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

Anonymous Author(s)
Affiliation
Address
email

Abstract

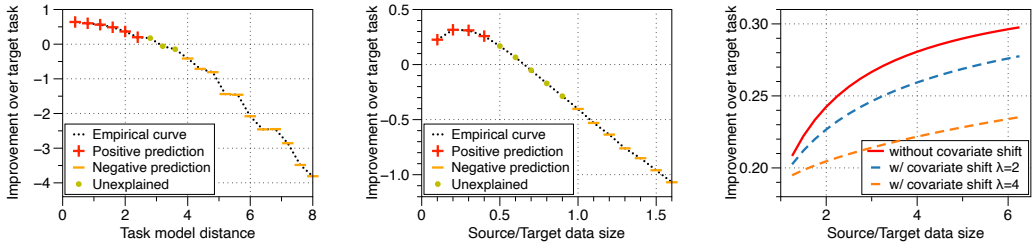
Multi-task learning is a powerful approach to solve complex tasks in many applications such as image and text classification. Yet, there is little rigorous understanding of when multi-task learning performs well, even compared to single-task learning. In this work, we study this question formally by considering three components including *task similarity*, *sample size* and *covariate shift* in the setting of high-dimensional linear regression. First, we show that whether multi-task learning performs better than single-task learning is determined by the bias-variance tradeoff of multi-task estimators. We analyze the bias-variance tradeoff by developing technical tools using random matrix theory. Second, we apply our tools to show that the performance of multi-task learning is negatively affected as a result of: (a) decreased task similarity; (b) increased source sample size; (c) covariate shift under increased source sample sizes. Finally, we validate the three components of our theory on image and text classification tasks. Inspired by our theory, we propose an incremental training scheduler for improving the efficiency of multi-task training for predicting a particular task.

1 Introduction

Multi-task learning represents a powerful paradigm to solve complex tasks in computer vision [1, 2], natural language processing [3, 4] and many other areas [5]. In many settings, multiple information sources are available to help solve a particular task. The performance of multi-task learning depends on the relationship between the information sources and the task [6]. When the information sources are heterogeneous, negative transfer–where multi-task learning (MTL) performs worse than single-task learning (STL)–has often been observed [7, 8, 9]. While many empirical approaches are proposed to mitigate negative transfer [5], a precise understanding of when negative transfer occurs remains elusive in the literature [10]. In this work, we show that negative transfer is determined by the bias-variance tradeoff of multi-task learning estimators for the high-dimensional linear regression setting. We develop new tools to analyze the tradeoff and explain negative transfer precisely.

Understanding 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 so that the error reduces as the data sizes of all tasks increase. For example, the sample sizes of all tasks are often assumed to be equal [11, 14, 15, 16]. On the other hand, uneven sample sizes (or dominating tasks) have been empirically observed to cause of negative transfer [17]. When all tasks are sufficiently similar, adding more labeled data improves the generalization performance for predicting a particular task [18]. To motivate our study, Figure 1 shows three examples for two linear regression tasks solved using a shared linear layer for both tasks and an output layer for each task. We observe that negative transfer occurs as (a) *task similarity*: tasks become more different; (b) *data size*: source/target data size increases. Furthermore, MTL performance is negatively affected when (c) *covariate shift*: the

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



(a) Task similarity (b) Sample size (c) Covariate shift

Figure 1: We observe that MTL performance (a) transitions from positive to negative transfer as *task model distance* increases (Prop. 3.1); (b) transitions from positive to negative transfer as source/target sample size increases (Prop. 3.2); (c) gets worse as *covariate shift* gets more severe (increasing λ) and source/target *sample size* increases (Prop. 3.4). The y-axis measures the loss of STL minus MTL.

covariance matrices of the two tasks become more different. The observations highlight the need to develop generalization bounds that scale tightly with properties of multiple task data.

Our main contribution is to develop technical tools to explain the above phenomena. We focus on predicting a particular task given multiple high-dimensional linear regression tasks [19, 20].

- **Task similarity:** We assume that each task follows a linear model with parameters $\beta_i \in \mathbb{R}^p$, for $1 \leqslant i \leqslant t$. Hence, task similarity can be measured by the distance between the model parameters.
- **Sample size:** Let $n_i = \rho_i \cdot p$ denote the sample size of task i, for a fixed value $\rho_i > 1$ that does not grow with p, for any $1 \leqslant i \leqslant p$. Hence, for a source task and a target task, their data ratio is measured by $n_1/n_2 = \rho_1/\rho_2$. Importantly, we assume that for the t-th (target) task, ρ_t is a small constant (say 2), to capture the need for adding more labeled data.
- **Covariate shift:** We assume that for every task $1 \leqslant i \leqslant t$, its features are random vectors with covariance matrix $\Sigma_i \in \mathbb{R}^{p\times p}$. Given two tasks i and j, we measure covariate shift by $\Sigma_i^{1/2}\Sigma_j^{-1/2}$.

We focus on the hard parameter sharing architecture with a linear shared layer for all tasks and a separate prediction head for each task [10, 21, 18]. Let $\hat{\beta}_t^{\text{MTL}}$ denote the optimal multi-task estimator for the target task, which is defined precisely in Section 2.1. We revisit the bias-variance tradeoff of the multi-task estimator. Interestingly, we observe that the variance of the multi-task estimator is always smaller than the variance of the single-task estimator, because of an increased sample size. The bias of the multi-task estimator results in a negative effect caused by the difference between β_t and the rest of $\{\beta_i\}_{i=1}^{t-1}$. Hence, the bias-variance tradeoff determines whether we observe positive or negative transfer. For the setting of two tasks, we show how the variance of the multi-task estimator scales with sample size and covariate shift in the following result.

Theorem 1.1 (Informal)**.** *In the setting of two tasks, the variance of the multi-task estimator $\hat{\beta}_t^{MTL}$ is equal to the following (times noise variance)*

$$\frac{1}{n_1+n_2}\cdot \operatorname{Tr}\left[(a_1\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} + a_2\,\mathrm{Id})^{-1}\right] + \mathrm{O}\left(p^{-1/2+o(1)}\right).$$

where a_1, a_2 are both fixed values that roughly scales with the sample sizes ρ_1, ρ_2, and satisfy $a_1 + a_2 = 1 - (\rho_1+\rho_2)^{-1}$ plus another deterministic equation. A similar result on the bias of $\hat{\beta}_t^{MTL}$ that scales with task similarity in addition to sample size and covariate shift holds.

Theorem 1.1 allows us to analyze of the bias-variance tradeoff of the multi-task estimator for two settings: (i) two tasks with arbitrary covariate shift; (ii) many tasks with no covariate shift.

Our next contribution is to use our newly developed tool to explain negative transfer in multi-task learning precisely. In Section 3, we explain the three phenomena in Figure 1 for a simplified isotropic setting of two tasks. It is crucial that the concentration error of Theorem 1.1 is small so that we can explain the transition phenomena in Figure 1a and 1b. The unexplained observations are caused by an error term that arises from the bias of $\hat{\beta}_t^{\text{MTL}}$ – we discuss these in Section 3 more precisely. Theorem 1.1 allows us to compare MTL performance under different covariate shift. When task models are the same and the covariate shift matrix belongs to a certain bounded set, we show that having no covariate shift yields the optimal MTL performance. Finally, we analyze the benefit of MTL for reducing the amount of labeled data needed to achieve comparable performance to STL, which is a key empirical finding of Taskonomy.

Our last contribution is to connect our theory to practical problems of interest. (i) We validate the three components of our theory in Section 3 and measure the data efficiency of multi-task learning on the sentiment analysis dataset. (ii) We provide a single-task based metric to predict positive or negative transfer in multi-task learning. While it is not well understood when multi-task learning provides positive transfer, we show that the STL results can help indicate and understand MTL results on ChestX-ray14 [1] and sentiment analysis datasets [23]. (iii) We design an incremental training schedule to improve the efficiency of multi-task training for predicting a particular task. We show that our training schedule reduces the computational cost by 55% compared to baseline multi-task training on the sentiment analysis dataset, while keeping the accuracy the same.

2 Multi-Task vs. Single-Task Learning: The Bias-Variance Tradeoff

We begin by describing our problem setup more formally. Then, we describe our main result for the bias-variance tradeoff of multi-task and transfer learning estimators in our setting.

2.1 Problem Setup

Recall that we have t labeled training datasets, 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. We consider the following linear multi-task learning architecture.

$$f(B; W_1, \ldots, W_t) = \sum_{i=1}^{t} \|X_i B W_i - Y_i\|^2, \tag{2.1}$$

where $B \in \mathbb{R}^{p \times r}$ and $W_i \in \mathbb{R}^r$. Here B provides a shared subspace for all tasks and W_i fits B suitably to each task. Following [18], we assume that $r < t$, because otherwise minimizing $f(\cdot)$ could result in BW_i to be the single-task optimum. Hence, applying equation (2.1) results in the same estimator as single-task learning. We define $\hat{\beta}_t^{\text{MTL}}$ by two steps: (i) minimizing $f(\cdot)$ over B; (ii) minimize $\{W_i\}_{i=1}^k$ over an independent sample of the training set, e.g. $O(p^{0.99})$ suffices. For more details, we refer the reader to Appendix A. For an estimator $\hat{\beta} \in \mathbb{R}^p$, we define the out-of-sample prediction loss as

$$L_t(\hat{\beta}) = \mathop{\mathbb{E}}_{x}\left[(x^\top \hat{\beta} - x^\top \beta_t)^2\right] = \left\|\mathbb{E}\left[\hat{\beta}\right] - \beta_t\right\|^2 + \mathbb{E}\left[\left\|\hat{\beta} - \mathbb{E}\left[\hat{\beta}\right]\right\|^2\right],$$

which can be further decomposed as the bias and variance of $\hat{\beta}$. The single-task estimator $\hat{\beta}_t^{\text{STL}}$ is given by $(X_t^\top X_t)^{-1} X_t^\top Y_t$. 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 small compared to $\{\rho_i\}_{i=1}^{t-1}$. This setting captures the need for adding more labeled data to reduce the prediction loss of the target task. A well-known result for this setting states that $L_t(\hat{\beta}_t^{\text{STL}}) = \sigma^2 \cdot \text{Tr}[(X_t^\top X_t)^{-1}\Sigma_t]$ is concentrated around $\frac{\sigma^2}{\rho_t - 1}$ (e.g. Chapter 6 of [22]), which scales with the sample size and noise level of the target task. However, this result only applies to a single task. Therefore, our goal is to extend this result to multiple tasks.

Notations. When there is no ambiguity, we drop the subscript t from $L_t(\hat{\beta}_t^{\text{MTL}})$ to $L(\hat{\beta}_t^{\text{MTL}})$ for simplicity. We refer to the first task as the source task when there are only two tasks. We say there is negative transfer if the prediction loss of $\hat{\beta}_t^{\text{MTL}}$ is larger than that of $\hat{\beta}_t^{\text{STL}}$, or positive transfer otherwise. For a matrix $X \in \mathbb{R}^{p_1 \times p_2}$, let $\lambda_{\min}(X)$ denote its minimum singular value. Let $\|X\|$ denote the spectral norm of X.

2.2 Analyzing the Tradeoff via Random Matrix Theory

To illustrate our intuition, we begin by considering the setting of two tasks with general covariance matrices. We decompose the test error of $\hat{\beta}_t^{\text{MTL}}$ on the target task into two parts as follows.

$$L(\hat{\beta}_t^{\text{MTL}}) = \hat{v}^2 \left\|\Sigma_2^{1/2}(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 (\beta_1 - \hat{v}\beta_2)\right\|^2 \tag{2.2}$$

$$+ \sigma^2 \cdot \text{Tr}\left[(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_2\right], \tag{2.3}$$

Here $\hat{v} \in \mathbb{R}$ denotes a fixed value that depends on the output layer weights W_1, W_2. The role of $\hat{v}$ is to scale the shared subspace B to fit each task. These are derived in Appendix A.

Equation (2.2) corresponds to the bias of $\hat{\beta}_t^{\text{MTL}}$, which captures how similar β_1 and β_2 are. Hence, the bias of $\hat{\beta}_t^{\text{MTL}}$ introduces a negative effect from adding the source labels. Equation (2.3) corresponds to the variance of $\hat{\beta}_t^{\text{MTL}}$, which is always smaller than the variance of $\hat{\beta}_t^{\text{STL}}$. Hence, whether $L(\hat{\beta}_t^{\text{MTL}}) < L(\hat{\beta}_t^{\text{STL}})$ is determined precisely by the tradeoff between the negative effect of the bias term and the positive effect of the variance term!

We provide a sharp analysis of the bias-variance tradeoff two tasks with general covariance matrices. We state our result for two tasks as follows.

Corollary 2.1. *For the setting of two tasks, let $M = \Sigma_1^{1/2}\Sigma_2^{-1/2}$, $\delta > 0$ be a desired error margin and $\rho_1 \gtrsim \frac{1}{\delta^2} \cdot \lambda_{\min}(M)^{-4}\|\Sigma_1\| \max(\|\beta_1\|^2, \|\beta_2\|^2)$. Let $\rho_2 > 1$ be a fixed value. There exists two deterministic functions Δ_{bias} and Δ_{var} that only depend on $\{\hat{v}, \Sigma_1, \Sigma_2, \rho_1, \rho_2, \beta_1, \beta_2\}$ such that*

- *If $\Delta_{bias} - \Delta_{var} < -\delta$, then w.h.p. over the randomness of X_1, X_2, we have $L(\hat{\beta}_t^{MTL}) < L(\hat{\beta}_t^{STL})$.*
- *If $\Delta_{bias} - \Delta_{var} > \delta$, then w.h.p. over the randomness of X_1, X_2, we have $L(\hat{\beta}_t^{MTL}) > L(\hat{\beta}_t^{STL})$.*

Corollary 2.1 applies to settings where large amounts of source task data is available but the target sample size is small. For such settings, we obtain a sharp transition from positive transfer to negative transfer determined by $\Delta_{\text{bias}} - \Delta_{\text{var}}$. While the general form of the threshold can be complex (as is preivous generalization bounds for MTL), they admit interpretable forms for simplified settings. This will be the focus of Section 3.

Proof Overview. We first describe the proof of Theorem 1.1. We use the Stieltjes transform method (or the resolvent method) in random matrix theory [25, 26, 27]. Roughly speaking, we study the resolvent $R(z) := [\Sigma_2^{-1/2}(X_1^\top X_1 + X_2^\top X_2)\Sigma_2^{-1/2} - z]^{-1}$ for $z \in \mathbb{C}$ around $z = 0$. Using the methods in [28, 29], we find the asymptotic limit, say $R_\infty(z)$, of $R(z)$ for any z as $p \to \infty$ with an almost optimal convergence rate. In particular, when $z = 0$, $\text{Tr}[R_\infty(0)]$ gives the right hand side of (**??**), which concludes Theorem 1.1. The details can be found in Appendix C and D.

