arXiv preprint arXiv:1702.08303*, 2017.

[9] 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.

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

[11] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[12] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

[13] Andreas Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7(Jan):117–139, 2006.

[14] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884, 2016.

[15] 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.

[16] 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.

[17] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[18] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

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

[20] 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.

[21] Vadim Ivanovich Serdobolskii. *Multiparametric statistics*. Elsevier, 2007.

[22] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[23] 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.

[24] 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.

[25] Tao Lei, Yu Zhang, Sida I Wang, Hui Dai, and Yoav Artzi. Simple recurrent units for highly parallelizable recurrence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4470–4481, 2018.

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

[27] 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.

[28] Theodore W Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 2003.

[29] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, 2nd edition, 2010.

[30] 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.

[31] Alexandru Nica and Roland Speicher. *Lectures on the combinatorics of free probability*, volume 13. Cambridge University Press, 2006.

[32] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1:457, 1967.

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

[34] 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.

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

[36] 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.

[37] 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.

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

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

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

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

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

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

[44] 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.

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

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

[47] 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.

[48] 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.

[49] 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.

A Supplementary Materials for the Technical Results

From [26], we know that we need to explicitly restrict the capacity r of B so that there is transfer between the two tasks. for the rest of the section, we shall consider the case when $r = 1$ we are considering the case of two tasks. Here, equation (1.1) simplifies to the following

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

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

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

As a remark, when $w_1 = w_2 = 1$, we recover the linear regression estimator. The advantage of using $f(B; w_1, w_2)$ is that if θ_1 is a scaling of θ_2, then this case can be solved optimally using equation (A.1) [27].

Defining the multi-task learning estimator. Suppose that the entries of ε_1 and ε_2 have variance σ^2. Using a validation set that is sub-sampled from the original training dataset, we get a validation loss as follows

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

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

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

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

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

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

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

We now state several helper lemmas to get a bound on the variance of $\hat{\beta}_t^{\text{STL}}$ and $\hat{\beta}_t^{\text{MTL}}$. Before that, we first need to fix the setting for the discussions in this section.

Assumption A.1. *We will consider two $n \times p$ random matrices of the form $X = Z\Sigma^{1/2}$, where Σ is a $p \times p$ deterministic positive definite symmetric matrices, and $Z = (z_{ij})$ is an $n \times p$ random matrix with real i.i.d. entries with mean zero and variance one. Note that the rows of X are i.i.d. centered random vectors with covariance matrix Σ. For simplicity, we assume that all the moments of z_{ij} exists, that is, for any $k \in \mathbb{N}$,*

$$\mathbb{E}|z_{ij}|^k \leqslant C_k, \quad 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p, \tag{A.4}$$

for some constant $C_k > 0$. We assume that $n = \rho p$ for some fixed constant $\rho > 1$. Without loss of generality, after a rescaling we can assume that the norm of Σ is bounded by a constant $C > 0$. Moreover, we assume that Σ is well-conditioned: $\kappa(\Sigma) \leqslant C$, where $\kappa(\cdot)$ denotes the condition number.

Here we have assumed (A.4) solely for simplicity of representation. if the entries of Z only have finite a-th moment for some $a > 4$, then all the results below still hold except that we need to replace $\mathrm{O}(p^{-\frac{1}{2}+\varepsilon})$ with $\mathrm{O}(p^{-\frac{1}{2}+\frac{2}{a}+\varepsilon})$ in the error bounds. We will not get deeper into this issue in this section, but refer the reader to Corollary C.8 below.

The first lemma, which is a folklore result in random matrix theory, helps to determine the asymptotic limit of $te(\hat{\beta}_t^{\mathrm{STL}})$, as $p \to \infty$. When the entries of X are multivariate Gaussian, this lemma recovers the classical result for the mean of inverse Wishart distribution [28]. For general non-Gaussian random matrices, it can be obtained from Stieltjes transform method; see e.g., Lemma 3.11 of [29]. Here we shall state a result obtained from Theorem 2.4 in [30], which gives an almost sharp error bound. Throughout the rest of this section, we shall say an event Ξ holds with high probability (whp) if $\mathbb{P}(\Xi) \to 1$ as $p \to \infty$.

Lemma A.2. *Suppose X satisfies assumption A.1. Let A be any $p \times p$ matrix that is independent of X. We have that for any constant $\varepsilon > 0$,*

$$\mathrm{Tr}\left[(X^\top X)^{-1} A\right] = \frac{1}{\rho - 1} \frac{1}{p} \mathrm{Tr}(\Sigma^{-1} A) + \mathrm{O}\left(\|A\| p^{-1/2+\varepsilon}\right) \tag{A.5}$$

with high probability.

We shall refer to random matrices of the form $X^\top X$ as sample covariance matrices following the standard notations in high-dimensional statistics. The second lemma extends Lemma A.2 for a single sample covariance matrices to the sum of two independent sample covariance matrices. It is the main random matrix theoretical input of this paper.

Lemma A.3. *Suppose $X_1 = Z_1 \Sigma_1^{1/2} \in \mathbb{R}^{n_1 \times p}$ and $X_2 = Z_2 \Sigma_2^{1/2} \in \mathbb{R}^{n_2 \times p}$ satisfy Assumption A.1 with $\rho_1 := n_1/p > 1$ and $\rho_2 := n_2/p > 1$. 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 descending order. Let A be any $p \times p$ matrix that is independent of X_1 and X_2. We have that for any constant $\varepsilon > 0$,*

$$\mathrm{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1} A\right] = \frac{1}{\rho_1 + \rho_2} \cdot \frac{1}{p} \mathrm{Tr}\left[(a_1 \Sigma_1 + a_2 \Sigma_2)^{-1} A\right] + \mathrm{O}\left(\|A\| p^{-1/2+\varepsilon}\right), \tag{A.6}$$

with high probability, where (a_1, a_2) is the solution to the following deterministic equations:

$$a_1 + a_2 = 1 - \frac{1}{\rho_1 + \rho_2}, \quad a_1 + \frac{1}{\rho_1 + \rho_2} \cdot \frac{1}{p} \sum_{i=1}^{p} \frac{\lambda_i^2 a_1}{\lambda_i^2 a_1 + a_2} = \frac{\rho_1}{\rho_1 + \rho_2}.$$

Finally, the last lemma describes the asymptotic limit of $(X_1^\top X_1 + X_2^\top X_2)^{-2}$.

Lemma A.4. *In the setting of Lemma A.3, let $\beta \in \mathbb{R}^p$ be any unit vector that is independent of X_1 and X_2. We have that for any constant $\varepsilon > 0$,*

$$\begin{aligned} &(n_1 + n_2)^2 \left\| \Sigma_2^{1/2} (X_1^\top X_1 + X_2^\top X_2)^{-1} \beta \right\|^2 \\ &= \beta^\top \Sigma_2^{-1/2} \frac{(1 + a_3)\,\mathrm{Id} + a_4 M^\top M}{(a_2 + a_1 M^\top M)^2} \Sigma_2^{-1/2} \beta + \mathrm{O}(n^{-1/2+\varepsilon}), \end{aligned} \tag{A.7}$$

with high probability, where a_3 and a_4 satisfy the following system of linear equations:

$$\left(\rho_2 a_2^{-2} - b_0\right) \cdot a_3 - b_1 \cdot a_4 = b_0, \quad \left(\rho_1 a_1^{-2} - b_2\right) \cdot a_4 - b_1 \cdot a_3 = b_1. \tag{A.8}$$

Here b_0, b_1 and b_2 are defined as

$$b_k := \frac{1}{p} \sum_{i=1}^{p} \frac{\lambda_i^{2k}}{(a_2 + \lambda_i^2 a_1)^2}, \quad k = 0, 1, 2.$$

The proof of Lemma A.3 and Lemma A.4 is a main focus of Section C. We remark that one can probably derive the same asymptotic result using free probability theory (see e.g. [31]), but our results (A.6) and (A.7) also give an almost sharp error bound $\mathrm{O}\left(p^{-1/2+\varepsilon}\right)$.

A.1 Proof of Two Tasks with General Covariance

In this section, we state and prove the formal version of Theorem 3.1. First, we need to introduce several quantities that will be used in our statement.