Using Theorem 1.1 over (2.3), we can calculate the amount of reduced variance compared to STL. The amount of reduced variance is given by Δ_{var}. For the bias term of equation (2.2), we apply a Gaussian concentration bound on $X_1^\top X_1$, whose expectation is $n_1^2\Sigma_1$. This results in the error term δ, which scales as $(1 \pm 1/\sqrt{\rho_1})^4$. Then, we applying a similar identity to Theorem 1.1 for bounding the bias term, noting that the derivative of $R(z)$ with respect to z can be approximated by $R'_\infty(z)$. This leads to a negative effect given by Δ_{bias}. The details are presented in Appendix A.2.

Next, we describe our result for more than two tasks with same features, i.e. $X_i = X$ for any i. This setting is prevalent in applications of multi-task learning to image classification, where there are multiple prediction labels/tasks for every image [1, 24].

Theorem 2.2. *Consider t data models $Y_i = X\beta_i + \varepsilon_i$, $i = 1, 2, \cdots, t$, that satisfy Assumption A.2. Let $U_r U_r^\top$ denote the best rank-r subspace approximation of $(B^\star)^\top \Sigma B^\star$, where $B^\star := [\beta_1, \beta_2, \ldots, \beta_t]$ and $U_r \in \mathbb{R}^{t\times r}$. Suppose $(B^\star)^\top \Sigma B^\star$ is of full rank in the sense that $\lambda_{\min}((B^\star)^\top \Sigma B^\star) \gtrsim \sigma^2$. Let $U_r(i)$ denote the i-th row vector of U_r. We have the following*

- *We have $L(\hat{\beta}_t^{MTL}) < L(\hat{\beta}_t^{STL})$ with high probability if*

$$\left(1 - \|U_r(t)\|^2\right) \frac{\sigma^2}{\rho - 1} > \|\Sigma(B^\star U_r U_r(t) - \beta_t)\|^2 + \mathrm{o}\left(\|B^\star\|^2 + \sigma^2\right).$$

- *We have $L(\hat{\beta}_t^{MTL}) > L(\hat{\beta}_t^{STL})$ with high probability if*

$$\left(1 - \|U_r(t)\|^2\right) \frac{\sigma^2}{\rho - 1} < \|\Sigma(B^\star U_r U_r(t) - \beta_t)\|^2 - \mathrm{o}\left(\|B^\star\|^2 + \sigma^2\right).$$

Remark. Theorem 1.1 extends a well-known result for the single-task setting when X_1, ρ_1, a_1 are all equal to zero [22]. The concentration error of our result that is order $O(p^{-1/2+o(1)})$ is nearly optimal.

3 Explaining Negative Transfer in an Isotropic Setting

We provide precise explanations to the phenomena of negative transfer in multi-task learning. We explain from three perspectives, including *task similarity*, *sample size* and *covariate shift*. We show

how negative transfer occurs by varying task similarity or sample size. Then we show that when source task sample size becomes large, covariate shift causes more negative effects.

3.1 Task Similarity

It is well-known since the seminal work of Caruana [6] that how well multi-task learning performs depends on task relatedness. We formalize this connection in the following simplified setting.

The isotropic model. Consider two tasks with isotropic covariances $\Sigma_1 = \Sigma_2 = \mathrm{Id}$. Recall that each task has sample 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 $i = 1, 2$. 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 Appendix A.

In the isotropic model, we show that as we increase the distance between β_1 and β_2, there is a transition from positive transfer to negative transfer in MTL. Our result below will provide an explanation to this phenomenon. We introduce the following notations.

$$\Psi(\beta_1, \beta_2) = \mathbb{E}\left[\|\beta_1 - \beta_2\|^2\right]/\sigma^2, \quad \Phi(\rho_1, \rho_2) = \frac{(\rho_1 + \rho_2 - 1)^2}{\rho_1(\rho_1 + \rho_2)(\rho_2 - 1)}.$$

Proposition 3.1 (Task model distance)**.** *In the isotropic model, suppose that $\rho_1 > 1$ and $\sigma_1 = \sigma_2 = \sigma$. Then we have that*

- *If $\Psi(\beta_1, \beta_2) < \gamma_+^{-1}\Phi(\rho_1, \rho_2)$, then w.h.p. over the randomness of X_1, X_2, $L(\hat{\beta}_t^{MTL}) < L(\hat{\beta}_t^{STL})$.*
- *If $\Psi(\beta_1, \beta_2) \geqslant \gamma_-^{-1}\Phi(\rho_1, \rho_2)$, then w.h.p. over the randomness of X_1, X_2, $L(\hat{\beta}_t^{MTL}) \geqslant L(\hat{\beta}_t^{STL})$.*

Here $\gamma_- = (1 - \mathrm{o}(1)) \cdot (1 - \rho_1^{-1/2})^4$ and $\gamma_+ = (1 + \mathrm{o}(1)) \cdot (1 + \rho_1^{-1/2})^4$. Concretely, if $\rho_1 > 40$, then $\gamma_- \in (1, 2)$ and $\gamma_+ \in (1/2, 1)$.

Proposition 3.1 simplifies Theorem 2.1 in the isotropic model, allowing for a more explicit statement of the bias-variance tradeoff. Concretely, $\Psi(\beta_1, \beta)$ and $\Phi(\rho_1, \rho_2)$ corresponds to Δ_{bias} and Δ_{var}, respectively. Prosposition 3.1 explains the transition observed in Figure 1a. Note that there are several unexplained observations near the transition threshold 0, which are caused by the concentration errors γ_+, γ_-. The proof of Proposition 3.1 can be found in Appendix B.1.

Algorithmic consequence. We can in fact extend the result to the cases where the noise variance is different. In this case, we will see that MTL is particularly effective. Concretely, suppose the noise variance σ_1^2 of task 1 differs from the noise variance σ_2^2 of task 2. If σ_1^2 is too large, the source task provides a negative transfer to the target. If σ_1^2 is small, the source task is more helpful. We leave the result to the Appendix in Proposition B.3. Inspired by the observation, we propose a single-task based metric to help understand MTL results using STL results.

- For each task, we train a single-task model. Let z_s and z_t be the prediction accuracy of each task, respectively. Let $\tau \in (0, 1)$ be a fixed threshold.
- If $z_s - z_t > \tau$, then we predict that there will be positive transfer when combining the two tasks using MTL. If $z_s - z_t < -\tau$, then we predict negative transfer.

3.2 Sample Size

In classical Rademacher or VC based theory of multi-task learning, the generalization bounds are usually presented for settings where the sample sizes are equal for all tasks [11, 13, 16]. 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 sample size. Our result below explains the phenomenon of Figure 1b.

Proposition 3.2 (Source/target sample size)**.** *In the isotropic model, assume that $\rho_1 > 40$ and $\rho_2 > 110$ are fixed constants, $\Psi(\beta_1, \beta_2) > 2/(\rho_2 - 1)$. We have that*

- *If $\frac{n_1}{n_2} = \frac{\rho_1}{\rho_2} < \frac{1}{\gamma_+} \cdot \frac{1 - 2\rho_2^{-1}}{\Psi(\beta_1, \beta_2)(\rho_2 - 1) - \gamma_+^{-1}}$, then w.h.p. $L(\hat{\beta}_t^{MTL}) < L(\hat{\beta}_t^{STL})$.*

- *If* $\frac{n_1}{n_2} = \frac{\rho_1}{\rho_2} > \frac{1}{\gamma_-} \cdot \frac{1-2\rho_2^{-1}}{\Psi(\beta_1,\beta_2)(\rho_2-1.5)-\gamma_-^{-1}}$*, then w.h.p.* $L(\hat{\beta}_t^{MTL}) > L(\hat{\beta}_t^{STL})$.

Proposition 3.2 describes the bias-variance tradeoff in terms of the data ratio n_1/n_2. From Figure 1b, we see that our result is able to explain the transition from positive to negative transfer. There are several unexplained observations near $y = 0$ caused by the errors γ_-, γ_+. The proof of Proposition 3.2 can be found in Appendix B.2.

Connection to Taskonomy. We use our tools to explain a key result of Taskonomy by Zamir et al.'18 [2], which shows that MTL can reduce the amount of labeled data needed to achieve comparable performance to STL. For $i = 1, 2$, let $\hat{\beta}_i^{\text{MTL}}(x)$ denote the estimator trained using $R \cdot n_i$ datapoints from every task. The data efficiency ratio roughly scales as

$$\arg\min_{x\in(0,1)} \; L_1(\hat{\beta}_1^{\text{MTL}}(x)) + L_2(\hat{\beta}_2^{\text{MTL}}(x)) \leqslant L_1(\hat{\beta}_1^{\text{STL}}) + L_2(\hat{\beta}_2^{\text{STL}}).$$

For example, the data efficiency ratio is 1 if there is negative transfer. Using our tools, we show that in the isotropic model, the data efficiency ratio is equal to

$$\frac{1}{\rho_1+\rho_2} + \frac{1}{(\rho_1+\rho_2)(\rho_1^{-1}+\rho_2^{-1}-\Theta(\Psi(\beta_1,\beta_2)))}.$$

Compared with Proposition 3.1, we see that when $\Psi(\beta_1, \beta_2)$ is smaller than $\rho_1^{-1} + \rho_2^{-1}$, the transfer is positive. Hence the data efficiency ratio quantifies how effective is the positive transfer using MTL. The precise statement can be found in Appendix B.2.

Algorithmic consequence. An interesting consequence of Proposition 3.2 is that $L(\hat{\beta}_t^{\text{MTL}})$ is not monotone in ρ_1. In particular, Figure 1b (and our analysis) shows that $L(\hat{\beta}_t^{\text{MTL}})$ behaves as a quadratic function over ρ_1. More generally, depending on how large $\Psi(\beta_1, \beta_2)$ is, $L(\hat{\beta}_t^{\text{MTL}})$ may also be monotonically increasing or decreasing. Based on this insight, we propose an incremental optimization schedule to improve MTL training efficiency.

- We divide the source task data into S batches. For S rounds, we incrementally add the source task data by adding one batch at a time.
- After training T epochs, if the validation accuracy becomes worse than the previous round's result, we terminate. Algorithm 1 in Appendix E describes the procedure in detail.

3.3 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 show that when n_1/n_2 is large, MTL with covariate shift can be suboptimal compared to MTL without covariate shift.

Example 3.3. We measure covariate shift by the matrix $M = \Sigma_1^{1/2}\Sigma_2^{-1/2}$. Assume that $\Psi(\beta_1, \beta_2) = 0$ for simplicity. 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 $1/\lambda$. Hence, λ measures the severity of covariate shift. Figure 1c shows a simulation of this setting by varying λ. We observe that as n_1/n_2 increases, the performance gap between the case of $M = \text{Id}$ and $\lambda = 2$ or 4 increases for MTL.

To compare different choices of M on the performance of $\hat{\beta}_t^{\text{MTL}}$, let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the singular values of M in decreasing order. Let $\mu_{\min} < \mu < \mu_{\max}$ are fixed values that do not grow with p. Consider the following bounded set

$$\mathcal{S}_\mu := \left\{ M \mid \prod_{i=1}^{p} \lambda_i \leqslant \mu^p, \mu_{\min} \leqslant \lambda_p, \lambda_1 \leqslant \mu_{\max} \right\},$$

Proposition 3.4 (Covariate shift)**.** *Assume that* $\Psi(\beta_1, \beta_2) = 0$ *and* $\rho_1, \rho_2 > 1$*. 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*

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

Models	MR, SST, SUBJ, CR, MPQA, TREC	
	MTL	**IncTrain**
MLP	100%	**xx%**
LSTM	**36%**	xx%

Table 1: Measuring the data efficiency ratio of muli-task learning.

Threshold	**Sentiment analysis**		**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

Table 2: Single-task learning results can help predict postive or negative transfer in multi-task learning.

The following proposition shows that when $n_1 \gg n_2$, $L(\hat{\beta}^{\text{MTL}})$ is minimized when there is no covariate shift between source and target task. The proof of Proposition 3.4 is left to Appendix B.3.

Algorithmic consequence. Our observation highlights the need to correct covariate shift when n_1/n_2 is large. Hence for such settings, we expect procedures that aim at correcting covariate shift to provide more significant gains. We consider a covariance alignment procedure proposed in Wu et al.'20 [18], which is designed for the purpose of correcting covariate shift. The idea is to add an alignment module between the input and the shared module B. This new module is then trained together with B and the output layers. We validate our insight on this procedure in the experiments.

4 Practical Connections: Detecting and Mitigating Negative Transfer

We connect our theory to pratical problems of interest. First, we validate the single-task based metric on text and image classification tasks. We show that for both datasets, single-task learning results can help predict positive or negative transfer in multi-task learning. Second, we show that our proposed incremental training schedule improves the training efficiency of multi-task learning for predicting a particular target task. Finally, we validate the three components of our theory and measure the data efficiency of multi-task learning on text classification tasks.

4.1 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[1]) followed by an LSTM or MLP layer proposed by [30].

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.

Training procedures and baselines. We use round-robin training schedule for the MTL model. We compare our incremental training scheduler of Algorithm 1 to the round-robin training scheduler. We focus on the prediction accuracy of a particular target task.

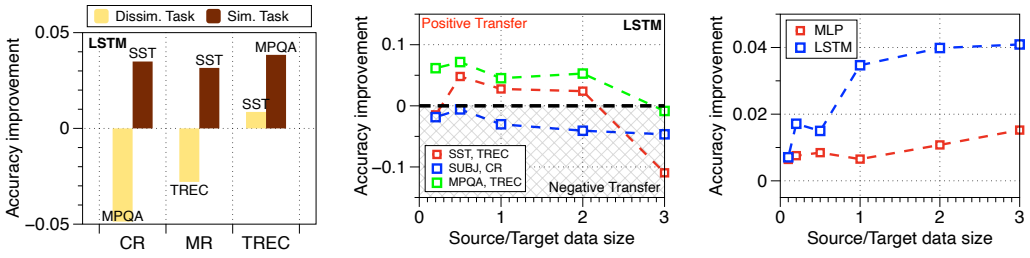
4.2 Experimental Results: Improving Data Efficiency

Improving multi-task training efficiency. We compare Algorithm 1 to performing multi-task learning for predicting a particular target dataset. Over six randomly selected task pairs from the sentiment analysis tasks, we find that Algorithm 1 requires only 45% of the computational cost to achieve the same performance on the target task, compared to the multi-task learning baseline.