Given the optimal ratio $\hat{v}$, let $\hat{M} = \hat{v}\Sigma_1^{1/2}\Sigma_2^{-1/2}$ denote the weighted covariate shift matrix. Denote by $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \ldots \geqslant \hat{\lambda}_p$ the eigenvalues of $\hat{M}^\top \hat{M}$. Let $(\hat{a}_1, \hat{a}_2)$ be the solutions to the following deterministic equations

$$\hat{a}_1 + \hat{a}_2 = 1 - \frac{1}{\rho_1 + \rho_2}, \quad \hat{a}_1 + \frac{1}{\rho_1 + \rho_2} \cdot \frac{1}{p} \sum_{i=1}^{p} \frac{\hat{\lambda}_i^2 \hat{a}_1}{\hat{\lambda}_i^2 \hat{a}_1 + \hat{a}_2} = \frac{\rho_1}{\rho_1 + \rho_2}. \tag{A.9}$$

After obtaining $(\hat{a}_1, \hat{a}_2)$, we can solve the following linear equations to get $(\hat{a}_3, \hat{a}_4)$:

$$\left(\rho_2 \hat{a}_2^{-2} - \hat{b}_0\right) \cdot \hat{a}_3 - \hat{b}_1 \cdot \hat{a}_4 = \hat{b}_0, \quad \left(\rho_1 \hat{a}_1^{-2} - \hat{b}_2\right) \cdot \hat{a}_4 - \hat{b}_1 \cdot \hat{a}_3 = \hat{b}_1. \tag{A.10}$$

where we denoted

$$\hat{b}_k := \frac{1}{p} \sum_{i=1}^{p} \frac{\hat{\lambda}_i^{2k}}{(\hat{a}_2 + \hat{\lambda}_i^2 \hat{a}_1)^2}, \quad k = 0, 1, 2.$$

Then we introduce the following matrix

$$\Pi = \frac{\rho_1^2}{(\rho_1 + \rho_2)^2} \cdot \hat{M} \frac{(1 + \hat{a}_3)\,\mathrm{Id} + \hat{a}_4 \hat{M}^\top \hat{M}}{(\hat{a}_2 + \hat{a}_1 \hat{M}^\top \hat{M})^2} \hat{M}^\top,$$

which is defined in a way such that *in certain sense* it is the asymptotic limit of the random matrix

$$\hat{v}\Sigma_1^{1/2} (\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_2 (\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_1^{1/2}.$$

We introduce two factors that will appear often in our statements and discussions:

$$\alpha_-(\rho_1) := (1 - \rho_1^{-1/2})^2, \quad \alpha_+(\rho_1) := (1 + \rho_1^{-1/2})^2.$$

In fact, α_-^2 and α_+^2 correspond to the largest and smallest singular values of $Z_1/\sqrt{n_1}$, respectively, as given by the famous Marčenko-Pastur law [32]. In particular, as ρ_1 increases, both α_- and α_+ will converge to 1 and $Z_1/\sqrt{n_1}$ will be more close to an isometry. Finally, we introduce the error term

$$\delta := \frac{\alpha_+^2(\rho_1) - 1}{\alpha_-^2(\rho_1)} \cdot \kappa^4(\hat{M}) \cdot \|\Sigma_1^{1/2}(\beta_s - \hat{w}\beta_t)\|^2,$$

where we recall that $\kappa(\hat{M})$ is the condition number of $\hat{M}$. Note that this factor converges to 0 as $\rho_1 \to \infty$.

Now we are ready to state our main result for two tasks with both covariate and model shift. It shows that the information transfer is solely determined by two deterministic quantities Δ_β and Δ_{var}, which give the change of model shift bias and the change of variance, respectively.

Theorem A.5. *For $i = 1, 2$, let $Y_i = X_i\beta_i + \varepsilon_i$ be two independent data models, where $X_i = Z_i\Sigma_i^{1/2} \in \mathbb{R}^{n_i \times p}$ satisfy Assumption A.1 with $\rho_i := n_i/p > 1$ being fixed constants, and $\varepsilon_i \in \mathbb{R}^{n_i}$ are random vectors with i.i.d. entries with mean zero, variance σ^2 and all moments as in* (A.4)*. Then with high probability, we have*

$$te(\hat{\beta}_t^{MTL}) \leqslant te(\hat{\beta}_t^{STL}) \quad \textit{when:} \quad \Delta_{var} - \Delta_\beta \geqslant \delta \tag{A.11}$$

$$te(\hat{\beta}_t^{MTL}) \geqslant te(\hat{\beta}_t^{STL}) \quad \textit{when:} \quad \Delta_{var} - \Delta_\beta \leqslant -\delta, \tag{A.12}$$

where

$$\Delta_{var} := \sigma^2 \left(\frac{1}{\rho_2 - 1} - \frac{1}{\rho_1 + \rho_2} \cdot \frac{1}{p} \operatorname{Tr}\left[(\hat{a}_1 \hat{M}^\top \hat{M} + \hat{a}_2 \operatorname{Id})^{-1} \right] \right) \tag{A.13}$$

$$\Delta_\beta := (\beta_s - \hat{v}\beta_t)^\top \Sigma_1^{1/2} \Pi \Sigma_1^{1/2} (\beta_s - \hat{v}\beta_t). \tag{A.14}$$

Under the simplifying setting in Section 4, we actually have an easier and sharper bound than Theorem A.5 as follows.

Lemma A.6. *In the setting of Theorem A.5, assume that $\Sigma_1 = \operatorname{Id}$, β_t is a random vector with i.i.d. entries with mean 0, variance κ^2 and all moments, and β_2 is a random vector such that $\beta_s - \beta_t$ is a random vector with i.i.d. entries with mean 0, variance d^2 and all moments. Denote $\widetilde{\Delta}_\beta := \left((1-\hat{v})^2\kappa^2 + d^2\right) \operatorname{Tr}[\Pi]$. Then we have*

$$te(\hat{\beta}_t^{MTL}) \leqslant te(\hat{\beta}_t^{STL}) \quad \textit{when:} \quad \Delta_{var} \geqslant (\alpha_+^2(\rho_1) + \mathrm{o}(1)) \cdot \widetilde{\Delta}_\beta,$$

$$te(\hat{\beta}_t^{MTL}) \geqslant te(\hat{\beta}_t^{STL}) \quad \textit{when:} \quad \Delta_{var} \leqslant (\alpha_-^2(\rho_1) - \mathrm{o}(1)) \cdot \widetilde{\Delta}_\beta.$$

Now we first give a proof of Theorem A.5 based on Lemma A.3.

Proof of Theorem A.5. [Todo: A proof outline; including the following key lemma.]

The proof is divided into four parts.

Part I: Bounding the bias from model shift. We relate the first term in equation (2.1) to Δ_β.

Proposition A.7. *In the setting of Theorem A.5, 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 $\Sigma_1 = \operatorname{Id}$ 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 (A.16). Recall that $X_1^\top X_1 = \Sigma_1^{1/2} Z_1^\top Z_1 \Sigma_1^{1/2}$. Denote by $\mathcal{E} = Z_1^\top Z_1 - n_1 \operatorname{Id}$. Let We have

$$\delta_1 = \delta_2 + 2\hat{w}^2 n_1 (\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t) + \hat{w}^2 \left\| \Sigma_2^{1/2} K \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} (\beta_s - \hat{w}\beta_t) \right\|^2 \tag{A.15}$$

Here we use the following on the second term in equation (A.15)

$$\begin{aligned}
&\left|(\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t)\right| \\
=&\left|\mathrm{Tr}\left[\mathcal{E} \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t)(\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2}\right]\right| \\
\leqslant&\|\mathcal{E}\| \cdot \left\|\Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t)(\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2}\right\|_\star \\
\leqslant&n_1 \left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1}\right) \cdot \left\|\Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \hat{w}\beta_t)(\beta_s - \hat{w}\beta_t)^\top \Sigma_1^{1/2}\right\|_\star \quad \text{(by equation (A.16))} \\
\leqslant&n_1 \left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1}\right) \left\|\Sigma_1^{1/2} K \Sigma_2 K \Sigma_1^{1/2}\right\| \cdot \left\|\Sigma_1^{1/2}(\beta_s - \hat{w}\beta_t)\right\|^2 \\
&\text{(since the matrix inside is rank 1)}
\end{aligned}$$

The third term in equation (A.15) 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 (A.15) 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 (A.16). □

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

$$-n_1 \left(2\sqrt{\frac{p}{n_1}} - \frac{p}{n_1}\right) \leqslant Z_1^T Z_1 - n_1 \,\mathrm{Id} \leqslant n_1 \left(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1}\right). \tag{A.16}$$

Part III: An ε-net argument. add some arguments with ε-net. □

add some remarks about the difficulty in random matrix theory

By applying a tighter bound to Part I of the above proof, we can conclude the proof Lemma A.6.

Proof of Lemma A.6. the proof for tighter bound .... □

A.2 Proof of Many Tasks with the Same Covariates

Theorem A.8. *Let $n = c \cdot p$. Let $X \in \mathbb{R}^{n \times p}$ and $Y_i = X\beta_i + \varepsilon_i$, for $i = 1, \dots, k$. 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 3.1, we can get a simple expression for the two functions Δ_{vari} and Δ_β. The proof of Theorem A.8 can be found in Appendix A.2.

For this setting, the problem reduces to the following.

$$f(B; W_1, \ldots, W_k) = \sum_{i=1}^{k} \|XBW_i - Y_i\|^2 . \tag{A.17}$$

In order to prove Theorem A.8, we will derive a closed form solution for equation (A.17).