Improving transfer learning training efficiency. We show that Algorithm 1 also applies to transfer learning settings. Compared to fine-tuning the source model on the target task, we show that our proposed method reduces the compute cost by $xx\%$, without sacrificing accuracy.

MTL improves data efficiency ratio. We measure the data efficiency ratio on the sentiment analysis tasks. In Table 1, we find that by performing multi-task learning, only 39% of the labeled data is needed to achieve comparable performance to single-task learning over all six tasks on LSTM.

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



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

Figure 2: Validating the three takeaways of Section 3 on sentiment analysis tasks. (a) Adding a semantically similar source task in MTL performs better than adding a dissimilar task. (b) As source/target data ratio increases, we observe a transition from positive to negative transfer. (c) As source/target data ratio increases, the performance gain from the covariance alignment procedure [18] over MTL increases for LSTM and MLP.

The data efficiency ratio of using MLP is 100% because the average performance of MTL is worse than the average of STL. We further show that applying incremental training helps reduce the data efficiency ratio to $xx\%$.

4.3 Ablation Studies

Understanding MTL results via STL results. We show that STL results can be used to help understand MTL results We validate the single-task based metric proposed in Section 3.1 for predicting positive or negative transfer in MTL. Table 2 shows the result on both the sentiment analysis and the ChestX-ray14 datasets. We find that using a threshold of $\tau = 0.1$, the STL results correctly predict positive or negative transfer with 75.6% accuracy and 38.8% recall among 150 task pairs. We observe similar results for 91 task pairs from the ChestX-ray14 dataset. The results show that STL performances are indicative of MTL performances.

Validation our theory. In Figure 2, we validate the three components of our theory on the sentiment analysis dataset. The experimental procedure is left to Appendix E.2.

5 Related Work

Multi-task learning. Adding a regularization over B, e.g. [14, 15]. Moreover, [31] observed that controlling the capacity can outperform the implicit capacity control of adding regularization over B.

Random matrix theory.

Transfer learning.

6 Conclusions and Discussions

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

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

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

[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] 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.

[15] 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.

[16] 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.

[17] 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.

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

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

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

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

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

[23] 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.

[24] 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.

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

[26] Terence Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Soc., 2012.

[27] László Erdos and Horng-Tzer Yau. A dynamical approach to random matrix theory. *Courant Lecture Notes in Mathematics*, 28, 2017.

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

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

[30] 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.

[31] 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.

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

[33] 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.

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

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

[36] 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.

[37] Shurong Zheng, Zhidong Bai, and Jianfeng Yao. Clt for eigenvalue statistics of large-dimensional general fisher matrices with applications. *Bernoulli*, 23(2):1130–1178, 2017.

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

[39] 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.

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

[41] 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.

[42] 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.

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

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

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

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

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

[48] 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.

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

[50] 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.

[51] 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

A.1 The multi-task learning estimator

From [18], 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 of two tasks with $r = 1$. Then equation (2.1) simplifies to

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

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

$$\begin{aligned}\hat{B}(w_1, w_2) &= (w_1^2 X_1^\top X_1 + w_2^2 X_2^\top X_2)^{-1}(w_1 X_1^\top Y_1 + w_2 X_2^\top Y_2) \qquad \text{(A.2)}\\ &= \frac{1}{w_2}\left(\frac{w_1^2}{w_2^2} X_1^\top X_1 + X_2^\top X_2\right)^{-1}\left(\frac{w_1}{w_2} X_1^\top Y_1 + X_2^\top Y_2\right)\\ &= \frac{1}{w_2}\left[\beta_t + \left(\frac{w_1^2}{w_2^2} X_1^\top X_1 + X_2^\top X_2\right)^{-1}\left(X_1^\top X_1\left(\frac{w_1}{w_2}\beta_1 - \frac{w_1^2}{w_2^2}\beta_2\right) + \left(\frac{w_1}{w_2} X_1^\top \varepsilon_1 + X_2^\top \varepsilon_2\right)\right)\right].\end{aligned}$$

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 β_1 is a scaling of β_2, then this case can be solved optimally using equation (A.1) [31].

Next we consider N_i independent samples of the training set $\{(\widetilde{x}_k^{(i)}, \widetilde{y}_k^{(i)}) : 1 \leqslant k \leqslant N_i\}$ from task-i, $i = 1, 2$. With these sample, we form the random matrices $\widetilde{X}_i \in \mathbb{R}^{N_i \times p}$ and $\widetilde{Y}_i \in \mathbb{R}^{N \times p}$, $i = 1, 2$, whose row vectors are given by $\widetilde{x}_k^{(i)}$ and $\widetilde{y}_k^{(i)}$. Here we assume that N_1 and N_2 satisfies $N_1/N_2 = n_1/n_2$ and $N_i \geqslant n_i^{1-\varepsilon_0}$ for some constant $\varepsilon_0 > 0$. Then we define the validation loss as

$$\widetilde{f}(\hat{B}; w_1, w_2) = \left\|\widetilde{X}_1 \hat{B} w_1 - \widetilde{Y}_1\right\|^2 + \left\|\widetilde{X}_2 \hat{B} w_2 - \widetilde{Y}_2\right\|^2. \tag{A.3}$$

Inserting (A.2) into (A.3), one can see that $\widetilde{f}$ only depends on the ratio $v := w_1/w_2$. Hence we will also write $\widetilde{f}(\hat{B}; v)$ in the following discussion.

Let $\hat{v} = \hat{w_1}/\hat{w_2}$ be the global minimizer of $\widetilde{f}(\hat{B}; v)$. We will define the multi-task learning estimator for the target task as

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

where $t = 2$ since we are considering the two task case, and it also stands for the "target task". The intuition for deriving $\hat{\beta}_t^{\text{MTL}}$ is akin to performing multi-task training in practice. Then the test loss of using $\hat{\beta}_t^{\text{MTL}}$ for the target task is

$$\begin{aligned}L(\hat{\beta}_t^{\text{MTL}}) = \;& \hat{v}^2 \left\|\Sigma_2^{1/2}(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1(\beta_1 - \hat{v}\beta_2)\right\|^2\\ &+ \sigma_2^2 \cdot \text{Tr}\left[\Sigma_2(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\right], \qquad \text{(A.4)}\end{aligned}$$

which is well-defined since it only depends on $\hat{v}$, but otherwise does not depend on $\hat{w}_1$ or $\hat{w}_2$ separately. Our goal is to study under model and covariate shifts, whether multi-task learning helps to learn the target task better than single-task learning. The baseline where we solve the target task with its own data is

$$L(\hat{\beta}_t^{\text{STL}}) = \sigma_2^2 \cdot \text{Tr}\left[\Sigma_2(X_2^\top X_2)^{-1}\right], \;\text{ where }\; \hat{\beta}_t^{\text{STL}} = (X_2^\top X_2)^{-1} X_2^\top Y_2.$$

One may observe that we can reduce $\widetilde{f}$ to an expression that is easier to handle using concentration of random vectors with i.i.d. entries. Before doing that, we first need to fix the setting for the following discussions, because we want to keep track of the error rate carefully instead of obtaining an asymptotic result only. First, we give the basic assumption for our main objects—the random matrices X_i, $i = 1, 2$.

Assumption A.1. *We will consider $n \times p$ 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 fixed $k \in \mathbb{N}$, there exists a constant $C_k > 0$ such that

$$\mathbb{E}|z_{ij}|^k \leqslant C_k, \quad 1 \leqslant i \leqslant n, \ \ 1 \leqslant j \leqslant p. \tag{A.5}$$

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.5) 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 some error bounds. We will not get deeper into this issue in this section, but refer the reader to Corollary C.8 below.

Then we make the following assumptions on the data models.

Assumption A.2. *For for some fixed $t \in \mathbb{N}$, let $Y_i = X_i\beta_i + \varepsilon_i$, $1 \leqslant i \leqslant t$, be independent data models, where X_i, β_i and ε_i are also independent of each other. Suppose that $X_i = Z_i\Sigma_i^{1/2} \in \mathbb{R}^{n_i \times p}$ 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_i^2 and all moments as in* (A.5)*.*

Throughout the appendix, we shall say an event Ξ holds with high probability (whp) if for any fixed $D > 0$, $\mathbb{P}(\Xi) \geqslant 1 - p^{-D}$ for large enough p. Moreover, we shall use $\mathrm{o}(1)$ to mean a small positive number that converges to 0 as $p \to \infty$.

Now suppose $Y_i = X_i\beta_i + \varepsilon_i$ and $\widetilde{Y}_i = \widetilde{X}_i\beta_i + \widetilde{\varepsilon}_i$, $i = 1, 2$, all satisfy Assumption A.2. Then we rewrite (A.3) as

$$\widetilde{f}(\hat{B}; v) = \sum_{i=1}^{2} \left\| \widetilde{X}_i\widetilde{\beta}_i - \widetilde{\varepsilon}_i \right\|^2, \quad \widetilde{\beta} := \hat{B}w_i - \beta_i.$$

Since $\widetilde{X}_i\widetilde{\beta}$ and $\widetilde{\varepsilon}_i$ are independent random vectors with i.i.d. centered entries, we can use the concentration estimate, Lemma D.6, to get that for any constant $\varepsilon > 0$,

$$\begin{aligned}\left| \left\| \widetilde{X}_i\widetilde{\beta}_i - \widetilde{\varepsilon}_i \right\|^2 - \mathop{\mathbb{E}}_{\widetilde{X}_i, \widetilde{\varepsilon}_i} \left[\left\| \widetilde{X}_i\widetilde{\beta}_i - \widetilde{\varepsilon}_i \right\|^2 \right] \right| &= \left| \left\| \widetilde{X}_i\widetilde{\beta}_i - \widetilde{\varepsilon}_i \right\|^2 - N_i(\widetilde{\beta}_i^\top \Sigma_i \widetilde{\beta}_i + \sigma_i^2) \right| \\ &\leqslant N_i^{1/2+\varepsilon}(\widetilde{\beta}_i^\top \Sigma_i \widetilde{\beta}_i + \sigma_i^2),\end{aligned}$$

with high probability. Thus we obtain that

$$\widetilde{f}(B; v) = \left[\sum_{i=1}^{2} N_i \left\| \Sigma_i^{1/2}(\hat{B}w_i - \beta_i) \right\|^2 + (N_1\sigma_1^2 + N_2\sigma_2^2) \right] \cdot \left(1 + \mathrm{O}(p^{-(1-\varepsilon_0)/2+\varepsilon}) \right),$$

where we also used $N_i \geqslant p^{-1+\varepsilon_0}$. Inserting (A.2) into the above expression and using again the concentration result, Lemma D.6, we obtain that

$$\sum_{i=1}^{2} N_i \left\| \Sigma_i^{1/2}(\hat{B}w_1 - \beta_i) \right\|^2 = val(\hat{B}; v) \cdot \left(1 + \mathrm{O}(p^{-1/2+\varepsilon}) \right)$$

with high probability, where

$$\begin{aligned}val(\hat{B}; v) &:= \mathop{\mathbb{E}}_{\varepsilon_1, \varepsilon_2} \left[\sum_{i=1}^{2} \left\| \Sigma_i^{1/2}(\hat{B}w_i - \beta_i) \right\|^2 \right] \\ &= N_1 \cdot \left\| \Sigma_1^{1/2} \left(v^2 X_1^\top X_1 + X_2^\top X_2 \right)^{-1} X_2^\top X_2 \left(\beta_1 - v\beta_2 \right) \right\|^2 \\ &+ N_2 \cdot v^2 \left\| \Sigma_2^{1/2} \left(v^2 X_1^\top X_1 + X_2^\top X_2 \right)^{-1} X_1^\top X_1 \left(\beta_1 - v\beta_2 \right) \right\|^2 \\ &+ N_1 \cdot v^2 \operatorname{Tr} \left[\Sigma_1 \left(v^2 X_1^\top X_1 + X_2^\top X_2 \right)^{-2} \left(\sigma_1^2 \cdot v^2 X_1^\top X_1 + \sigma_2^2 \cdot X_2^\top X_2 \right) \right] \\ &+ N_2 \cdot \operatorname{Tr} \left[\Sigma_2 \left(v^2 X_1^\top X_1 + X_2^\top X_2 \right)^{-2} \left(\sigma_1^2 \cdot v^2 X_1^\top X_1 + \sigma_2^2 \cdot X_2^\top X_2 \right) \right].\end{aligned}$$

In sum, we have obtained that

$$\widetilde{f}(\hat{B}; v) = \Big[val(\hat{B}; v) + (N_1\sigma_1^2 + N_2\sigma_2^2)\Big] \cdot \Big(1 + \mathrm{O}(p^{-(1-\varepsilon_0)/2+\varepsilon})\Big) \tag{A.6}$$

Hence to minimize $\widetilde{f}$, it suffices to minimize $val(\hat{B}; v)$ over v.

We now state several helper lemmas to get estimates on $L(\hat{\beta}_t^{\mathrm{STL}})$ and $L(\hat{\beta}_t^{\mathrm{MTL}})$. The first lemma, which is a folklore result in random matrix theory, helps to determine the asymptotic limit of $L(\hat{\beta}_t^{\mathrm{STL}})$, 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 [32]. For general non-Gaussian random matrices, it can be obtained from Stieltjes transform method; see e.g., Lemma 3.11 of [25]. Here we shall state a result obtained from Theorem 2.4 in [33], which gives an almost sharp error bound.

Lemma A.3. *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 $\varepsilon > 0$,*

$$\mathrm{Tr}\left[(X^\top X)^{-1} A\right] = \frac{1}{\rho - 1}\frac{1}{p}\mathrm{Tr}(\Sigma^{-1}A) + \mathrm{O}\left(\|A\| p^{-1/2+\varepsilon}\right) \tag{A.7}$$

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.3 for a single sample covariance matrix to the sum of two independent sample covariance matrices. It is the main random matrix theoretical input of this paper.

Lemma A.4. *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 $\rho_2 := n_2/p > 1$ being fixed constants. 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 $\varepsilon > 0$,*

$$\mathrm{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1} A\right] = \frac{1}{\rho_1 + \rho_2}\frac{1}{p}\mathrm{Tr}\left[(a_1\Sigma_1 + a_2\Sigma_2)^{-1}A\right] + \mathrm{O}\left(\|A\| p^{-1/2+\varepsilon}\right) \tag{A.8}$$

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

$$a_1 + a_2 = 1 - \frac{1}{\rho_1 + \rho_2}, \quad a_1 + \frac{1}{\rho_1 + \rho_2} \cdot \frac{1}{p}\sum_{i=1}^{p} \frac{\lambda_i^2 a_1}{\lambda_i^2 a_1 + a_2} = \frac{\rho_1}{\rho_1 + \rho_2}. \tag{A.9}$$

Finally, the last lemma describes the asymptotic limit of $(X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_2(X_1^\top X_1 + X_2^\top X_2)^{-1}$, which will be needed when we estimate the first term on the right-hand side of (A.4).

Lemma A.5. *In the setting of Lemma A.4, let $\beta \in \mathbb{R}^p$ be any vector that is independent of X_1 and X_2. We have that for any constant $\varepsilon > 0$,*

$$\begin{aligned} &(n_1 + n_2)^2 \left\|\Sigma_2^{1/2}(X_1^\top X_1 + X_2^\top X_2)^{-1}\beta\right\|^2 \\ &= \beta^\top \Sigma_2^{-1/2} \frac{(1 + a_3)\,\mathrm{Id} + a_4 M^\top M}{(a_1 M^\top M + a_2)^2} \Sigma_2^{-1/2}\beta + \mathrm{O}(p^{-1/2+\varepsilon}\|\beta\|^2), \end{aligned} \tag{A.10}$$

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

$$\left(\rho_2 a_2^{-2} - b_0\right) \cdot a_3 - b_1 \cdot a_4 = b_0, \quad \left(\rho_1 a_1^{-2} - b_2\right) \cdot a_4 - b_1 \cdot a_3 = b_1. \tag{A.11}$$

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.4 and Lemma A.5 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. [34]), but our results (A.8) and (A.10) also give an almost sharp error bound $\mathrm{O}\left(p^{-1/2+\varepsilon}\right)$.

A.2 Proof for Two Tasks with General Covariates

In this section, we state and prove the formal version of Theorem 2.1, which covers the two tasks case with $t = 2$. In this section, we consider the case where the entries of ε_1 and ε_2 have the same variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

First, we introduce several quantities that will be used in our statement, and they are also related to the quantities in Lemma A.4 and Lemma A.5. 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, and $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \ldots \geqslant \hat{\lambda}_p$ be the eigenvalues of $\hat{M}^\top \hat{M}$. Define $(\hat{a}_1, \hat{a}_2)$ as the solution to the following system of deterministic equations,

$$\hat{a}_1 + \hat{a}_2 = 1 - \frac{1}{\rho_1 + \rho_2}, \quad \hat{a}_1 + \frac{1}{\rho_1 + \rho_2} \cdot \frac{1}{p} \sum_{i=1}^{p} \frac{\hat{\lambda}_i^2 \hat{a}_1}{\hat{\lambda}_i^2 \hat{a}_1 + \hat{a}_2} = \frac{\rho_1}{\rho_1 + \rho_2}. \tag{A.12}$$

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

$$\left(\rho_2 \hat{a}_2^{-2} - \hat{b}_0\right) \cdot \hat{a}_3 - \hat{b}_1 \cdot \hat{a}_4 = \hat{b}_0, \quad \left(\rho_1 \hat{a}_1^{-2} - \hat{b}_2\right) \cdot \hat{a}_4 - \hat{b}_1 \cdot \hat{a}_3 = \hat{b}_1. \tag{A.13}$$

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

$$\Pi = \frac{\rho_1^2}{(\rho_1 + \rho_2)^2} \cdot \hat{M} \frac{(1 + \hat{a}_3)\,\mathrm{Id} + \hat{a}_4 \hat{M}^\top \hat{M}}{(\hat{a}_1 \hat{M}^\top \hat{M} + \hat{a}_2)^2} \hat{M}^\top. \tag{A.14}$$

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

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

In fact, $\alpha_-(\rho_1)$ and $\alpha_+(\rho_1)$ correspond to the largest and smallest singular values of $Z_1/\sqrt{n_1}$, respectively, as given by the famous Marčenko-Pastur law [35]. 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

$$\delta \equiv \delta(\hat{v}) := \frac{\alpha_+^2(\rho_1) - 1}{\alpha_-^2(\rho_1) \lambda_{\min}^2(\hat{M})} \cdot \|\Sigma_1^{1/2}(\beta_1 - \hat{v}\beta_2)\|^2, \tag{A.15}$$

where $\lambda_{\min}(\hat{M})$ is the smallest singular value of $\hat{M}$. Note that this factor converges to 0 as ρ_1 increases.

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 determined by two deterministic quantities Δ_{bias} and Δ_{var}, which give the change of model shift bias and the change of variance, respectively.

Theorem A.6. *Consider two data models $Y_i = X_i\beta_i + \varepsilon_i$, $i = 1, 2$, that satisfy Assumption A.2. With high probability, we have*

$$L(\hat{\beta}_t^{MTL}) \leqslant L(\hat{\beta}_t^{STL}) \quad \textit{when:} \quad \Delta_{var} - \Delta_{bias} \geqslant \delta \tag{A.16}$$

$$L(\hat{\beta}_t^{MTL}) \geqslant L(\hat{\beta}_t^{STL}) \quad \textit{when:} \quad \Delta_{var} - \Delta_{bias} \leqslant -\delta, \tag{A.17}$$

where

$$\Delta_{var} := \sigma^2 \left(\frac{1}{\rho_2 - 1} - \frac{1}{\rho_1 + \rho_2} \cdot \frac{1}{p} \operatorname{Tr}\left[(\hat{a}_1 \hat{M}^\top \hat{M} + \hat{a}_2\,\mathrm{Id})^{-1} \right] \right) \tag{A.18}$$

$$\Delta_{bias} := (\beta_1 - \hat{v}\beta_2)^\top \Sigma_1^{1/2} \Pi \Sigma_1^{1/2} (\beta_1 - \hat{v}\beta_2). \tag{A.19}$$

For the isotropic model in Section 3, we actually have an easier and sharper bound than Theorem A.6 as follows.

Lemma A.7. *In the setting of Theorem A.6, assume that $\Sigma_1 = \mathrm{Id}$, β_2 is a random vector with i.i.d. entries with mean 0, variance κ^2 and all moments, and β_1 is a random vector such that $(\beta_1 - \beta_2)$ is a random vector with i.i.d. entries with mean 0, variance d^2 and all moments. Denote $\Delta^\star_{bias} := \left((1-\hat{v})^2\kappa^2 + d^2\right) \mathrm{Tr}\left[\Pi\right]$. Then we have*

$$L(\hat{\beta}_t^{MTL}) \leqslant L(\hat{\beta}_t^{STL}) \quad \textit{when:} \quad \Delta_{var} \geqslant (\alpha_+^2(\rho_1) + \mathrm{o}(1)) \cdot \Delta^\star_{bias},$$
$$L(\hat{\beta}_t^{MTL}) \geqslant L(\hat{\beta}_t^{STL}) \quad \textit{when:} \quad \Delta_{var} \leqslant (\alpha_-^2(\rho_1) - \mathrm{o}(1)) \cdot \Delta^\star_{bias}.$$

Now we give the proof of Theorem A.6 based on Lemma A.4 and Lemma A.5.

Proof of Theorem A.6. Note that

$$L(\hat{\beta}_t^{\mathrm{STL}}) - L(\hat{\beta}_t^{\mathrm{MTL}}) = \sigma^2 \left(\mathrm{Tr}\left[(X_2^\top X_2)^{-1}\Sigma_2\right] - \mathrm{Tr}\left[(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_2\right]\right)$$
$$- \hat{v}^2 \left\|\Sigma_2^{1/2}(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1(\beta_1 - \hat{v}\beta_2)\right\|^2 =: \delta_{\mathrm{var}}(\hat{v}) - \delta_{\mathrm{bias}}(\hat{v}).$$

The proof is divided into the following four steps.

(i) We first consider $\hat{M} \equiv \hat{M}(v) = v\Sigma_1^{1/2}\Sigma_2^{-1/2}$ for a fixed $v \geqslant 0$. Then we use Lemma A.3 and Lemma A.4 to calculate the variance reduction $\delta_{\mathrm{var}}(v)$, which will lead to the Δ_{var} term.

(ii) Using the approximate isometry property of X_1 (see (A.22) below), we will bound the bias term $\delta_{\mathrm{bias}}(v)$ through

$$\widetilde{\delta}_{\mathrm{bias}}(v) := v^2 n_1^2 \left\|\Sigma_2^{1/2}(v^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_1(\beta_1 - v\beta_2)\right\|^2. \tag{A.20}$$

(iii) We use Lemma A.5 to calculate (A.20), which will lead to the Δ_{bias} term.

(iv) Finally we use a standard ε-net argument to extend the above results to $\hat{M} = \hat{v}\Sigma_1^{1/2}\Sigma_2^{-1/2}$ for a possibly random $\hat{v}$ which depends on Y_1 and Y_2.

Step I: Variance reduction. Let $\hat{M} = v\Sigma_1^{1/2}\Sigma_2^{-1/2}$ for any fixed constant $v \geqslant 0$. Using Lemma A.4, we can obtain that for any constant $\varepsilon > 0$,

$$\sigma^2 \cdot \mathrm{Tr}\left[(X_2^\top X_2)^{-1}\Sigma_2\right] = \frac{\sigma^2}{\rho_2 - 1}\left(1 + \mathrm{O}(p^{-1/2+e})\right),$$

and

$$\sigma^2 \cdot \mathrm{Tr}\left[(v^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_2\right] = \frac{\sigma^2}{\rho_1 + \rho_2} \cdot \frac{1}{p}\,\mathrm{Tr}\left[(\hat{a}_1 \hat{M}^\top \hat{M} + \hat{a}_2\,\mathrm{Id})^{-1}\right]\left(1 + \mathrm{O}(p^{-1/2+e})\right),$$

with high probability, where $\hat{a}_1$ and $\hat{a}_2$ satify (A.12). Combining them, we get

$$\delta_{\mathrm{var}}(v) = \Delta_{\mathrm{var}}(v) + \mathrm{O}(\sigma^2 p^{-1/2+e}) \quad \text{whp}, \tag{A.21}$$

where $\Delta_{\mathrm{var}}(v)$ is defined as in (A.18) but with $\hat{v}$ replaced by v.

Step II: Bounding the bias term. In this step, we shall use the following the following bounds on the singular values of Z_1: for any fixed $\varepsilon > 0$, we have

$$\alpha_-(\rho_1) - \mathrm{O}(p^{-1/2+e}) \preceq \frac{Z_1^T Z_1}{n_1} \preceq \alpha_+(\rho_1) + \mathrm{O}(p^{-1/2+e}) \tag{A.22}$$

with high probability. In fact, $Z_1^T Z_1$ 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 $[\alpha_-(\rho_1) - \mathrm{o}(1), \alpha_+(\rho_1) + \mathrm{o}(1)]$ with probability $1 - \mathrm{o}(1)$ [36]. For the estimate (A.22) we used [33, Theorem 2.10] to get a stronger probability bound.

Next we shall use (A.22) to approximate $\delta_{\mathrm{bias}}(v)$ with $\widetilde{\delta}_{\mathrm{bias}}(v)$ in (A.20).

Lemma A.8. *In the setting of Theorem A.6, we denote by* $K = (v^2 X_1^\top X_1 + X_2^\top X_1)^{-1}$*, and*

$$\delta_\varepsilon(v) := n_1^2 v^2 \left\|\Sigma_1^{1/2} K \Sigma_2 K \Sigma_1^{1/2}\right\| \cdot \left\|\Sigma_1^{1/2}(\beta_1 - v\beta_2)\right\|^2.$$

Then we have whp,

$$\left|\delta_{bias}(v) - \widetilde{\delta}_{bias}(v)\right| \leqslant \left(\alpha_+^2(\rho_1) - 1 + \mathrm{O}(p^{-1/2+\varepsilon})\right) \delta_\varepsilon.$$

Proof. Denote by $\mathcal{E} = Z_1^\top Z_1 - n_1 \,\mathrm{Id}$. Then we can write

$$\begin{aligned}\delta_{\text{bias}}(v) - \widetilde{\delta}_{\text{bias}}(v) &= 2v^2 n_1 (\beta_1 - v\beta_2)^\top \Sigma_1^{1/2} \mathcal{E} \left(\Sigma_1^{1/2} K \Sigma_2 K \Sigma_1^{1/2}\right) \Sigma_1^{1/2}(\beta_1 - v\beta_2) \\ &\quad + v^2 \left\|\Sigma_2^{1/2} K \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2}(\beta_1 - v\beta_2)\right\|^2. \end{aligned} \tag{A.23}$$

Using (A.22), we can bound

$$\|\mathcal{E}\| \leqslant \left(\alpha_+(\rho_1) - 1 + \mathrm{O}(p^{-1/2+\varepsilon})\right) n_1, \quad \text{whp.}$$

Thus we can estimate that

$$\begin{aligned}|\delta_{\text{bias}}(v) - \widetilde{\delta}_{\text{bias}}(v)| &\leqslant v^2 \left(2n_1\|\mathcal{E}\| + \|\mathcal{E}\|^2\right) \left\|\Sigma_1^{1/2} K \Sigma_2 K \Sigma_1^{1/2}\right\| \left\|\Sigma_1^{1/2}(\beta_1 - v\beta_2)\right\|^2 \\ &= v^2 \left[(n_1 + \|\mathcal{E}\|)^2 - n_1^2\right] \left\|\Sigma_1^{1/2} K \Sigma_2 K \Sigma_1^{1/2}\right\| \left\|\Sigma_1^{1/2}(\beta_1 - v\beta_2)\right\|^2 \\ &\leqslant v^2 n_1^2 \left[\alpha_+^2(\rho_1) + \mathrm{O}(p^{-1/2+\varepsilon}) - 1\right] \left\|\Sigma_1^{1/2} K \Sigma_2 K \Sigma_1^{1/2}\right\| \left\|\Sigma_1^{1/2}(\beta_1 - v\beta_2)\right\|^2, \end{aligned}$$

which concludes the proof by the definition of δ_ε. □

Note by (A.22), we have with high probability,

$$\begin{aligned} v^2 n_1^2 \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1^{1/2} &= \hat{M} \frac{1}{(\hat{M}^\top Z_1^\top Z_1 \hat{M} + Z_2^\top Z_2)^2} \hat{M}^\top \\ &\preceq n_1^2 \hat{M} \frac{1}{\left[n_1 \alpha_-(\rho_1) \hat{M}^\top \hat{M} + n_2 \alpha_-(\rho_2) + \mathrm{O}(p^{1/2+\varepsilon})\right]^2} \hat{M}^\top \\ &\preceq \left[\alpha_-^2(\rho_1) \hat{M}\hat{M}^\top + 2\frac{\rho_2}{\rho_1} \alpha_-(\rho_1)\alpha_-(\rho_2) + 2\left(\frac{\rho_2}{\rho_1}\right)^2 \alpha_-^2(\rho_2)(\hat{M}\hat{M}^\top)^{-1}\right]^{-1} + \mathrm{O}(p^{-1/2+\varepsilon}) \\ &\prec [\alpha_-^2(\rho_1)\lambda_{\min}^2(\hat{M})]^{-1} \cdot (1-c) \end{aligned}$$

for some small enough constant $c > 0$. Together with Lemma A.8, we get with high probability,

$$\left|\delta_{\text{bias}}(v) - \widetilde{\delta}_{\text{bias}}(v)\right| \leqslant (1-c)\delta(v) \tag{A.24}$$

for some small constant $c > 0$, where recall $\delta(v)$ defined in (A.15).