Proof of Theorem A.8. 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 $\hat{\beta}_i^{\text{MTL}}$ as $B^\star U_r U_r(i)$. Now we get that

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

By using Lemma A.2, we conclude the proof. □

A.3 Proof of the Transfer Learning Setting

Proposition A.9. *Let $X_i \in \mathbb{R}^{n_i \times p}$ and $Y_i = X_i \beta_i + \varepsilon_i$, for $i = 1, \cdots, t$, where ε_i are random vectors with i.i.d. entries of mean zero and variance σ^2. Let $\rho_i > 1$, $1 \leqslant i \leqslant t$ be fixed constants. Assume that $\max_{1 \leqslant i \leqslant t} \|\beta_i\| = \mathrm{O}(1)$, $\sigma^2 = \mathrm{O}(1)$ and*

$$\left\| \left[(B_{t-1}^\star)^\top B_{t-1}^\star \right]^{-1} \right\| = \mathrm{O}(1), \quad B_{t-1}^\star := [\beta_1, \beta_2, \ldots, \beta_{t-1}]. \tag{A.18}$$

Suppose that all random variables have subexponential decay. Then we have with high probability,

$$te(BW_t) = \|\varepsilon_\beta + \sigma^2 \varepsilon_{var}^{(1)}\|^2 + \sigma^2 \|\varepsilon_{var}^{(2)}\|^2 + \mathrm{O}(p^{-1/2+\varepsilon}), \tag{A.19}$$

where $B_{t-1}^\star := [\beta_1, \beta_2, \ldots, \beta_{t-1}]$, *and*

$$\begin{aligned}
\varepsilon_\beta &:= \Sigma_k^{1/2} \left\{ 1 - B_{t-1}^\star \left[(B_{t-1}^\star)^\top B_{t-1}^\star \right]^{-1} (B_{t-1}^\star)^\top \right\} \beta_t, \\
\varepsilon_{var}^{(1)} &:= \Sigma_k^{1/2} B_{t-1}^\star \left[(B_{t-1}^\star)^\top B_{t-1}^\star \right]^{-1} \mathcal{M}_1 \left[(B_{t-1}^\star)^\top B_{t-1}^\star + \sigma^2 \mathcal{M}_1 \right]^{-1} (B_{t-1}^\star)^\top \beta_t, \\
\varepsilon_{var}^{(2)} &:= \mathcal{M}_2^{1/2} \left[(B_{t-1}^\star)^\top B_{t-1}^\star + \sigma^2 \mathcal{M}_1 \right]^{-1} (B_{t-1}^\star)^\top \beta_t.
\end{aligned}$$

Here $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *are* $k \times k$ *diagonal matrices with* $(\mathcal{M}_1)_{ii} = \frac{1}{\rho_i - 1} \cdot \frac{1}{p} \operatorname{Tr}\left(\Sigma_i^{-1}\right)$ *and* $(\mathcal{M}_2)_{ii} = \frac{1}{\rho_i - 1} \cdot \frac{1}{p} \operatorname{Tr}\left(\Sigma_i^{-1} \Sigma_k\right)$.

Proof of Proposition A.9. For each task i, $1 \leqslant i \leqslant t-1$, we can find that

$$\hat{\beta}_i = \beta_i + (X_i^\top X_i)^{-1} X_i^\top \varepsilon_i.$$

Then we calculate that

$$\|\hat{\beta}_i\|^2 = \|\beta_i\|^2 + 2\beta_i^\top \cdot (X_i^\top X_i)^{-1} X_i^\top \varepsilon_i + \|(X_i^\top X_i)^{-1} X_i^\top \varepsilon_i\|^2.$$

We can use concentration of random vector with independent entries, Lemma D.6, to get that for any deterministic vector β with $\|\beta\| = \mathrm{O}(1)$,

$$\beta^\top (X_i^\top X_i)^{-1} X_i^\top \varepsilon_i \leqslant p^{\varepsilon/2} \cdot \sigma \left[\beta^\top (X_i^\top X_i)^{-1} \beta \right]^{1/2} \leqslant \sigma p^{-1/2+\varepsilon}, \tag{A.20}$$

with high probability for any constant $\varepsilon > 0$, where we used Lemma D.2 to bound

$$\|(X_i^\top X_i)^{-1}\| \lesssim n_i^{-1} \quad \text{whp.} \tag{A.21}$$

Similarly, using Lemma D.6 we get with high probability,

$$\begin{aligned}
\left| \|(X_i^\top X_i)^{-1} X_i^\top \varepsilon_i\|^2 - \sigma^2 \operatorname{Tr}\left((X_i^\top X_i)^{-1} \right) \right| &\leqslant p^{\varepsilon/2} \cdot \sigma^2 \left\{ \operatorname{Tr}\left[X_i (X_i^\top X_i)^{-2} X_i^\top \right]^2 \right\}^{1/2} \\
&= p^{\varepsilon/2} \cdot \sigma^2 \left[\operatorname{Tr}(X_i^\top X_i)^{-2} \right]^{1/2} \leqslant \sigma^2 p^{-1/2+\varepsilon},
\end{aligned} \tag{A.22}$$

and for $i \neq j$,

$$\begin{aligned}
\left| \varepsilon_j^\top X_j (X_j^\top X_j)^{-1} (X_i^\top X_i)^{-1} X_i^\top \varepsilon_i \right| &\leqslant p^{\varepsilon/2} \cdot \sigma^2 \left\{ \operatorname{Tr}\left[(X_i^\top X_i)^{-1} (X_j^\top X_j)^{-1} \right] \right\}^{1/2} \\
&\leqslant \sigma^2 p^{-1/2+\varepsilon},
\end{aligned} \tag{A.23}$$

With (A.20)-(A.23), we get that with high probability,

$$\begin{aligned}
\|\hat{\beta}_i\|^2 &= \|\beta_i\|^2 + \sigma^2 \cdot \operatorname{Tr}\left((X_i^\top X_i)^{-1} \right) + \mathrm{O}(\sigma p^{-1/2+\varepsilon}) \\
&= \|\beta_i\|^2 + \frac{\sigma^2}{\rho_i - 1} \cdot \frac{1}{p} \operatorname{Tr}\left(\Sigma_i^{-1} \right) + \mathrm{O}(\sigma p^{-1/2+\varepsilon}),
\end{aligned} \tag{A.24}$$

and

$$\hat{\beta}_i^\top \hat{\beta}_j = \beta_i^\top \beta_j + \mathrm{O}(\sigma p^{-1/2+\varepsilon}), \quad i \neq j. \tag{A.25}$$

With (A.24) and (A.25), we obtain that with high probability,

$$B^\top B = (B_{t-1}^\star)^\top B_{t-1}^\star + \sigma^2 \mathcal{M}_1 + \mathrm{O}(\sigma p^{-1/2+\varepsilon}), \tag{A.26}$$

where $\mathrm{O}(\sigma p^{-1/2+\varepsilon})$ means a $(t-1) \times (t-1)$ matrix, say $\mathcal{E}$, satisfying $\|\mathcal{E}\| \leqslant C\sigma p^{-1/2+\varepsilon}$. Notice that by (A.18), we also have with high probability,

$$\|(B^\top B)^{-1}\| \lesssim \left\| \left[(B_{t-1}^\star)^\top B_{t-1}^\star \right]^{-1} \right\| = \mathrm{O}(1).$$

Moreover, using (A.20) we get that

$$B^\top \beta = (B_{t-1}^\star)^\top \beta + \mathrm{O}(\sigma p^{-1/2+\varepsilon}), \tag{A.27}$$

for any deterministic vector β with $\|\beta\| = \mathrm{O}(1)$.

Now we are ready to calculate $te(BW_t)$ using the above concentrations results. For the t-th target model, by optimizing over W_t we get

$$\hat{W}_t = (B^\top X_t^\top X_t B)^{-1} B^\top X_t^\top Y_k = (B^\top X_t^\top X_t B)^{-1} B^\top X_t^\top (X_t \beta_t + \varepsilon_k).$$

We then calculate that

$$\begin{aligned} \mathop{\mathbb{E}}_{\varepsilon_k} \left[\left\| \Sigma_k^{1/2} (B\hat{W}_t - \beta_t) \right\|^2 \right] &= \left\| \Sigma_k^{1/2} \left[\mathrm{Id} - B(B^\top X_t^\top X_t B)^{-1} B^\top X_t^\top X_t \right] \beta_t \right\|^2 \\ &\quad + \sigma^2 \cdot \mathrm{Tr}\left[\Sigma_k (B^\top X_t^\top X_t B)^{-1} \right]. \end{aligned} \tag{A.28}$$

By the concentration of random vector with independent entries or the restricted isometry property for $X_t B$, we have with high probability,

$$B^\top X_t^\top X_t B = n_k \left(B^\top B + \mathrm{O}(p^{-1/2+\varepsilon}) \right), \quad B^\top X_t^\top X_t \beta_t = n_k \left(B^\top \beta_t + \mathrm{O}(p^{-1/2+\varepsilon}) \right).$$

Together with (A.21), (A.26) and (A.27), we can simplify the right-hand side of (A.28) as

$$\begin{aligned} &\mathop{\mathbb{E}}_{\varepsilon_k} \left[\left\| \Sigma_k^{1/2} (B\hat{W}_k - \beta_t) \right\|^2 \right] = \left\| \Sigma_k^{1/2} \beta_t - \Sigma_k^{1/2} B (B^\top B)^{-1} B^\top \beta_t \right\|^2 + \mathrm{O}(p^{-1/2+\varepsilon}) \\ &= \left\| \Sigma_k^{1/2} \beta_t - \Sigma_k^{1/2} B \left[(B_{t-1}^\star)^\top B_{t-1}^\star + \sigma^2 \mathcal{M}_1 \right)^{-1} \right] (B_{t-1}^\star)^\top \beta_t \right\|^2 + \mathrm{O}(p^{-1/2+\varepsilon}). \end{aligned} \tag{A.29}$$

As in (A.26), we can show by concentration that

$$B^\top \Sigma_k B = (B_{t-1}^\star)^\top \Sigma_k B_{t-1}^\star + \sigma^2 \mathcal{M}_2 + \mathrm{O}(\sigma p^{-1/2+\varepsilon}). \tag{A.30}$$

Together with (A.26) and (A.27), we can further simplify (A.29) as

$$\begin{aligned} &\mathop{\mathbb{E}}_{\varepsilon_i, \forall 1 \leqslant i \leqslant k} \left[\left\| \Sigma_k^{1/2} (B\hat{W}_k - \beta_t) \right\|^2 \right] \\ &= \mathrm{O}(p^{-1/2+\varepsilon}) + \left\| \Sigma_k^{1/2} \beta_t \right\|^2 - 2\beta_t^\top \Sigma_k B_{t-1}^\star \left[(B_{t-1}^\star)^\top B_{t-1}^\star + \sigma^2 \mathcal{M}_1 \right]^{-1} (B_{t-1}^\star)^\top \beta_t \\ &+ \beta_t^\top B_{t-1}^\star \left[(B_{t-1}^\star)^\top B_{t-1}^\star + \sigma^2 \mathcal{M}_1 \right]^{-1} \left((B_{t-1}^\star)^\top \Sigma_k B_{t-1}^\star + \sigma^2 \mathcal{M}_2 \right) \left[(B_{t-1}^\star)^\top B_{t-1}^\star + \sigma^2 \mathcal{M}_1 \right]^{-1} (B_{t-1}^\star)^\top \beta_t \\ &= \| \varepsilon_\beta + \sigma^2 \varepsilon_{\text{var}}^{(1)} \|^2 + \sigma^2 \| \varepsilon_{\text{var}}^{(2)} \|^2 + \mathrm{O}(p^{-1/2+\varepsilon}). \end{aligned}$$

This concludes the proof. □

B Supplementary Materials for the Theoretical Implications

B.1 Proofs for Section 4.1

For the ratio $w = w_1/w_2$, we define the function

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

Under the setting of Lemma A.6, again using concentration for random vectors with i.i.d. entries, Lemma D.6, we can obtain that for the validation loss in (A.2),

$$val(\hat{B}; w_1, w_2) = val(w) \left(1 + \mathrm{O}(p^{-1/2+\varepsilon}) \right) \tag{B.1}$$

with high probability for any constant $\varepsilon > 0$. Thus for the following discussions, it suffices focus on the behavior of $val(w)$. Let $\hat{w}$ the minimizer of $val(w)$. The proof will consist of two main steps.

- First, we show that the optimal ratio $\hat{w}$ is close to 1. Then (B.1) gives that $|\hat{v} - \hat{w}| = \mathrm{O}(p^{-1/2+\varepsilon})$ whp.
- Second, we plug $\hat{v}$ back into $te(\hat{\beta}_t^{\text{MTL}})$ and use Lemma A.6 to show the result.

For the first step, we will prove the following result.

Lemma B.1. *We have that the minimizer for $val(w)$ satisfies*

$$|\hat{w} - 1| \leqslant C\left(\frac{d^2}{\kappa^2} + \frac{\sigma^2}{p\kappa^2}\right) \tag{B.2}$$

for some constant $C > 0$.

Proof. First it is easy to observe that $val(w) < val(-w)$ for $w > 0$. Hence it suffices to assume that $w \geqslant 0$.

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

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

Notice that

$$\mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}(X_i^\top X_i)^2\right],\ i = 1, 2, \quad \text{and} \quad \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-1}\right]$$

are increasing functions in w. Hence taking derivative of $val(w)$ with respect to w, we obtain that

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

where the matrix $\mathcal{A}$ is

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

Using the estimate (A.16), we get that $\mathcal{A}$ is lower bounded as

$$\begin{aligned} \mathcal{A} \succeq\ &-\frac{4d^2}{w^5}n_1n_2(\alpha_+(\rho_2) + \mathrm{o}(1))^2 + \left[\frac{2\,(w-1)}{w^3}\kappa^2 - \frac{2d^2}{w^3}\right]n_1^2(\alpha_-(\rho_1) - \mathrm{o}(1))^2 \\ &- 2\frac{\sigma^2}{w^3}n_1(\alpha_+(\rho_1) + \mathrm{o}(1)) \succ 0, \end{aligned}$$

as long as

$$w > w_1 := 1 + \frac{d^2}{\kappa^2} + \frac{\sigma^2}{n_1\kappa^2}\frac{\alpha_+(\rho_1) + \mathrm{o}(1)}{\alpha_-^2(\rho_1)} + \frac{2d^2}{\kappa^2}\frac{\rho_2(\alpha_+^2(\rho_2) + \mathrm{o}(1))}{\rho_1\alpha_-^2(\rho_1)}.$$

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

Then we consider the case $w \leqslant 1$, and the proof is similar as above. Taking derivative of $val(w)$, we obtain that

$$\begin{aligned} val'(w) &\leqslant n_1\left[2\,(w-1)\,\kappa^2\right] \cdot \mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-2}(X_2^\top X_2)^2\right] \\ &\quad + n_2\left[2wd^2 + 2w\,(w-1)\,(2w-1)\kappa^2\right] \cdot \mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ &\quad + n_1(2w\sigma^2) \cdot \mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-2}X_2^\top X_2\right] \\ &= n_1\,\mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-1}\mathcal{B}\right], \end{aligned} \tag{B.3}$$

where the matrix $\mathcal{B}$ is

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

Using the estimate (A.16), we get that $\mathcal{B}$ is upper bounded as

$$\mathcal{B} \preceq -2(1-w)\kappa^2 n_2^2(\alpha_-(\rho_2) - \mathrm{o}(1))^2 + 2wd^2 n_1 n_2(\alpha_+(\rho_1) + \mathrm{o}(1))^2 + 2w\sigma^2 n_2(\alpha_+(\rho_2) + \mathrm{o}(1)) \prec 0,$$

as long as

$$w < w_2 := 1 - \frac{d^2}{\kappa^2}\frac{\rho_1(\alpha_+(\rho_1) + \mathrm{o}(1))^2}{\rho_2 \alpha_-^2(\rho_2)} - \frac{\sigma^2}{n_2\kappa^2}\frac{\alpha_+(\rho_2) + \mathrm{o}(1)}{\alpha_-^2(\rho_2)}.$$

Hence $val'(w) < 0$ on $[0, w_2)$, i.e. $val(w)$ is strictly decreasing for $w < w_2$. Hence we must have $\hat{w} \leqslant w_2$.

In sum, we obtain that $w_2 \leqslant w \leqslant w_1$. Note that under our assumptions, we have

$$\max(|w_1 - 1|, |w_2 - 1|) = \mathrm{O}\left(\frac{d^2}{\kappa^2} + \frac{\sigma^2}{p\kappa^2}\right),$$

which concludes the proof. □

Combining (B.2) and (B.1), we obtain that whp,

$$|\hat{v} - 1| = \mathrm{O}\left(\mathcal{E}\right), \quad \mathcal{E} := \frac{d^2}{\kappa^2} + \frac{\sigma^2}{p\kappa^2} + p^{-1/2+\varepsilon} \tag{B.4}$$

Inserting it into (A.3) and using a similar concentration result as in (B.1), we get whp,

$$\begin{aligned} te(\hat{\beta}_t^{\mathrm{MTL}}) = {} & (1 + \mathrm{O}(\mathcal{E}))\cdot\left[d^2 + \mathrm{O}\left(\mathcal{E}^2\kappa^2\right)\right]\mathrm{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ & + (1 + \mathrm{O}(\mathcal{E}))\cdot\sigma^2\,\mathrm{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right]. \end{aligned}$$

In order to study the phenomenon of bias-variance trade-off, we need the bias term with d^2 and the variance term with σ^2 to be of the same order. With estimate (A.16), we see that

$$\mathrm{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \sim p, \quad \mathrm{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right] \sim \frac{p}{n_1 + n_2}.$$

Hence we need to choose that $p \cdot d^2 \sim \sigma^2$. On the other hand, we want the error term $\mathcal{E}^2\kappa^2$ to be much smaller than d^2, which leads to the condition $p^{-1+2\varepsilon}\kappa^2 \ll d^2 \ll \kappa^2$. The above considerations lead to the following choices of parameters: there exists a constant $c > 0$ such that

$$pd^2 \sim \sigma^2 \sim 1, \quad p^{-1+c}\kappa^2 \leqslant d^2 \leqslant p^c\kappa^2. \tag{B.5}$$

Under this choice, we can write

$$\begin{aligned} te(\hat{\beta}_t^{\mathrm{MTL}}) = {} & (1 + \mathrm{O}(n^{-\varepsilon}))\cdot d^2\,\mathrm{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ & + (1 + \mathrm{O}(n^{-\varepsilon}))\cdot\sigma^2\,\mathrm{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right] \end{aligned} \tag{B.6}$$

whp for some constant $\varepsilon > 0$.