Step III: The limit of $\widetilde{\delta}_{\text{bias}}(v)$. Using Lemma A.5 with Σ_1 and M replaced by $v^2\Sigma_1$ and $\hat{M}$, we obtain that

$$\begin{aligned}\widetilde{\delta}_{\text{bias}}(v) &= \frac{\rho_1^2}{(\rho_1+\rho_2)^2} \cdot v^2 (\beta_1 - v\beta_2)^\top \Sigma_1 \Sigma_2^{-1/2} \frac{(1+\hat{a}_3)\,\mathrm{Id} + \hat{a}_4 \hat{M}^\top \hat{M}}{(a_1 \hat{M}^\top \hat{M} + a_2)^2} \Sigma_2^{-1/2} \Sigma_1 (\beta_1 - v\beta_2) + \mathrm{O}(p^{-1/2+\varepsilon}) \\ &= (\beta_1 - v\beta_2)^\top \Sigma_1^{1/2} \Pi \Sigma_1^{1/2} (\beta_1 - v\beta_2) + \mathrm{O}(p^{-1/2+\varepsilon}) =: \Delta_{\text{bias}}(v) + \mathrm{O}(p^{-1/2+\varepsilon}), \end{aligned}$$

with high probability. Together with and (A.21) and (A.24), we obtain that whp,

$$\begin{cases} \delta_{\text{var}}(v) > \delta_{\text{bias}}(v), & \text{if } \ \Delta_{\text{var}}(v) - \Delta_{\text{bias}}(v) \geqslant \delta(v), \\ \delta_{\text{var}}(v) < \delta_{\text{bias}}(v), & \text{if } \ \Delta_{\text{var}}(v) - \Delta_{\text{bias}}(v) \leqslant -\delta(v). \end{cases} \tag{A.25}$$

Step IV: An ε-net argument. Finally, it remains to extend the above result to $v = \hat{v}$, which is random and depends on X_1 and X_2. We first show that for any fixed constant $C_0 > 0$, there exists a high probability event Ξ on which (A.25) holds uniformly for all $v \in [0, C_0]$. In fact, for a large constant $C_1 > 0$, we consider v belonging to a discrete set

$$V := \{v_k = kp^{-1} : 1 \leqslant k \leqslant C_0 p + 1\}.$$

Then using the arguments for the first three steps and a simple union bound, we get that (A.25) holds simultaneously for all $v \in V$ with high probability. On the other hand, by (A.22) the event

$$\Xi_1 := \left\{ \alpha_-(\rho_1)/2 \preceq \frac{Z_1^T Z_1}{n_1} \preceq 2\alpha_+(\rho_1),\ \alpha_-(\rho_2)/2 \preceq \frac{Z_2^T Z_2}{n_2} \preceq 2\alpha_+(\rho_2) \right\}$$

holds with high probability. Now it is easy to check that on Ξ_1, for all $v_k \leqslant v \leqslant v_{k+1}$ we have the following estimates:

$$|\delta_{\text{var}}(v) - \delta_{\text{var}}(v_k)| \lesssim p^{-1}\delta_{\text{var}}(v_k),\ \ |\delta_{\text{bias}}(v) - \delta_{\text{bias}}(v_k)| \lesssim p^{-1}\delta_{\text{bias}}(v_k),\ \ |\delta(v) - \delta(v_k)| \lesssim p^{-1}\delta(v_k),$$
$$|\Delta_{\text{bias}}(v) - \Delta_{\text{bias}}(v_k)| \lesssim p^{-1}\Delta_{\text{bias}}(v_k),\ \ |\Delta_{\text{var}}(v) - \Delta_{\text{var}}(v_k)| \lesssim p^{-1}\Delta_{\text{var}}(v_k).$$

Then a simple application of triangle inequality gives that the event

$$\Xi_2 = \{\text{(A.25) holds simultaneously for all } 0 \leqslant v \leqslant C_0\}$$

holds with high probability. On the other hand, on Ξ_1 one can see that for any small constant $\varepsilon > 0$,

$$|\delta_{\text{var}}(v) - \delta_{\text{var}}(C_0)| \leqslant \varepsilon\delta_{\text{var}}(C_0), \quad |\delta_{\text{bias}}(v) - \delta_{\text{bias}}(C_0)| \leqslant \varepsilon\delta_{\text{bias}}(C_0), \quad |\delta(v) - \delta(C_0)| \leqslant \varepsilon\delta(C_0),$$
$$|\Delta_{\text{bias}}(v) - \Delta_{\text{bias}}(C_0)| \leqslant \varepsilon\Delta_{\text{bias}}(C_0), \quad |\Delta_{\text{var}}(v) - \Delta_{\text{var}}(C_0)| \leqslant \varepsilon\Delta_{\text{var}}(C_0),$$

for all $v \geqslant C_0$ as long as C_0 is chosen large enough depending on ε. Together with the estimate at C_0, we get that (A.25) holds simultaneously for all $v \geqslant 0$ on the high probability event $\Xi_1 \cap \Xi_2$. This concludes the proof since v must be one of the positive values. □

Remark A.9. One can see from the above proof that the main error, δ, of Theorem A.6 comes from approximating δ_{bias} by $\widetilde{\delta}_{\text{bias}}$ in (A.24). In order to improve this estimate and obtain an exact asymptotic result as for the δ_{var} term, one needs to study the singular value distribution of the following random matrix:

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

In fact, the eigenvalues of $\mathcal{X} := (X_1^\top X_1)^{-1} X_2^\top X_2$ have been studied in the name of Fisher matrices; see e.g. [37]. However, since $\mathcal{X}$ is not symmetric, it is known that the singular values of $\mathcal{X}$ are different from its eigenvalues. To the best of our knowledge, the asymptotic singular value behavior of $\mathcal{X}$ is still unknown in random matrix theory literature, and the study of the singular values of $\mathcal{X} + v^2$ will be even harder. We leave this problem to future study.

By replacing (A.24) with a tighter bound in Step II of the above proof, we can conclude the proof of Lemma A.7.

Proof of Lemma A.7. For any fixed $v \geqslant 0$, $\beta_1 - v\beta_2$ is a random vector with i.i.d. entries with mean 0 and variance $(1-v)^2\kappa^2 + d^2$. Then using the concentration result, Lemma D.6, we get that for any constant $\varepsilon > 0$,

$$\begin{aligned}
&\left|\delta_{\text{bias}}(v) - [(1-v)^2\kappa^2 + d^2]\operatorname{Tr}(\mathcal{K}^\top\mathcal{K})\right| \\
&= \left|(\beta_1 - v\beta_2)^\top \mathcal{K}^\top\mathcal{K}(\beta_1 - v\beta_2) - [(1-v)^2\kappa^2 + d^2]\operatorname{Tr}(\mathcal{K}^\top\mathcal{K})\right| \\
&\leqslant p^{\varepsilon}[(1-v)^2\kappa^2 + d^2]\left\{\operatorname{Tr}\left[(\mathcal{K}^\top\mathcal{K})^2\right]\right\}^{1/2} \lesssim p^{1/2+\varepsilon}[(1-v)^2\kappa^2 + d^2],
\end{aligned} \tag{A.26}$$

where we denoted $\mathcal{K} := v\Sigma_2^{1/2}(v^2 X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1$, and in the last step we used $\|\mathcal{K}\| = \mathrm{O}(1)$ by (A.22). Now for $\operatorname{Tr}(\mathcal{K}^\top\mathcal{K})$, we rewrite it as

$$v^2[(1-v)^2\kappa^2 + d^2]\operatorname{Tr}\left[(v^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_2(v^2 X_1^\top X_1 + X_2^\top X_2)^{-1}(X_1^\top X_1)^2\right].$$

Recalling that $\Sigma_1 = \mathrm{Id}$ and bounding $(X_1^\top X_1)^2 = (Z_1^\top Z_1)^2$ using (A.22) again, we obtain that

$$\delta^\star_{\text{bias}}(v)\cdot(\alpha_-^2(\rho_1) - \mathrm{O}(p^{-1/2+\varepsilon})) \leqslant [(1-v)^2\kappa^2 + d^2]\operatorname{Tr}(\mathcal{K}^\top\mathcal{K}) \leqslant \delta^\star_{\text{bias}}(v)\cdot(\alpha_+^2(\rho_1) + \mathrm{O}(p^{-1/2+\varepsilon})), \tag{A.27}$$

where

$$\delta^{\star}_{\text{bias}}(v) := n_1^2 v^2[(1-v)^2\kappa^2 + d^2]\,\mathrm{Tr}\left[(v^2X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_2(v^2X_1^\top X_1 + X_2^\top X_2)^{-1}\right].$$

Note that $\delta^{\star}_{\text{bias}}(v) \sim 1$, hence combining (A.26) and (A.27) we get

$$\delta^{\star}_{\text{bias}}(v)\cdot(\alpha_-^2(\rho_1) - \mathrm{O}(p^{-1/2+\varepsilon})) \leqslant \delta_{\text{bias}}(v) \leqslant \delta^{\star}_{\text{bias}}(v)\cdot(\alpha_+^2(\rho_1) + \mathrm{O}(p^{-1/2+\varepsilon})). \tag{A.28}$$

Now we can replace the estimate (A.24) with this stronger estimate, and repeat all the other parts of the proof of Theorem A.6 to conclude Lemma A.7. In particular, one can calculate $\delta^{\star}_{\text{bias}}(v)$ using Lemma A.5 and get the $\Delta^{\star}_{\text{bias}}(v)$ term, We omit the details. □

A.3 Proof for Many Tasks with the Same Covariate

Proof of Theorem 2.2. In this setting, we need to study the following loss function:

$$f(B; W_1, \ldots, W_t) = \sum_{i=1}^{t} \|XBW_i - Y_i\|^2. \tag{A.29}$$

For any fixed $W_1, W_2, \ldots, W_t \in \mathbb{R}^r$, we can derive a closed form solution for B as

$$\begin{aligned}\hat{B}(W_1,\ldots,W_t) &= (X^\top X)^{-1}X^\top\left(\sum_{i=1}^{t} Y_i W_i^\top\right)(\mathcal{W}\mathcal{W}^\top)^{-1}\\ &= (B^\star\mathcal{W}^\top)(\mathcal{W}\mathcal{W}^\top)^{-1} + (X^\top X)^{-1}X^\top\left(\sum_{i=1}^{t}\varepsilon_i W_i^\top\right)(\mathcal{W}\mathcal{W}^\top)^{-1},\end{aligned}$$

where we denote $\mathcal{W} \in \mathbb{R}^{r\times t}$ as $\mathcal{W} = [W_1, W_2, \ldots, W_t]$. Then as in (A.6), we pick N independent samples of the training set for each task with $N \geqslant n^{1-\varepsilon_0}$, and use concentration to get the validation loss as

$$\widetilde{f}(\hat{B}; \mathcal{W}) = N\left[val(\mathcal{W}) + t\sigma^2\right]\cdot\left(1 + \mathrm{O}(p^{-(1-\varepsilon_0)/2+\varepsilon})\right). \tag{A.30}$$

Here $val(\mathcal{W})$ is defined as

$$val(\mathcal{W}) := \mathop{\mathbb{E}}_{\varepsilon_j, \forall 1\leqslant j\leqslant t}\left[\sum_{i=1}^{t}\left\|\Sigma^{1/2}(\hat{B}W_i - \beta_i)\right\|^2\right] = \delta_{\text{bias}}(\mathcal{W}) + \delta_{\text{var}}(\mathcal{W}),$$

where the model shift bias term $\delta_{\text{bias}}(\mathcal{W})$ is given by

$$\delta_{\text{bias}}(\mathcal{W}) := \sum_{i=1}^{t}\left\|\Sigma^{1/2}\left((B^\star\mathcal{W}^\top)(\mathcal{W}\mathcal{W}^\top)^{-1}W_i - \beta_i\right)\right\|^2,$$

and the variance term $\delta_{\text{var}}(\mathcal{W})$ can be calculated as

$$\delta_{\text{var}}(\mathcal{W}) := \sigma^2\cdot\mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right].$$

It suffices to minimize $\delta_{\text{bias}}(\mathcal{W})$ over $\mathcal{W}$, since both $tN\sigma^2$ and $\delta_{\text{var}}(\mathcal{W})$ do not depend on the weights. We denote $Q := \mathcal{W}^\top(\mathcal{W}\mathcal{W}^\top)^{-1}\mathcal{W} \in \mathbb{R}^{k\times k}$, whose (i,j)-th entry is equal to $W_i^\top(\mathcal{W}\mathcal{W}^\top)^{-1}W_j$. Now we can write $\delta_{\text{bias}}(\mathcal{W})$ succinctly as

$$\delta_{\text{bias}}(\mathcal{W}) = \left\|\Sigma^{1/2}B^\star(Q - \mathrm{Id})\right\|_F^2.$$

From this equation we can solve the minimizer optimally as $Q_0 = U_r U_r^\top$. On the other hand, let $\hat{\mathcal{W}}$ be the minimizer of $\widetilde{f}$, and denote $\hat{Q} := \hat{\mathcal{W}}^\top(\hat{\mathcal{W}}\hat{\mathcal{W}}^\top)^{-1}\hat{\mathcal{W}}$. We claim that $\hat{Q}$ satisfies

$$\|Q_0^{-1}\hat{Q} - \mathrm{Id}\,\| = \mathrm{o}(1) \quad \text{whp.} \tag{A.31}$$

In fact, if (A.31) does not hold, then using the condition $\lambda_{\min}((B^\star)^\top\Sigma B^\star) \gtrsim \sigma^2$ and that $\delta_{\text{var}}(\mathcal{W}) = \mathrm{O}(\sigma^2)$ by (A.22), we obtain that

$$val(\hat{\mathcal{W}}) + t\sigma^2 > (val(\mathcal{W}_0) + t\sigma^2)\cdot(1 + \mathrm{o}(1)) \;\Rightarrow\; \widetilde{f}(\hat{B}; \hat{\mathcal{W}}) > \widetilde{f}(\hat{B}; \mathcal{W}_0),$$

that is, $\hat{\mathcal{W}}$ is not a minimizer. This leads to a contradiction.