With (B.6) and Lemma A.6, we can prove Proposition 4.1, which gives a transition threshold with respect to the ratio between the model bias and the noise level. With slight abuse of notations, we shall write $\hat{a}_i$, $\hat{b}_k$ and $\hat{M}$ as a_i, b_k and M, respectively, throughout the rest of this section.

Proof of Proposition 4.1. In the setting of Proposition 4.1, we have $M = \Sigma_1^{1/2}\Sigma_2^{-1/2} = \mathrm{Id}$. Then solving equations (A.9) and (A.10) with $\hat{\lambda}_i = 1$, we get that

$$a_1 = \frac{\rho_1(\rho_1 + \rho_2 - 1)}{(\rho_1 + \rho_2)^2}, \quad a_2 = \frac{\rho_2(\rho_1 + \rho_2 - 1)}{(\rho_1 + \rho_2)^2}, \tag{B.7}$$

$$a_3 = \frac{\rho_2}{(\rho_1 + \rho_2)(\rho_1 + \rho_2 - 1)}, \quad a_4 = \frac{\rho_1}{(\rho_1 + \rho_2)(\rho_1 + \rho_2 - 1)}. \tag{B.8}$$

Using Lemma A.2 and Lemma A.3, we can track the reduction of variance from $\hat{\beta}_t^{\text{MTL}}$ to $\hat{\beta}_t^{\text{STL}}$ as

$$\begin{aligned}\delta_{\text{var}} &:= \sigma^2 \operatorname{Tr}\left[(X_2^\top X_2)^{-1}\right] - (1+\mathrm{O}(n^{-\varepsilon}))\cdot\sigma^2 \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right] \\ &= \Delta_{\text{var}}\cdot(1+\mathrm{O}(n^{-\varepsilon}))\end{aligned} \tag{B.9}$$

with high probability, where

$$\Delta_{\text{var}} := \sigma^2\left(\frac{1}{\rho_2-1} - \frac{1}{\rho_1+\rho_2}\cdot\frac{1}{a_1+a_2}\right) = \sigma^2\cdot\frac{\rho_1}{(\rho_2-1)(\rho_1+\rho_2-1)}.$$

Next for the model shift bias

$$\delta_\beta := (1+\mathrm{O}(n^{-\varepsilon}))\cdot d^2 \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right],$$

we can get from Lemma A.6 (or rather the proof of Lemma A.6) that

$$\alpha_-^2(\rho_1) - \mathrm{o}(1) \leqslant \frac{\delta_\beta}{\Delta_\beta} \leqslant \alpha_+^2(\rho_1) + \mathrm{o}(1), \tag{B.10}$$

where

$$\Delta_\beta := pd^2\cdot\frac{\rho_1^2}{(\rho_1+\rho_2)^2}\cdot\frac{1+a_3+a_4}{(a_1+a_2)^2} = pd^2\cdot\frac{\rho_1^2(\rho_1+\rho_2)}{(\rho_1+\rho_2-1)^3}.$$

With (B.9) and (B.10), we conclude that if

$$pd^2\cdot\frac{\rho_1^2(\rho_1+\rho_2)}{(\rho_1+\rho_2-1)^3}\cdot\left(\alpha_+^2(\rho_1)+\mathrm{o}(1)\right) < \sigma^2\cdot\frac{\rho_1}{(\rho_2-1)(\rho_1+\rho_2-1)}, \tag{B.11}$$

we have that $\delta_{\text{var}} > \delta_\beta$, which implies that $te(\hat{\beta}_t^{\text{MTL}})$ is lower than $te(\hat{\beta}_t^{\text{STL}})$. We can simplify (B.11) to

$$d^2 < \frac{\sigma^2}{p}\frac{(\rho_1+\rho_2-1)^2}{\rho_1(\rho_1+\rho_2)(\rho_2-1)}\cdot\left(\left(1+\sqrt{\frac{1}{\rho_1}}\right)^{-4} - \mathrm{o}(1)\right),$$

Plugging into $\rho_1 > 40$, we obtain the first statement of Proposition 4.1.

On the other hand, if

$$pd^2\cdot\frac{\rho_1^2(\rho_1+\rho_2)}{(\rho_1+\rho_2-1)^3}\cdot\left(\alpha_-^2(\rho_1)-\mathrm{o}(1)\right) > \sigma^2\cdot\frac{\rho_1}{(\rho_2-1)(\rho_1+\rho_2-1)}, \tag{B.12}$$

we have that $\delta_{\text{var}} < \delta_\beta$, which implies that $te(\hat{\beta}_t^{\text{MTL}})$ is larger than $te(\hat{\beta}_t^{\text{STL}})$. We can simplify (B.12) to

$$d^2 > \frac{\sigma^2}{p}\frac{(\rho_1+\rho_2-1)^2}{\rho_1(\rho_1+\rho_2)(\rho_2-1)}\cdot\left(\left(1-\sqrt{\frac{1}{\rho_1}}\right)^{-4} + \mathrm{o}(1)\right),$$

Plugging into $\rho_1 > 40$, we obtain the second statement of Proposition 4.1. □

Then we prove Proposition 4.2, which describes the effect of source task data size on the information transfer.

Proof of Proposition 4.2. Recall that under the setting of Example **??**, (B.9) and (B.10) hold. Then we get that $te(\hat{\beta}_t^{\text{MTL}}) < te(\hat{\beta}_t^{\text{STL}})$ whp if

$$\left(pd^2\frac{c_1(c_1+c_2)(c_2-1)}{(c_1+c_2-1)^2} + \mathrm{O}\left(p^{1/2+\varepsilon}d^2\right)\right)\left(1+\sqrt{\frac{1}{c_1}}\right)^4 < \sigma^2\left(1+\mathrm{O}\left(p^{-1/2+\varepsilon}\right)\right); \tag{B.13}$$

otherwise if

$$\left(pd^2\frac{c_1(c_1+c_2)(c_2-1)}{(c_1+c_2-1)^2} + \mathrm{O}\left(p^{1/2+\varepsilon}d^2\right)\right)\left(1-\sqrt{\frac{1}{c_1}}\right)^4 > \sigma^2\left(1+\mathrm{O}\left(p^{-1/2+\varepsilon}\right)\right), \tag{B.14}$$

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

Now we prove the first statement of Proposition 4.2. 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.14) 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.13) holds as long as p is large enough, and hence $te(\hat{\beta}_t^{\mathrm{MTL}}) < te(\hat{\beta}_t^{\mathrm{STL}})$ holds. □

Proof of Proposition 4.3. 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.1 we also have that

$$|\hat{v}-1| = \mathrm{O}(p^{-1}) \tag{B.15}$$

in the setting Proposition 4.3. In fact, Lemma B.1 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.15), Lemma A.3 and Lemma A.6, we get that whp,

$$(te(\hat{\beta}_t^{\mathrm{MTL}}))(M_0) = \frac{\sigma^2}{c_1+c_2}\cdot\frac{1}{p}\operatorname{Tr}\left(\frac{1}{a_1M_0^\top M_0+a_2}\right)+\Delta_\beta(M_0),$$

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 (A.9), 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 (A.10) and (**??**), 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

$$(te(\hat{\beta}_t^{\text{MTL}}))(M_0) = te(M_0) + \mathcal{E}, \tag{B.16}$$

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.16), we conclude (**??**). □

Theoretical justification of Example **??**. In the setting of Example **??**, equations in (A.9) become

$$a_1 + a_2 = 1 - \frac{p}{n_1 + n_2}, \;\; a_1 + \frac{p}{2(n_1 + n_2)} \cdot \left(\frac{a_1}{a_1 + \lambda^2 a_2} + \frac{a_1}{a_1 + \frac{a_2}{\lambda^2}}\right) = \frac{n_1}{n_1 + n_2}. \tag{B.17}$$

It's not hard to verify that there is only one valid solution (a_1, a_2) to (B.17). After solving these, we get the test error for the target task as follows.

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

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.17), where we denote $\gamma = p/(n_1+n_2)$. Then for any λ, the test error in (B.18) 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.17), 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.17)), and hence $te(\lambda) > te(1)$. Otherwise if $n_1 < n_2$, we have $a_1 < (1-\gamma)/2$, and hence $te(\lambda) < te(1)$. □

B.2 Proofs for Section 4.2

With similar techniques, we then show Proposition 4.6, which gives a transition threshold with respect to the difference between the noise levels of the two tasks.

Proof of Proposition 4.6. In the setting of Proposition 4.6, the validation loss and the test error become

$$\begin{aligned} val(\hat{B}; w_1, w_2) = n_1\cdot &\left\|\Sigma_1^{1/2}\left(\frac{w_1^2}{w_2^2}X_1^\top X_1 + X_2^\top X_2\right)^{-1} X_2^\top X_2\left(\beta_s - \frac{w_1}{w_2}\beta_t\right)\right\|^2 \\ &+ n_1\sigma^2\cdot\frac{w_1^2}{w_2^2}\operatorname{Tr}\left[\left(\frac{w_1^2}{w_2^2}X_1^\top X_1 + X_2^\top X_2\right)^{-2}\left(\sigma_1^2\frac{w_1^2}{w_2^2}X_1^\top X_1 + \sigma_2^2 X_2^\top X_2\right)\right] \\ &+ n_2\cdot\frac{w_1^2}{w_2^2}\left\|\Sigma_2^{1/2}\left(\frac{w_1^2}{w_2^2}X_1^\top X_1 + X_2^\top X_2\right)^{-1} X_1^\top X_1\left(\beta_s - \frac{w_1}{w_2}\beta_t\right)\right\|^2 \\ &+ n_2\sigma^2\cdot\operatorname{Tr}\left[\left(\frac{w_1^2}{w_2^2}X_1^\top X_1 + X_2^\top X_2\right)^{-2}\left(\sigma_1^2\frac{w_1^2}{w_2^2}X_1^\top X_1 + \sigma_2^2 X_2^\top X_2\right)\right], \end{aligned}$$

and

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

where $\hat{v} = \hat{w_1}/\hat{w_2}$ is the global minimizer of $val(\hat{B}; w_1, w_2)$. Again using concentration of random vector with i.i.d. entries, Lemma D.6, we can rewrite $te(\hat{\beta}_t^{\text{MTL}})$ as

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

with high probability for any constant $\varepsilon > 0$.

In the current setting, we can also show that (B.4) holds for $\hat{v}$. Since the proof is almost the same as the one for Lemma B.1, we omit the details. Thus under the choice parameters in (B.5), $te(\hat{\beta}_t^{\text{MTL}})$ can be simplified as in (B.6):

$$\begin{aligned} te(\hat{\beta}_t^{\text{MTL}}) = & (1+\mathrm{O}(n^{-\varepsilon}))\cdot d^2\operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ & + (1+\mathrm{O}(n^{-\varepsilon}))\cdot\sigma_2^2\operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right] \\ & + (1+\mathrm{O}(n^{-\varepsilon}))\cdot(\sigma_1^2-\sigma_2^2)\operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2}X_1^\top X_1\right]. \end{aligned} \tag{B.19}$$

Then we write

$$te(\hat{\beta}_t^{\text{STL}}) - te(\hat{\beta}_t^{\text{MTL}}) = \delta_{\text{var}} - \delta_\beta - \delta_{\text{var}}^{(2)},$$

where

$$\delta_{\text{var}} := \sigma_2^2 \operatorname{Tr}\left[(X_2^\top X_2)^{-1}\right] - (1 + \mathrm{O}(n^{-\varepsilon})) \cdot \sigma_2^2 \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right]$$

satisfies (B.9) but with σ^2 replaced with σ_2^2,

$$\delta_\beta := (1 + \mathrm{O}(n^{-\varepsilon})) \cdot d^2 \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right]$$

satisfies (B.10), and

$$\delta_{\text{var}}^{(2)} := (1 + \mathrm{O}(n^{-\varepsilon})) \cdot (\sigma_1^2 - \sigma_2^2) \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2} X_1^\top X_1\right].$$

To estimate this new term, we use the same arguments as in the proof of Lemma A.6: we first replace $X_1^\top X_1$ with $n_1 \operatorname{Id}$ up to some error using (A.16), and then apply Lemma A.4 to calcualte $\operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2}\right]$. This process leads to the following estimates on $\delta_{\text{var}}^{(2)}$:

$$\alpha_-(\rho_1) - \mathrm{o}(1) \leqslant \frac{\delta_{\text{var}}^{(2)}}{\Delta_{\text{var}}^{(2)}} \leqslant \alpha_+(\rho_1) + \mathrm{o}(1), \tag{B.20}$$

where

$$\Delta_{\text{var}}^{(2)} := (\sigma_1^2 - \sigma_2^2) \frac{\rho_1(\rho_1 + \rho_2)}{(\rho_1 + \rho_2 - 1)^3}.$$

Next we compare δ_{var} with $\delta_\beta + \delta_{\text{var}}^{(2)}$. Our main goal is to see how the extra $\delta_{\text{var}}^{(2)}$ affects the information transfer in this case.

Note that the condition $d^2 < \frac{\sigma_2^2}{2p} \cdot \Phi(\rho_1, \rho_2)$ means the we have $\delta_{\text{var}} > \delta_\beta$ by Proposition 4.1. Hence if $\sigma_1^2 \leqslant \sigma_2^2$, then $\delta_{\text{var}}^{(2)} < 0$ and we always have $\delta_{\text{var}} > \delta_\beta + \delta_{\text{var}}^{(2)}$, which gives $te(\hat{\beta}_t^{\text{MTL}}) < te(\hat{\beta}_t^{\text{STL}})$.

It remains to consider the case $\sigma_1^2 \geqslant \sigma_2^2$.

Positive transfer. By (B.9), (B.10) and (B.20) above, if

$$\begin{aligned} &\sigma_2^2 \cdot \frac{\rho_1}{(\rho_2 - 1)(\rho_1 + \rho_2 - 1)} \cdot (1 - \mathrm{o}(1)) \\ &> pd^2 \cdot \frac{\rho_1^2(\rho_1 + \rho_2)}{(\rho_1 + \rho_2 - 1)^3} \left(1 + \sqrt{\frac{1}{\rho_1}}\right)^4 + (\sigma_1^2 - \sigma_2^2) \cdot \frac{\rho_1(\rho_1 + \rho_2)}{(\rho_1 + \rho_2 - 1)^3} \left(1 + \sqrt{\frac{1}{\rho_1}}\right)^2, \end{aligned} \tag{B.21}$$

then we have $\delta_{\text{var}} > \delta_\beta + \delta_{\text{var}}^{(2)}$ whp, which gives $te(\hat{\beta}_t^{\text{MTL}}) < te(\hat{\beta}_t^{\text{STL}})$. We can solve (B.21) to get

$$\sigma_1^2 < -pd^2 \cdot \rho_1 \left(1 + \sqrt{\frac{1}{\rho_1}}\right)^2 \cdot (1 - \mathrm{o}(1)) + \sigma_2^2 \left[1 + \rho_1 \Phi(\rho_1, \rho_2) \left(1 + \sqrt{\frac{1}{\rho_1}}\right)^{-2}\right] \cdot (1 - \mathrm{o}(1)).$$

Plugging into $\rho_1 > 50$, we obtain the first claim of Proposition 4.6 for positive transfer.

Negative transfer. On the other hand, if

$$\begin{aligned} &\sigma_2^2 \cdot \frac{\rho_1}{(\rho_2 - 1)(\rho_1 + \rho_2 - 1)} \cdot (1 + \mathrm{o}(1)) \\ &< pd^2 \cdot \frac{\rho_1^2(\rho_1 + \rho_2)}{(\rho_1 + \rho_2 - 1)^3} \left(1 - \sqrt{\frac{1}{\rho_1}}\right)^4 + (\sigma_1^2 - \sigma_2^2) \cdot \frac{\rho_1(\rho_1 + \rho_2)}{(\rho_1 + \rho_2 - 1)^3} \left(1 - \sqrt{\frac{1}{\rho_1}}\right)^2, \end{aligned} \tag{B.22}$$

then we have $\delta_{\text{var}} < \delta_\beta + \delta_{\text{var}}^{(2)}$ whp, which gives $te(\hat{\beta}_t^{\text{MTL}}) > te(\hat{\beta}_t^{\text{STL}})$. We can solve (B.22) to get

$$\sigma_1^2 > -pd^2 \cdot \rho_1 \left(1 - \sqrt{\frac{1}{\rho_1}}\right)^2 \cdot (1 + \mathrm{o}(1)) + \sigma_2^2 \left[1 + \rho_1 \Phi(\rho_1, \rho_2) \left(1 - \sqrt{\frac{1}{\rho_1}}\right)^{-2}\right] \cdot (1 + \mathrm{o}(1)).$$

Plugging into $\rho_1 > 50$, we obtain the second claim of Proposition 4.6 for negative transfer. □

Then we prove Proposition 4.5, which gives precise upper and lower bounds on the data efficiency ratio.