In sum, we have solved that $\hat{\beta}_i^{\mathrm{MTL}} = B^\star\left(U_r U_r(i) + \mathrm{o}(1)\right)$. Inserting it into the definition of the test loss, we get that

$$\begin{aligned} L(\hat{\beta}_t^{\mathrm{MTL}}) &= \left\|\Sigma^{1/2}\left((B^\star\hat{\mathcal{W}}^\top)(\hat{\mathcal{W}}\hat{\mathcal{W}}^\top)^{-1}\hat{W}_t - \beta_2\right)\right\|^2 + \sigma^2\hat{W}_t^\top(\hat{\mathcal{W}}\hat{\mathcal{W}}^\top)^{-1}\hat{W}_t \cdot \mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right] \\ &= \left\|\Sigma^{1/2}\left(B^\star U_r U_r(t) - \beta_2\right)\right\|^2 + \mathrm{o}\left(\|B^\star\|^2\right) + \sigma^2\|U_r(t)\|^2\,\mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right]\cdot(1+\mathrm{o}(1)) \\ &= \left\|\Sigma^{1/2}\left(B^\star U_r U_r(t) - \beta_2\right)\right\|^2 + \frac{\sigma^2}{\rho-1}\|U_r(t)\|^2 + \mathrm{o}\left(\|B^\star\|^2 + \sigma^2\right), \end{aligned}$$

with high probability, where we used Lemma A.3 in the last step. Similar, by Lemma A.3 we have

$$L(\hat{\beta}_t^{\mathrm{MTL}}) = \frac{\sigma^2}{\rho-1}\cdot(1+\mathrm{o}(1)).$$

Combining the above two estimates, we conclude the proof. □

A.4 Proof for the Transfer Learning Setting

Theorem A.10. *Consider t data models $Y_i = X\beta_i + \varepsilon_i$, $i = 1, 2, \cdots, t$, that satisfy Assumption A.2. Moreover, assume that $\max_{1\leqslant i\leqslant t}\|\beta_i\| = \mathrm{O}(1)$, $\sigma^2 = \mathrm{O}(1)$, and*

$$\left\|\left[(B_{t-1}^\star)^\top B_{t-1}^\star\right]^{-1}\right\| = \mathrm{O}(1), \quad B_{t-1}^\star := [\beta_1, \beta_2, \ldots, \beta_{t-1}]. \tag{A.32}$$

Then for any constant $\varepsilon > 0$, we have

$$L(BW_t) = \|\varepsilon_\beta + \sigma^2\varepsilon_{var}^{(1)}\|^2 + \sigma^2\|\varepsilon_{var}^{(2)}\|^2 + \mathrm{O}(p^{-1/2+\varepsilon}) \tag{A.33}$$

with high probability, where

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

Here $\mathcal{M}_1$ and $\mathcal{M}_2$ are $(t-1)\times(t-1)$ diagonal matrices with

$$(\mathcal{M}_1)_{ii} = \frac{1}{\rho_i - 1}\cdot\frac{1}{p}\,\mathrm{Tr}\left(\Sigma_i^{-1}\right), \quad \textit{and} \quad (\mathcal{M}_2)_{ii} = \frac{1}{\rho_i-1}\cdot\frac{1}{p}\,\mathrm{Tr}\left(\Sigma_i^{-1}\Sigma_t\right).$$

Proof of Theorem A.10. For each task i, $1 \leqslant i \leqslant t-1$, we can find that

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

We first calculate $B^\top B$, where recall that $B = [\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{t-1}]$.

For the diagonal entries of $B^\top B$ we can calculate that

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

Using concentration of random vector with i.i.d. centered entries, Lemma D.6, we get that for any deterministic vector β with $\|\beta\| = \mathrm{O}(1)$,

$$\beta^\top(X_i^\top X_i)^{-1}X_i^\top\varepsilon_i \leqslant p^{\varepsilon/2}\cdot\sigma\left[\beta^\top(X_i^\top X_i)^{-1}\beta\right]^{1/2} \leqslant \sigma p^{-1/2+\varepsilon}, \tag{A.34}$$

with high probability, where we used (A.22) to bound

$$\|(X_i^\top X_i)^{-1}\| \lesssim n_i^{-1} \quad \text{whp.} \tag{A.35}$$

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

$$\begin{aligned} \left|\|(X_i^\top X_i)^{-1}X_i^\top\varepsilon_i\|^2 - \sigma^2\,\mathrm{Tr}\left[(X_i^\top X_i)^{-1}\right]\right| &\leqslant p^{\varepsilon/2}\cdot\sigma^2\,\mathrm{Tr}^{1/2}\left\{\left[X_i(X_i^\top X_i)^{-2}X_i^\top\right]^2\right\} \\ &= p^{\varepsilon/2}\cdot\sigma^2\,\mathrm{Tr}^{1/2}\left[(X_i^\top X_i)^{-2}\right] \leqslant \sigma^2 p^{-1/2+\varepsilon}, \end{aligned} \tag{A.36}$$

and for $i \neq j$,

$$\begin{aligned}\left|\varepsilon_j^\top X_j (X_j^\top X_j)^{-1}(X_i^\top X_i)^{-1} X_i^\top \varepsilon_i\right| &\leqslant p^{\varepsilon/2}\cdot \sigma^2 \operatorname{Tr}^{1/2}\left[(X_i^\top X_i)^{-1}(X_j^\top X_j)^{-1}\right] \\ &\leqslant \sigma^2 p^{-1/2+\varepsilon}.\end{aligned} \tag{A.37}$$

With (A.34)-(A.37), we get that with high probability,

$$\hat\beta_i^\top \hat\beta_j = \beta_i^\top \beta_j + \mathrm{O}(\sigma p^{-1/2+\varepsilon}), \quad i \neq j, \tag{A.38}$$

and

$$\begin{aligned}\|\hat\beta_i\|^2 &= \|\beta_i\|^2 + \sigma^2 \operatorname{Tr}\left[(X_i^\top X_i)^{-1}\right] + \mathrm{O}(\sigma p^{-1/2+\varepsilon}) \\ &= \|\beta_i\|^2 + \frac{\sigma^2}{\rho_i - 1}\cdot \frac{1}{p}\operatorname{Tr}\left(\Sigma_i^{-1}\right) + \mathrm{O}(\sigma p^{-1/2+\varepsilon}),\end{aligned} \tag{A.39}$$

where we used Lemma A.3 in the second step.

With (A.38) and (A.39), we obtain that with high probability,

$$B^\top B = (B_{t-1}^\star)^\top B_{t-1}^\star + \sigma^2 \mathcal{M}_1 + \mathrm{O}(\sigma p^{-1/2+\varepsilon}), \tag{A.40}$$

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.32), we have that

$$\|(B^\top B)^{-1}\| = \mathrm{O}(1) \quad \text{whp.}$$

Moreover, using (A.34) we get

$$B^\top \beta = (B_{t-1}^\star)^\top \beta + \mathrm{O}(\sigma p^{-1/2+\varepsilon}), \tag{A.41}$$

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

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

$$\hat W_t = (B^\top X_t^\top X_t B)^{-1} B^\top X_t^\top Y_k = (B^\top X_t^\top X_t B)^{-1} B^\top X_t^\top (X_t\beta_t + \varepsilon_t).$$

We then calculate that

$$\begin{aligned}\mathop{\mathbb{E}}_{\varepsilon_t}\left[\left\|\Sigma_t^{1/2}(B\hat W_t - \beta_t)\right\|^2\right] &= \left\|\Sigma_t^{1/2}\left[\mathrm{Id} - B(B^\top X_t^\top X_t B)^{-1} B^\top X_t^\top X_t\right]\beta_t\right\|^2 \\ &\quad + \sigma^2 \cdot \operatorname{Tr}\left[\Sigma_t (B^\top X_t^\top X_t B)^{-1}\right].\end{aligned} \tag{A.42}$$

Using the restricted isometry property for X_t on the subspace spanned by the column vectors of B, we get that with high probability,

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

Together with (A.35), (A.40) and (A.41), we can simplify the right-hand side of (A.42) as

$$\begin{aligned}&\mathop{\mathbb{E}}_{\varepsilon_t}\left[\left\|\Sigma_t^{1/2}(B\hat W_t - \beta_t)\right\|^2\right] = \left\|\Sigma_t^{1/2}\beta_t - \Sigma_t^{1/2} B(B^\top B)^{-1} B^\top \beta_t\right\|^2 + \mathrm{O}(p^{-1/2+\varepsilon}) \\ &= \left\|\Sigma_t^{1/2}\beta_t - \Sigma_t^{1/2} B\left[(B_{t-1}^\star)^\top B_{t-1}^\star + \sigma^2\mathcal{M}_1\right]^{-1}(B_{t-1}^\star)^\top \beta_t\right\|^2 + \mathrm{O}(p^{-1/2+\varepsilon}).\end{aligned} \tag{A.43}$$

Then as in (A.40), we can show by concentration and Lemma A.3 that

$$B^\top \Sigma_t B = (B_{t-1}^\star)^\top \Sigma_t B_{t-1}^\star + \sigma^2 \mathcal{M}_2 + \mathrm{O}(\sigma p^{-1/2+\varepsilon}). \tag{A.44}$$

Combining (A.40), (A.41) and (A.44), we can further take the expectation and simplify (A.43) as

$$\begin{aligned}&\mathop{\mathbb{E}}_{\varepsilon_i, \forall 1\leqslant i\leqslant t}\left[\left\|\Sigma_t^{1/2}(B\hat W_t - \beta_t)\right\|^2\right] \\ &= \left\|\Sigma_t^{1/2}\beta_t\right\|^2 - 2\beta_t^\top \Sigma_t B_{t-1}^\star \widetilde\beta_t + \widetilde\beta_t^\top\left((B_{t-1}^\star)^\top \Sigma_t B_{t-1}^\star + \sigma^2\mathcal{M}_2\right)\widetilde\beta_t + \mathrm{O}(p^{-1/2+\varepsilon}).\end{aligned}$$

where we abbreviated

$$\widetilde\beta_t := \left[(B_{t-1}^\star)^\top B_{t-1}^\star + \sigma^2\mathcal{M}_1\right]^{-1}(B_{t-1}^\star)^\top \beta_t.$$

Finally, rearranging terms and performing some basic calculations we can conclude the proof. □

B Supplementary Materials for the Theoretical Implications

B.1 Supplement to Section 3.1

We define the function

$$\begin{aligned} val(v) = \frac{\rho_1}{\rho_2}\left[d^2 + (v-1)^2\kappa^2\right] \cdot \mathrm{Tr}\left[(v^2 X_1^\top X_1 + X_2^\top X_2)^{-2}(X_2^\top X_2)^2\right] \\ + v^2\left[d^2 + (v-1)^2\kappa^2\right] \cdot \mathrm{Tr}\left[(v^2 X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ + \left(\frac{\rho_1}{\rho_2}v^2 + 1\right)\sigma^2 \cdot \mathrm{Tr}\left[(v^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\right]. \end{aligned}$$

For the isotropic model with $\sigma_1^2 = \sigma_2^2 = \sigma^2$, using concentration for random vectors with i.i.d. entries, Lemma D.6, we can obtain that $val(\hat{B}; w_1, w_2) = val(v) \cdot \big(1 + \mathrm{O}(p^{-1/2+\varepsilon})\big)$. Hence the validation loss in (A.6) reduces to

$$\widetilde{f}(\hat{B}; v) = \left[N_2 \cdot val(v) + (N_1\sigma_1^2 + N_2\sigma_2^2)\right] \cdot \Big(1 + \mathrm{O}(p^{-(1-\varepsilon_0)/2+\varepsilon})\Big) \tag{B.1}$$

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

- First, we show that $\hat{w}$ is close to 1, and then (B.1) implies that $\hat{v}$ is also close to 1.
- Second, we plug $\hat{v}$ back into $L(\hat{\beta}_2^{\mathrm{MTL}})$ and use Lemma A.7 to show the result.

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

Lemma B.1. *For the isotropic model, the minimizer for $val(v)$ satisfies*

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

for some constant $C > 0$.

Proof. To be consistent with the notation $\hat{w}$, we shall change the name of the argument to w in the 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

$$\begin{aligned} val(w) = \frac{\rho_1}{\rho_2}\left[\frac{d^2}{w^4} + \frac{(w-1)^2}{w^4}\kappa^2\right] \cdot \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}(X_2^\top X_2)^2\right] \\ + \left[\frac{d^2}{w^2} + \frac{(w-1)^2}{w^2}\kappa^2\right] \cdot \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ + \frac{\rho_1}{\rho_2}\sigma^2 \cdot \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-1}\right] + \sigma^2 \cdot \mathrm{Tr}\left[(w^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\right]. \end{aligned}$$

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

$$\begin{aligned} val'(w) \geqslant \frac{\rho_1}{\rho_2}\left[\frac{2(w-1)(2-w)}{w^5}\kappa^2 - \frac{4d^2}{w^5}\right] \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}(X_2^\top X_2)^2\right] \\ + \left[\frac{2\,(w-1)}{w^3}\kappa^2 - \frac{2d^2}{w^3}\right] \cdot \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ - 2\frac{\sigma^2}{w^3} \cdot \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}X_1^\top X_1\right] = \mathrm{Tr}\left[(X_1^\top X_1 + w^{-2}X_2^\top X_2)^{-2}\mathcal{A}\right], \end{aligned}$$

where the matrix $\mathcal{A}$ is

$$\mathcal{A} := \frac{\rho_1}{\rho_2}\left[\frac{2(w-1)(2-w)}{w^5}\kappa^2 - \frac{4d^2}{w^5}\right](X_2^\top X_2)^2 + \left[\frac{2\,(w-1)}{w^3}\kappa^2 - \frac{2d^2}{w^3}\right](X_1^\top X_1)^2 - 2\frac{\sigma^2}{w^3}X_1^\top X_1.$$

Using the estimate (A.22), we get that $\mathcal{A}$ is lower bounded as

$$\mathcal{A} \succeq -\frac{4d^2}{w^5} n_1 n_2 (\alpha_+(\rho_2) + \mathrm{o}(1))^2 + \left[\frac{2\,(w-1)}{w^3} \kappa^2 - \frac{2d^2}{w^3} \right] n_1^2 (\alpha_-(\rho_1) - \mathrm{o}(1))^2 - 2\frac{\sigma^2}{w^3} n_1 (\alpha_+(\rho_1) + \mathrm{o}(1)) \succ 0,$$

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

$$\begin{aligned} val'(w) &\leqslant \frac{\rho_1}{\rho_2} \left[2\,(w-1)\,\kappa^2 \right] \cdot \mathrm{Tr}\left[(w^2 X_1^\top X_1 + X_2^\top X_2)^{-2} (X_2^\top X_2)^2 \right] \\ &\quad + \left[2wd^2 + 2w\,(w-1)\,(2w-1)\kappa^2 \right] \cdot \mathrm{Tr}\left[(w^2 X_1^\top X_1 + X_2^\top X_2)^{-2} (X_1^\top X_1)^2 \right] \\ &\quad + \frac{\rho_1}{\rho_2} (2w\sigma^2) \cdot \mathrm{Tr}\left[(w^2 X_1^\top X_1 + X_2^\top X_2)^{-2} X_2^\top X_2 \right] \\ &= \frac{\rho_1}{\rho_2} \mathrm{Tr}\left[(w^2 X_1^\top X_1 + X_2^\top X_2)^{-1} \mathcal{B} \right], \end{aligned} \tag{B.3}$$

where the matrix $\mathcal{B}$ is

$$\mathcal{B} = 2\,(w-1)\,\kappa^2 (X_2^\top X_2)^2 + \frac{\rho_2}{\rho_1} \left[2wd^2 + 2w\,(w-1)\,(2w-1)\kappa^2 \right] (X_1^\top X_1)^2 + 2w\sigma^2 X_2^\top X_2.$$

Using the estimate (A.22), 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 $\hat{w} \geqslant w_2$.