Proof of Proposition 4.5. Suppose we have reduced number of datapoints—αn_1 for task 1 and αn_2 for task 2 with $n_1 = \rho_1 p$ and $n_2 = \rho_2 p$. In the setting of Proposition 4.5, we still have (B.2). Then using Lemmas A.3 and A.6, and recalling (B.7) and (B.8), we get that whp,

$$te_i(\hat{\beta}(\alpha)) = \sigma^2 \left(\frac{1}{\alpha(\rho_1+\rho_2)-1} + \mathrm{O}\left(p^{-1/2+\varepsilon}\right) \right) + \Delta_\beta^{(i)}, \quad i = 1, 2, \tag{B.23}$$

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

$$\left(1 - \sqrt{\frac{1}{\alpha\rho_i}}\right)^4 \leqslant \Delta_\beta^{(i)} / \left[pd^2 \cdot \frac{(\alpha\rho_i)^2 \cdot \alpha(\rho_1+\rho_2)}{[\alpha(\rho_1+\rho_2)-1]^3} \cdot \left(1 + \mathrm{O}\left(p^{-1/2+\varepsilon}\right)\right) \right] \leqslant \left(1 + \sqrt{\frac{1}{\alpha\rho_i}}\right)^4.$$

On the other hand, using Lemma A.2, we have whp

$$te_i(\hat{\beta}_t^{\mathrm{STL}}) = \frac{\sigma^2}{\rho_i - 1}\left(1 + \mathrm{O}\left(p^{-1/2+\varepsilon}\right)\right), \quad i = 1, 2. \tag{B.24}$$

Comparing (B.23) and (B.24), we immediately obtain a trivial lower bound $\alpha^* \geqslant \alpha_l - \mathrm{o}(1)$, where

$$\alpha_l := \frac{1}{\rho_1+\rho_2}\left[\frac{2(\rho_1-1)(\rho_2-1)}{\rho_1+\rho_2-2} + 1\right] \geqslant \frac{\min(\rho_1,\rho_2)}{\rho_1+\rho_2}.$$

In fact, one can see that if $\alpha \leqslant \alpha_l$, then we have

$$\frac{2\sigma^2}{\alpha(\rho_1+\rho_2)-1} \geqslant \frac{\sigma^2}{\rho_1-1} + \frac{\sigma^2}{\rho_2-1},$$

that is $te_1(\hat{\beta}(\alpha)) + te_2(\hat{\beta}(\alpha))$ is larger than $te_1(\hat{\beta}_t^{\mathrm{STL}}) + te_2(\hat{\beta}_t^{\mathrm{STL}})$ even if we do not take into account the model shift bias terms $\Delta_\beta^{(i)}$. Next we try to get more precise bounds on α^*. In the following discussions, we only consider α such that $\alpha\rho \geqslant \alpha_l\rho \geqslant \min(\rho_1,\rho_2)$.

The upper bound. From (B.23) and (B.24), we see that $\alpha^* \leqslant \alpha$ if α satisfies

$$\begin{aligned} &(1+\mathrm{o}(1)) \cdot \sum_{i=1}^{2} pd^2 \frac{(\alpha\rho_i)^2 \cdot \alpha(\rho_1+\rho_2)}{[\alpha(\rho_1+\rho_2)-1]^3} \left(1 + \sqrt{\frac{1}{\alpha\rho_i}}\right)^4 \\ &\leqslant \frac{\sigma^2}{\rho_1-1} + \frac{\sigma^2}{\rho_2-1} - \frac{2\sigma^2}{\alpha(\rho_1+\rho_2)-1}. \end{aligned} \tag{B.25}$$

For the sum on the left-hand side, we can rewrite it as

$$pd^2 \frac{1}{[1-(\alpha\rho)^{-1}]^3} \sum_{i=1}^{2} \left(\sqrt{\frac{\rho_i}{\rho_1+\rho_2}} + \sqrt{\frac{1}{\alpha\rho}}\right)^4,$$

where $\rho := \rho_1 + \rho_2$.

In order to solve (B.25), we now consider the case $\min(\rho_1,\rho_2) \geqslant 200$. With some basic calculations, one can show that in this case

$$\frac{1}{[1-(\alpha\rho)^{-1}]^3} \sum_{i=1}^{2} \left(\sqrt{\frac{\rho_i}{\rho_1+\rho_2}} + \sqrt{\frac{1}{\alpha\rho}}\right)^4 < \frac{\rho_1^2+\rho_2^2}{(\rho_1+\rho_2)^2} + 0.32.$$

Thus the following inequality implies (B.25):

$$\left(\frac{\rho_1^2+\rho_2^2}{(\rho_1+\rho_2)^2} + 0.32\right) pd^2 < \frac{\sigma^2}{\rho_1-1} + \frac{\sigma^2}{\rho_2-1} - \frac{2\sigma^2}{\alpha(\rho_1+\rho_2)-1}. \tag{B.26}$$

In particular, if

$$\left(\frac{\rho_1^2+\rho_2^2}{(\rho_1+\rho_2)^2} + 0.32\right) pd^2 < \frac{\sigma^2}{\rho_1-1} + \frac{\sigma^2}{\rho_2-1} - \frac{2\sigma^2}{(\rho_1+\rho_2)-1},$$

that is, we have positive transfer when using all the data, then we can solve from (B.26) the following upper bound on α^*:

$$\begin{aligned}\alpha^* &< \frac{1}{\rho_1+\rho_2}\left[\frac{2}{\frac{1}{\rho_1-1}+\frac{1}{\rho_2-1}-\left(\frac{\rho_1^2+\rho_2^2}{(\rho_1+\rho_2)^2}+0.32\right)\frac{pd^2}{\sigma^2}}+1\right]\\ &< \frac{1}{\rho_1+\rho_2}\left[\frac{2}{\frac{1}{\rho_1}+\frac{1}{\rho_2}-\left(\frac{\rho_1^2+\rho_2^2}{(\rho_1+\rho_2)^2}+\frac{1}{3}\right)\frac{pd^2}{\sigma^2}}+1\right].\end{aligned}$$

The lower bound. From (B.23) and (B.24), we see that $\alpha^* \geqslant \alpha$ if α satisfies

$$\begin{aligned}&(1-\mathrm{o}(1))\cdot\sum_{i=1}^{2} pd^2\frac{(\alpha\rho_i)^2\cdot\alpha(\rho_1+\rho_2)}{[\alpha(\rho_1+\rho_2)-1]^3}\left(1-\sqrt{\frac{1}{\alpha\rho_i}}\right)^4\\ &\geqslant \frac{\sigma^2}{\rho_1-1}+\frac{\sigma^2}{\rho_2-1}-\frac{2\sigma^2}{\alpha(\rho_1+\rho_2)-1}.\end{aligned} \tag{B.27}$$

We then follow similar arguments as the above proof for the upper bound.

In order to solve (B.27), we consider the case $\min(\rho_1,\rho_2)\geqslant 200$. With some basic calculations, one can show that the sum on the left-hand side of (B.27) satisfies

$$\frac{1}{[1-(\alpha\rho)^{-1}]^3}\sum_{i=1}^{2}\left(\sqrt{\frac{\rho_i}{\rho_1+\rho_2}}-\sqrt{\frac{1}{\alpha\rho}}\right)^4 > \frac{\rho_1^2+\rho_2^2}{(\rho_1+\rho_2)^2}-0.26.$$

Thus the following inequality implies (B.27):

$$\left(\frac{\rho_1^2+\rho_2^2}{(\rho_1+\rho_2)^2}-0.26\right)pd^2 > \frac{\sigma^2}{\rho_1-1}+\frac{\sigma^2}{\rho_2-1}-\frac{2\sigma^2}{\alpha(\rho_1+\rho_2)-1}. \tag{B.28}$$

There are two cases: if

$$\left(\frac{\rho_1^2+\rho_2^2}{(\rho_1+\rho_2)^2}-0.26\right)pd^2 \geqslant \frac{\sigma^2}{\rho_1-1}+\frac{\sigma^2}{\rho_2-1},$$

then we always have negative transfer for all choice of $0\leqslant\alpha\leqslant 1$; otherwise, we can solve from (B.28) the following lower bound on α^*:

$$\begin{aligned}\alpha^* &> \frac{1}{\rho_1+\rho_2}\left[\frac{2}{\frac{1}{\rho_1-1}+\frac{1}{\rho_2-1}-\left(\frac{\rho_1^2+\rho_2^2}{(\rho_1+\rho_2)^2}-0.26\right)\frac{pd^2}{\sigma^2}}+1\right]\\ &> \frac{1}{\rho_1+\rho_2}\left[\frac{2}{\frac{1}{\rho_1}+\frac{1}{\rho_2}-\left(\frac{\rho_1^2+\rho_2^2}{(\rho_1+\rho_2)^2}-\frac{1}{3}\right)\frac{pd^2}{\sigma^2}}+1\right].\end{aligned}$$

This concludes the proof. □

C Proof of Lemma A.3 and Lemma A.4

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

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

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

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

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

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

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

We assume that Σ_1 and Σ_2 have eigendecompositions

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

where the eigenvalues satisfy that

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

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

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

where the singular values satisfy that

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

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

We summarize our basic assumptions here for future reference.

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

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

C.1 Notations

We will use the following notion of stochastic domination, which was first introduced in [33] and subsequently used in many works on random matrix theory, such as [30, 34, 35, 36, 37, 38]. It simplifies the presentation of the results and their proofs by systematizing statements of the form "ξ is bounded by ζ with high probability up to a small power of n".