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. □

For the rest of this section, we choose the parameters that satisfy the following relations:

$$pd^2 \sim \sigma^2 \sim 1, \quad p^{-1+c_0}\sigma^2 \leqslant \kappa^2 \leqslant p^{-\varepsilon_0 - c_0}\sigma^2, \tag{B.4}$$

for some small constant $c_0 > 0$. We will explain below why we make this choice. Before that, we first show the following estimate on the optimizer $\hat{v}$: with high probability,

$$|\hat{v} - 1| = \mathrm{O}\left(\mathcal{E}\right), \quad \mathcal{E} := \frac{d^2}{\kappa^2} + \frac{\sigma^2}{p\kappa^2} + p^{-1/2+\varepsilon_0/2+2\varepsilon}. \tag{B.5}$$

In fact, from the proof of Lemma B.1 above, one can check that if $C\mathcal{E} \leqslant |w - \hat{w}| \leqslant 2C\mathcal{E}$ for a large enough constant $C > 1$, then $|val'(w)| \gtrsim pd^2$. Moreover, under the choice (B.4) we have

$$val(w) = \mathrm{O}(pd^2), \quad \text{for} \quad |w - \hat{w}| \leqslant 2C\mathcal{E}.$$

Thus we obtain that for $|w - \hat{w}| \geqslant 2C\mathcal{E}$,

$$|val(w) - val(\hat{w})| \geqslant |val(w) - \min(val(w_1), val(w_2))| \gtrsim pd^2\mathcal{E} \gtrsim \mathcal{E} \cdot val(\hat{w}),$$

which leads to $\widetilde{f}(\hat{B}; w) > \widetilde{f}(\hat{B}; \hat{w})$ whp by (B.1). Thus w cannot be a minimizer of $\widetilde{f}(\hat{B}; v)$, and we must have $|\hat{v} - \hat{w}| \leqslant 2C\mathcal{E}$. Together with (B.2), we conclude (B.5).

Inserting (B.5) into (A.4) and applying Lemma D.6 to $(\beta_1 - \hat{v}\beta_s)$ again, we get whp,

$$\begin{aligned} L(\hat{\beta}_t^{\text{MTL}}) = (1 + \mathrm{O}(\mathcal{E})) \cdot \left[d^2 + \mathrm{O}\left(\mathcal{E}^2\kappa^2\right)\right] \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ + (1 + \mathrm{O}(\mathcal{E})) \cdot \sigma^2 \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right]. \end{aligned} \tag{B.6}$$

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.22), we see that

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

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 $p^{-1+\varepsilon_0+4\varepsilon}\kappa^2 \ll d^2 \ll \kappa^2$. The above considerations lead to the choices of parameters in (B.4). Moreover, under (B.4) we can simplify (B.6) to

$$\begin{aligned} L(\hat{\beta}_2^{\text{MTL}}) = (1 + \mathrm{O}(n^{-\varepsilon})) \cdot d^2 \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ + (1 + \mathrm{O}(n^{-\varepsilon})) \cdot \sigma^2 \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right] \end{aligned} \tag{B.7}$$

whp for some constant $\varepsilon > 0$.

With (B.7) and Lemma A.7, we can prove Proposition 3.1, 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 throughout the rest of this section.

Proof of Proposition 3.1. In the setting of Proposition 3.1, we have $M = \Sigma_1^{1/2}\Sigma_2^{-1/2} = \mathrm{Id}$. Then solving equations (A.12) and (A.13) with $\hat{\lambda}_i = 1$, we get that

$$a_1 = \frac{\rho_1(\rho_1 + \rho_2 - 1)}{(\rho_1 + \rho_2)^2}, \quad a_2 = \frac{\rho_2(\rho_1 + \rho_2 - 1)}{(\rho_1 + \rho_2)^2}, \tag{B.8}$$

$$a_3 = \frac{\rho_2}{(\rho_1 + \rho_2)(\rho_1 + \rho_2 - 1)}, \quad a_4 = \frac{\rho_1}{(\rho_1 + \rho_2)(\rho_1 + \rho_2 - 1)}. \tag{B.9}$$

Using Lemma A.3 and Lemma A.4, we can track the reduction of variance from $\hat{\beta}_2^{\text{MTL}}$ to $\hat{\beta}_2^{\text{STL}}$ as

$$\begin{aligned} \delta_{\text{var}} &:= \sigma^2 \operatorname{Tr}\left[(X_2^\top X_2)^{-1}\right] - (1 + \mathrm{O}(n^{-\varepsilon})) \cdot \sigma^2 \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right] \\ &= \Delta_{\text{var}} \cdot (1 + \mathrm{O}(n^{-\varepsilon})) \end{aligned} \tag{B.10}$$

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

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

we can get from Lemma A.7 (or rather the proof of it) that

$$\alpha_-^2(\rho_1) - \mathrm{o}(1) \leqslant \frac{\delta_{\text{bias}}}{\Delta_{\text{bias}}} \leqslant \alpha_+^2(\rho_1) + \mathrm{o}(1), \tag{B.11}$$

where

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

Note that

$$L(\hat{\beta}_2^{\text{STL}}) - L(\hat{\beta}_2^{\text{MTL}}) = \delta_{\text{var}} - \delta_{\text{bias}}. \tag{B.12}$$

Then we can track its sign using (B.10) and (B.11).

Positive transfer. With (B.10) and (B.11), we conclude that if

$$pd^2 \cdot \frac{\rho_1^2(\rho_1 + \rho_2)}{(\rho_1 + \rho_2 - 1)^3} \cdot \left(\alpha_+^2(\rho_1) + \mathrm{o}(1)\right) < \sigma^2 \cdot \frac{\rho_1}{(\rho_2 - 1)(\rho_1 + \rho_2 - 1)}, \tag{B.13}$$

we have that $\delta_{\text{var}} > \delta_{\text{bias}}$, which implies $L(\hat{\beta}_2^{\text{MTL}}) < L(\hat{\beta}_2^{\text{STL}})$. We can simplify (B.13) to

$$\frac{pd^2}{\sigma^2} < \Phi(\rho_1, \rho_2) \cdot \left(\alpha_+^2(\rho_1) + \mathrm{o}(1)\right)^{-1}, \tag{B.14}$$

Since $\Psi(\beta_1, \beta_2) = pd^2/\sigma^2$, it gives the first statement of Proposition 3.1.

Negative transfer. On the other hand, if

$$pd^2 \cdot \frac{\rho_1^2(\rho_1 + \rho_2)}{(\rho_1 + \rho_2 - 1)^3} \cdot \left(\alpha_-^2(\rho_1) - \mathrm{o}(1)\right) > \sigma^2 \cdot \frac{\rho_1}{(\rho_2 - 1)(\rho_1 + \rho_2 - 1)}, \tag{B.15}$$

we have that $\delta_{\text{var}} < \delta_{\text{bias}}$, which implies $L(\hat{\beta}_2^{\text{MTL}}) > L(\hat{\beta}_2^{\text{STL}})$. We can simplify (B.15) to

$$\Psi(\beta_1, \beta_2) = \frac{pd^2}{\sigma^2} > \Phi(\rho_1, \rho_2) \cdot \left(\alpha_-^2(\rho_1) - \mathrm{o}(1)\right)^{-1}, \tag{B.16}$$

which gives the second statement of Proposition 3.1. □

B.2 Supplement to Section 3.2

We first prove Proposition 3.2, which describes the effect of source/task data ratio on the information transfer.

Proof of Proposition 3.2. Following the above proof of Proposition 3.1, we see that $L(\hat{\beta}_2^{\text{MTL}}) < L(\hat{\beta}_2^{\text{STL}})$ whp if (B.14) holds, while $L(\hat{\beta}_2^{\text{MTL}}) > L(\hat{\beta}_2^{\text{STL}})$ whp if (B.16) holds.

We first explain the meaning of the condition

$$\Psi(\beta_1, \beta_2) > 2/(\rho_2 - 1). \tag{B.17}$$

Notice that the function

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

is strictly decreasing with respect to ρ_1 as long as $\rho_2 > 2$, and $\Phi(\rho_1, \rho_2)$ converges to $(\rho_2 - 1)^{-1}$ as $\rho_1 \to \infty$. Moreover, we notice that $\left(\alpha_-^2(\rho_1) - \mathrm{o}(1)\right)^{-1} < 2$ for $\rho_1 > 40$. Hence (B.17) implies that (B.16) holds for all large enough ρ_1. The transition from positive transfer when ρ_1 is small to negative transfer when ρ_1 is large is described by the two bounds in Proposition 3.2.

The two bounds follows directly from (B.14) and (B.16). We will use the following trivial inequalities

$$\frac{(\rho_2 - 1)\rho_1}{\rho_1 + \rho_2 - 2} \cdot \left(1 - \frac{1}{(\rho_1 + \rho_2 - 2)^2}\right) \leqslant \Phi(\rho_1, \rho_2) \leqslant \frac{(\rho_2 - 1)\rho_1}{\rho_1 + \rho_2 - 2}. \tag{B.18}$$

Positive transfer. With (B.18), we see that (B.14) is implied by the following inequality:

$$\Psi(\beta_1, \beta_2) \cdot \frac{(\rho_2 - 1)\rho_1}{\rho_1 + \rho_2 - 2} < \gamma_+^{-1}. \tag{B.19}$$

then we can solve (B.19) to get

$$\rho_1 < \frac{\rho_2 - 2}{\Psi(\beta_1, \beta_2) \cdot \gamma_+(\rho_2 - 1) - 1}. \tag{B.20}$$

This gives the first statement of Proposition 3.2.

Note that if we require the RHS of (B.20) to be larger than 40, that is, (B.20) is not a null condition. Then together with (B.17), we get

$$\rho_2 - 2 > (2\gamma_+ - 1)\rho_1.$$

Plugging into $\rho_1 > 40$, we get $\rho_2 \geqslant 106$. This gives a constraint on ρ_2.

Negative transfer. With (B.18), we see that (B.16) is implied by the following inequality:

$$\Psi(\beta_1, \beta_2) \cdot \frac{(\rho_2 - 1)\rho_1}{\rho_1 + \rho_2 - 2}\left(1 - \frac{1}{(\rho_1 + \rho_2 - 2)^2}\right) > \Psi(\beta_1, \beta_2) \cdot \frac{(\rho_2 - 1.5)\rho_1}{\rho_1 + \rho_2 - 2} > \gamma_-^{-1}. \tag{B.21}$$

where we used $(1-(\rho_1+\rho_2-2)^{-2})(\rho_2-1)>\rho_2-1.5$ for $\rho_1>40$ and $\rho_2>110$. Then we can solve (B.21) to get

$$\rho_1 > \frac{(\rho_2-2)\sigma^2}{\Psi(\beta_1,\beta_2)\cdot\gamma_-(\rho_2-1.5)-1}, \tag{B.22}$$

which gives the second statement of Proposition 3.2. We remark that condition (B.17) implies $\Psi(\beta_1,\beta_2)\cdot\gamma_-(\rho_2-1.5)>1$, so (B.22) does not give a trivial bound. □

Next we state Proposition B.2, which gives precise upper and lower bounds on the data efficiency ratio for taskonomy.

Proposition B.2 (Labeled data efficiency)**.** *In the isotropic model, assume that* $\rho_1,\rho_2>10$ *and* $\Psi(\beta_1,\beta_2)<(5\rho_1)^{-1}+(5\rho_2)^{-1}$. *Then the data efficiency ratio* $x^\star$ *satisfies*

$$x_l \leqslant x^\star \leqslant \frac{1}{\rho_1+\rho_2}\left(\frac{2}{(\rho_1-1)^{-1}+(\rho_2-1)^{-1}-5\Psi(\beta_1,\beta_2)}+1\right),$$

where we denoted

$$x_l := \frac{1}{\rho_1+\rho_2}\left(\frac{2}{(\rho_1-1)^{-1}+(\rho_2-1)^{-1}}+1\right).$$

Proof of Proposition B.2. Suppose we have reduced number of datapoints—xn_1 for task 1 and xn_2 for task 2 with $n_1=\rho_1 p$ and $n_2=\rho_2 p$. Then all the results in the proof of Proposition 3.1 still hold, except that we need to replace (ρ_1,ρ_2) with $(x\rho_1,x\rho_2)$. More precisely, we have

$$a_1=\frac{\rho_1(x\rho_1+x\rho_2-1)}{x(\rho_1+\rho_2)^2},\quad a_2=\frac{\rho_2(x\rho_1+\alpha\rho_2-1)}{x(\rho_1+\rho_2)^2},$$
$$a_3=\frac{\rho_2}{(\rho_1+\rho_2)(x\rho_1+x\rho_2-1)},\quad a_4=\frac{\rho_1}{(\rho_1+\rho_2)(x\rho_1+x\rho_2-1)}.$$

Moreover, with high probability,

$$L_i(\hat\beta_i^{\mathrm{MTL}}(x))=\frac{\sigma^2}{x(\rho_1+\rho_2)-1}\left(1+\mathrm{o}(1)\right)+\delta_{\mathrm{bias}}^{(i)},\quad i=1,2. \tag{B.23}$$

Here the model shift biases $\delta_{\mathrm{bias}}^{(i)}$ satisfy that

$$\alpha_-^2(\alpha\rho_i)-\mathrm{o}(1)\leqslant\delta_{\mathrm{bias}}^{(i)}/\Delta_{\mathrm{bias}}^{(i)}\leqslant\alpha_+^2(\alpha\rho_i)+\mathrm{o}(1),\quad i=1,2,$$

where $\Delta_{\mathrm{bias}}^{(i)}$ are defined as

$$\Delta_{\mathrm{bias}}^{(i)}:=pd^2\frac{(x\rho_i)^2\cdot x(\rho_1+\rho_2)}{[x(\rho_1+\rho_2)-1]^3},\quad i=1,2,.$$

On the other hand, using Lemma A.3 we have whp,

$$L_i(\hat\beta_i^{\mathrm{STL}})=\frac{\sigma^2}{\rho_i-1}\left(1+\mathrm{o}(1)\right),\quad i=1,2. \tag{B.24}$$

Comparing (B.23) and (B.24), we immediately obtain the lower bound $x^*\geqslant x_l$. In fact, one can see that if $x<x_l$, then we have

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

that is, $L_1(\hat\beta(\alpha))+L_2(\hat\beta(\alpha))$ is larger than $L_1(\hat\beta_t^{\mathrm{STL}})+L_2(\hat\beta_t^{\mathrm{STL}})$ even if we do not take into account the model shift bias terms $\delta_{\mathrm{bias}}^{(i)}$.