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

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

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

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

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

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

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

Lemma C.3 (Lemma 3.2 in [30])**.** *Let ξ and ζ be families of nonnegative random variables.*

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

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

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

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

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

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

Our main goal is to study the following matrix inverse

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

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

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

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

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

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

Now we introduce a convenient self-adjoint linearization trick. This idea dates back at least to Girko, see e.g., the works [39, 40, 41] and references therein. It has been proved to be useful in studying the local laws of random matrices of the Gram type [42, 43, 38, 44]. We define the following $(p+n) \times (p+n)$ self-adjoint block matrix, which is a linear function of X:

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

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

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

For simplicity of notations, we define the index sets

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

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

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

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

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

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

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

Next we introduce the spectral decomposition of G. Let

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

be a singular value decomposition of W, where

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Moreover, the solution satisfies

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

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

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

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

Suppose they satisfy the system of equations

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

for some (random) errors satisfying

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which imply that

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

Moreover, we trivially have

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

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

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

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

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

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

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

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

Using Corollary C.8, we can complete the proof of Lemma A.3 and Lemma A.4.

Proof of Lemma A.3. In the setting of Lemma A.3, we can write

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

where $\widetilde{\Sigma}_1 := w^2\Sigma_1$, Σ_2, Z_1 and Z_2 satisfy Assumption C.1. Here the extra n^{-1} is due to the choice of the variances—in the setting of Lemma A.3 the variances of the entries of $Z_{1,2}$ are equal to 1, while in (C.1) they are taken to be n^{-1}. As in (C.6), we assume that $M := \widetilde{\Sigma}_1^{1/2}\Sigma_2^{-1/2}$ has singular value decomposition

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

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

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

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

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

where by (C.18),

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

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

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

with probability $1 - \mathrm{o}(1)$. This concludes (A.6) if we rename $r_1 b_2 \to a_1$ and $r_2 b_3 \to a_2$. For (A.6), 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 (A.9). This gives (**??**) by (A.6). □

Proof of Lemma A.4. In the setting of Lemma A.4, we can write

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where we used $M^\top M = V\Lambda^2 V^\top$ in the second step. This concludes Lemma A.4. □

D Proof of Theorem C.7

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

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

In this case, the problem is reduced to proving the anisotropic local law for G with $U = \mathrm{Id}$ and $V = \mathrm{Id}$, such that the entries of Y_1 and Y_2 are independent. This can be handled using the standard resolvent methods as in e.g. [30, 45, 46]. To go from the Gaussian case to the general X case, we will adopt a continuous self-consistent comparison argument developed in [38].

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

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

and prove the following result.

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

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

D.1 Basic estimates

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

First, note that $Z_1^\top Z_1$ (resp. $Z_2^\top Z_2$) is a standard sample covariance matrix, and it is well-known that its nonzero eigenvalues are all inside the support of the Marchenko-Pastur law $[(1-\sqrt{d_1})^2, (1+\sqrt{d_1})^2]$ (resp. $[(1-\sqrt{d_2})^2, (1+\sqrt{d_2})^2]$) with probability $1 - \mathrm{o}(1)$ [47]. In our proof, we shall need a slightly stronger probability bound, which is given by the following lemma. Denote the nonzero eigenvalues of $Z_1^\top Z_1$ and $Z_2^\top Z_2$ by $\lambda_1(Z_1^\top Z_1) \geqslant \cdots \geqslant \lambda_{p\wedge n_1}(Z_1^\top Z_1)$ and $\lambda_1(Z_2^\top Z_2) \geqslant \cdots \geqslant \lambda_p(Z_2^\top Z_2)$.

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

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

and

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

Proof. This lemma essentially follows from [30, Theorem 2.10], although the authors considered the case with $q \prec n^{-1/2}$ only. The results for larger q follows from [48, Lemma 3.12], but only the bounds for the largest eigenvalues are given there in order to avoid the issue with the smallest eigenvalue when d_2 is close to 1. However, under the assumption (C.3), the lower bound for the smallest eigenvalue follows from the exactly the same arguments as in [48]. Hence we omit the details. □

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

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

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

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

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

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

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

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

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

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

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

Definition D.4 (Minors)**.** *For any $(p+n)\times(p+n)$ matrix $\mathcal{A}$ and $\mathbb{T}\subseteq\mathcal{I}$, we define the minor $\mathcal{A}^{(\mathbb{T})}:=(\mathcal{A}_{\mathfrak{a}\mathfrak{b}}:\mathfrak{a},\mathfrak{b}\in\mathcal{I}\setminus\mathbb{T})$ 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* (C.13)*)*

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

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

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

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

Then we record the following resolvent identities.

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

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

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

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

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

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

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

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

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

Proof. All these identities can be proved directly using Schur's complement formula. The reader can also refer to, for example, [38, Lemma 4.4]. □

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

Lemma D.6 (Lemma 3.8 of [49])**.** *Let (x_i), (y_j) be independent families of centered and independent random variables, and (A_i), (B_{ij}) be families of deterministic complex numbers. Suppose the entries x_i, y_j have variance at most n^{-1} and satisfy the bounded support condition (C.8) with $q\leqslant n^{-\phi}$ for some constant $\phi>0$. Then we have the following bound:*

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

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

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

D.2 Entrywise local law

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

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

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

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

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

Step 1: Large deviations estimates. In this step, we prove some (almost) optimal large deviations estimates on the off-diagonal entries of G, and on the following Z variables. In analogy to [49, Section 3] and [38, Section 5], we introduce the Z variables

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

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

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

and

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

For simplicity, we introduce the following random error

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

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

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

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

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

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

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

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

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

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

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

and

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

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

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

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

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

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

We define the following z-dependent event

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

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

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

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

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

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

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

Then we claim the following result.

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

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

Proof. By (D.8), (D.16) and (D.17), we obtain that for $i \in \mathcal{I}_1$, $\mu \in \mathcal{I}_2$ and $\nu \in \mathcal{I}_3$,

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

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

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

where we recall (D.7), and

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

By (D.11) we can bound that

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

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

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

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

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

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

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

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

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

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

As a byproduct, we obtain from the two estimates that

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

Together with (D.22), we get that

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

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

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

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

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

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

which further implies

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

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

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

First we claim that it suffices to show that

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

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

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

and

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

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

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

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

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

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

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

Proof. By (C.15), we get

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof of Proposition D.1. With (D.15), one can repeat the polynomialization method in [30, Section 5] to get the anisotropic local law (C.26) for G_0. The proof is exactly the same, except for some minor notation difference, so we omit the details. □

D.3 Anisotropic local law

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

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

for any deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}}$ and $z \in \mathbf{D}$. Here we abbreviated $Z := \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ and $Z^{Gauss} := \begin{pmatrix} Z_1^{Gauss} \\ Z_2^{Gauss} \end{pmatrix}$. We prove the above statement using a continuous comparison argument introduced in [38]. The proof is similar to the ones in Sections 7-8 of [38], so we only give an outline without writing down all the details.

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

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

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

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

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

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

We also introduce the matrices

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

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

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

provided all the expectations exist.

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

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

Here for simplicity of notations, we introduce the following notation of generalized entries: for $\mathbf{u},\mathbf{v} \in \mathbb{R}^{\mathcal{I}}$, we shall denote $G_{\mathbf{u}\mathbf{v}} := \mathbf{u}^\top G \mathbf{v}$. 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 (D.48) for any fixed $p \in 2\mathbb{N}$ and constant $\varepsilon > 0$:

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

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

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

Together with Markov's inequality, we conclude (C.26). In order to prove (D.50), we compare $Z^{\theta,Z^0_{\mu i}}_{(\mu i)}$ and $Z^{\theta,Z^1_{\mu i}}_{(\mu i)}$ via a common $Z^{\theta,0}_{(\mu i)}$, i.e. we will prove that

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

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

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

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

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

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

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

with high probability.

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

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

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

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

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

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

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

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

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

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

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

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

Lemma D.12. *Suppose that*

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

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

Proof. The proof is the same as the one for [38, Lemma 7.16]. □

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

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

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

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

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

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

Then we can calculate the derivative

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\frac{1}{6} \mathbb{E}\, f_{(\mu i)}^{(3)}(0)\, \mathbb{E} \left(Z_{i\mu}^a \right)^3.$$

Note that $\mathbb{E} \left(Z_{i\mu}^a \right)^3$ is of order $n^{-3/2}$, while the sum over i and μ in (D.51) provides a factor n^2. In fact, $\mathbb{E}\, f_{(\mu i)}^{(3)}(0)$ 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 [38, Section 8]. The argument for our case is almost the same except for some notational differences, so we omit the details.

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

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

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

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

where

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

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

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

which implies

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

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

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

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

With direct calculation, one can get the derivative

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

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

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

and

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

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

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

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

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

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

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

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

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

With the contraction mapping, we have the bound

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

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

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

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

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

We can calculate that

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

At $z = 0$, we have

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

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

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

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

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