Then we try to obtain an upper bound on x^*. In the following discussions, we only consider x such that $x\geqslant\alpha_l$. In particular, we have $x(\rho_1+\rho_2)\geqslant x_l(\rho_1+\rho_2)\geqslant\min(\rho_1,\rho_2)$.

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

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

We rewrite the inequality as

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

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

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

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

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 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. □

B.3 Supplement to Section 3.3

Proof of Proposition 3.4. Let

$$M_0 := \arg\min_{M \in \mathcal{S}_\mu} g(M).$$

We now calculate $g(M_0)$. With the same arguments as in Lemma B.1 we can show that (B.5) holds. Moreover, if the parameters are chosen such that $p^c\kappa^2 \leqslant \sigma^2 \leqslant p^{1-c}\kappa^2$ (recall (B.4)), we can simplify

$$g(M_0) = (1 + \mathrm{O}(p^{-\varepsilon})) \cdot \sigma^2 \operatorname{Tr}\left[\Sigma_2 (X_1^\top X_1 + X_2^\top X_2)^{-1}\right],$$

with high probability for any constant $\varepsilon > 0$. 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 Lemma A.4, we can obtain that with high probability,

$$g(M_0) = \frac{\sigma^2}{\rho_1 + \rho_2} \cdot \frac{1}{p} \operatorname{Tr}\left(\frac{1}{a_1(M_0) \cdot M_0^\top M_0 + a_2(M_0)}\right) \cdot \left(1 + \mathrm{O}(p^{-\varepsilon})\right). \tag{B.29}$$

From equation (A.9), it is easy to obtain the following estimates on $a_1(M)$ and $a_2(M)$ for any $M \in \mathcal{S}_\mu$:

$$\frac{\rho_1 - 1}{\rho_1 + \rho_2} < a_1(M) < \frac{\rho_1 + \rho_2 - 1}{\rho_1 + \rho_2}, \quad a_2(M) < \frac{\rho_2}{\rho_1 + \rho_2}. \tag{B.30}$$

Inserting (B.30) into (B.29) and using $\lambda(M_0^\top M_0) \geqslant \mu_{\min}^2$, we obtain that with high probability,

$$\left(1 + \frac{\rho_2}{(\rho_1 - 1)\mu_{\min}^2}\right)^{-1} \widetilde{g}(M_0) \cdot \left(1 - \mathrm{O}(p^{-\varepsilon})\right) \leqslant g(M_0) \leqslant \widetilde{g}(M_0) \cdot \left(1 + \mathrm{O}(p^{-\varepsilon})\right), \tag{B.31}$$

where

$$\widetilde{g}(M_0) := \frac{\sigma^2}{(\rho_1 + \rho_2) a_1(M_0)} \cdot \frac{1}{p} \operatorname{Tr}\left(\frac{1}{M_0^\top M_0}\right).$$

By 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

$$\widetilde{g}(M_0) \leqslant \frac{\sigma^2}{\mu^2(\rho_1 + \rho_2) a_1(M_0)}. \tag{B.32}$$

On the other hand, when $M = \mu\,\mathrm{Id}$, applying Lemma A.4 we can obtain that with high probability,

$$\begin{aligned} g(\mu\,\mathrm{Id}) &= \frac{\sigma^2}{\rho_1 + \rho_2} \cdot \frac{1}{p} \operatorname{Tr}\left(\frac{1}{\mu^2 a_1(\mu\,\mathrm{Id}) + a_2(\mu\,\mathrm{Id})}\right) \cdot \left(1 + \mathrm{O}(p^{-\varepsilon})\right) \\ &\leqslant \frac{\sigma^2}{\mu^2(\rho_1 + \rho_2) a_1(\mu\,\mathrm{Id})}. \end{aligned} \tag{B.33}$$

Combining (B.30), (B.31), (B.32) and (B.33), we conclude the proof. □

We observe the following two phases as we increase n_1/p in Figure 1c. 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 $\Sigma_1 = \Sigma_2$. We provide a theoretical justification below.

Theoretical justification of Example 3.3. We denote the test error as $L_\lambda(\hat{\beta}_t^{\mathrm{MTL}})$ in the setting where M has $p/2$ singular values that are equal to λ and $p/2$ singular values that are equal to $1/\lambda$. Then equations in (A.9) become

$$a_1 + a_2 = 1 - \frac{1}{\rho_1 + \rho_2}, \;\; a_1 + \frac{1}{2(\rho_1 + \rho_2)} \cdot \left(\frac{a_1}{a_1 + \lambda^2 a_2} + \frac{a_1}{a_1 + \frac{a_2}{\lambda^2}}\right) = \frac{\rho_1}{\rho_1 + \rho_2}. \tag{B.34}$$

After solving these, with (B.29) we get that whp,

$$L_\lambda(\hat{\beta}_t^{\text{MTL}}) = \frac{\sigma^2}{2(\rho_1+\rho_2)}(1+\mathrm{O}(p^{-\varepsilon}))\cdot g(\lambda), \quad g(\lambda) := \frac{1}{\frac{a_1}{\lambda^2}+a_2} + \frac{1}{a_1\lambda^2+a_2}. \tag{B.35}$$

We now study the behavior of g as λ changes. We abbreviate $\gamma := (\rho_1+\rho_2)^{-1}$. Then with (B.34), we can rewrite

$$g(\lambda) = \frac{1}{\frac{a_1}{\lambda^2}+(1-\gamma-a_1)} + \frac{1}{a_1\lambda^2+(1-\gamma-a_1)}.$$

We can compute that

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

From this expressions, we observe the following behaviors.

(i) If $n_1 > n_2$, we have $a_1 > (1-\gamma)/2$ (because $a_1 > a_2$ as observed from the equation (B.34)). Hence $g(\lambda) > g(1)$, which gives $L_\lambda(\hat{\beta}_t^{\text{MTL}}) > L_1(\hat{\beta}_t^{\text{MTL}})$.

(ii) If $n_1 < n_2$, we have $a_1 < (1-\gamma)/2$. Hence $g(\lambda) < g(1)$, which gives $L_\lambda(\hat{\beta}_t^{\text{MTL}}) < L_1(\hat{\beta}_t^{\text{MTL}})$.

(iii) If $n_1 = n_2$, we have $g(\lambda) = g(1) = 2/(1-\gamma)$, which explains why the curves in Figure 1 (c) all cross at the point $n_1 = n_2$.

These justify the observations in Figure 1 (c), □

B.4 Labeled data de-noising

Then we show Proposition B.3, which gives a transition threshold with respect to the difference between the noise levels of the two tasks.

Proposition B.3 (Labeled data de-noising)**.** *In the isotropic model, assume that* $\rho_1 > 40$ *and* $\mathbb{E}\left[\|\beta_1-\beta_2\|^2\right] < \frac{1}{2}\sigma_2^2\cdot\Phi(\rho_1,\rho_2)$*. We derive the following transition as a parameter of* σ_1^2*:*

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

As a corollary, if $\sigma_1^2 \leqslant \sigma_2^2$, then we always get positive transfer. The proof of Proposition B.3 is similar to Proposition 3.1.

Proof of Proposition B.3. In the setting of Proposition B.3, 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

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

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 $L(\hat{\beta}_t^{\text{MTL}})$ as

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

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

In the current setting, we can also show that (B.5) 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.4), $L(\hat{\beta}_t^{\text{MTL}})$ can be simplified as in (B.7):

$$\begin{aligned} L(\hat{\beta}_t^{\text{MTL}}) = \; & (1 + \mathrm{O}(n^{-\varepsilon})) \cdot d^2 \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2} (X_1^\top X_1)^2\right] \\ & + (1 + \mathrm{O}(n^{-\varepsilon})) \cdot \sigma_2^2 \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right] \\ & + (1 + \mathrm{O}(n^{-\varepsilon})) \cdot (\sigma_1^2 - \sigma_2^2) \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2} X_1^\top X_1\right]. \end{aligned} \tag{B.36}$$

Then we write

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

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.10) but with σ^2 replaced with σ_2^2,

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

satisfies (B.11), 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.7: we first replace $X_1^\top X_1$ with $n_1 \operatorname{Id}$ up to some error using (A.22), and then apply Lemma A.5 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 $\delta_{\text{var}}^{(2)}$:

$$\alpha_-(\rho_1) - \mathrm{o}(1) \leqslant \frac{\delta_{\text{var}}^{(2)}}{\Delta_{\text{var}}^{(2)}} \leqslant \alpha_+(\rho_1) + \mathrm{o}(1), \tag{B.37}$$

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 $\delta_{\text{bias}} + \delta_{\text{var}}^{(2)}$. 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 $\delta_{\text{var}} > \delta_{\text{bias}}$ by Proposition 3.1. Hence if $\sigma_1^2 \leqslant \sigma_2^2$, then $\delta_{\text{var}}^{(2)} < 0$ and we always have $\delta_{\text{var}} > \delta_{\text{bias}} + \delta_{\text{var}}^{(2)}$, which gives $L(\hat{\beta}_t^{\text{MTL}}) < L(\hat{\beta}_t^{\text{STL}})$. It remains to consider the case $\sigma_1^2 \geqslant \sigma_2^2$.

Positive transfer. By (B.10), (B.11) and (B.37) above, if

$$\begin{aligned} & \sigma_2^2 \cdot \frac{\rho_1}{(\rho_2 - 1)(\rho_1 + \rho_2 - 1)} \cdot (1 - \mathrm{o}(1)) \\ & > pd^2 \cdot \frac{\rho_1^2(\rho_1 + \rho_2)}{(\rho_1 + \rho_2 - 1)^3} \left(1 + \sqrt{\frac{1}{\rho_1}}\right)^4 + (\sigma_1^2 - \sigma_2^2) \cdot \frac{\rho_1(\rho_1 + \rho_2)}{(\rho_1 + \rho_2 - 1)^3} \left(1 + \sqrt{\frac{1}{\rho_1}}\right)^2, \end{aligned} \tag{B.38}$$

then we have $\delta_{\text{var}} > \delta_{\text{bias}} + \delta_{\text{var}}^{(2)}$ whp, which gives $L(\hat{\beta}_t^{\text{MTL}}) < L(\hat{\beta}_t^{\text{STL}})$. We can solve (B.38) 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 B.3 for positive transfer.

Negative transfer. On the other hand, if

$$\begin{aligned}&\sigma_2^2 \cdot \frac{\rho_1}{(\rho_2-1)(\rho_1+\rho_2-1)} \cdot (1+\mathrm{o}(1)) \\ &< pd^2 \cdot \frac{\rho_1^2(\rho_1+\rho_2)}{(\rho_1+\rho_2-1)^3}\left(1-\sqrt{\frac{1}{\rho_1}}\right)^4 + (\sigma_1^2-\sigma_2^2)\cdot\frac{\rho_1(\rho_1+\rho_2)}{(\rho_1+\rho_2-1)^3}\left(1-\sqrt{\frac{1}{\rho_1}}\right)^2,\end{aligned} \tag{B.39}$$

then we have $\delta_{\text{var}} < \delta_{\text{bias}} + \delta_{\text{var}}^{(2)}$ whp, which gives $L(\hat{\beta}_t^{\text{MTL}}) > L(\hat{\beta}_t^{\text{STL}})$. We can solve (B.39) 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 B.3 for negative transfer. □

C Proof of Lemma A.4 and Lemma A.5

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 [38] and subsequently used in many works on random matrix theory, such as [33, 39, 40, 41, 42, 28]. 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 [33])**.** *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 [43, 44, 45] and references therein. It has been proved to be useful in studying the local laws of random matrices of the Gram type [46, 47, 28, 48]. 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

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

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

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

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

Suppose they satisfy the system of equations

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

for some (random) errors satisfying

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Now $\widehat{Z}_1$ and $\widehat{Z}_2$ satisfy the assumptions in Theorem 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.4 and Lemma A.5.

Proof of Lemma A.4. In the setting of Lemma A.4, 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.4 the variances of the entries of $Z_{1,2}$ are equal to 1,

while in (C.1) they are taken to be n^{-1}. As in (C.6), we assume that $M := \widetilde{\Sigma}_1^{1/2}\Sigma_2^{-1/2}$ has singular value decomposition

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

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

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

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

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

where by (C.18),

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

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

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

with probability $1 - \mathrm{o}(1)$. This concludes (A.8) if we rename $r_1 b_2 \to a_1$ and $r_2 b_3 \to a_2$. For (A.8), 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.12). This gives (**??**) by (A.8). □

Proof of Lemma A.5. In the setting of Lemma A.5, 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

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

where we used $M^\top M = V\Lambda^2 V^\top$ in the second step. This concludes Lemma A.5. □

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

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

In this case, the problem is reduced to proving the anisotropic local law for G with $U = \mathrm{Id}$ and $V = \mathrm{Id}$, such that the entries of Y_1 and Y_2 are independent. This can be handled using the standard resolvent methods as in e.g. [33, 49, 29]. To go from the Gaussian case to the general X case, we will adopt a continuous self-consistent comparison argument developed in [28].

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 $1 - \mathrm{o}(1)$ [36]. 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 [33, Theorem 2.10], although the authors considered the case with $q \prec n^{-1/2}$ only. The results for larger q follows from [50, 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 [50]. 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)*)*

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

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

Then we record the following resolvent identities.

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

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

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

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

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

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

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

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

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

Proof. All these identities can be proved directly using Schur's complement formula. The reader can also refer to, for example, [28, Lemma 4.4]. □

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

Lemma D.6 (Lemma 3.8 of [51])**.** *Let (x_i), (y_j) be independent families of centered and independent random variables, and (A_i), (B_{ij}) be families of deterministic complex numbers. Suppose the entries x_i, y_j have variance at most n^{-1} and satisfy the bounded support condition (C.8) with $q \leqslant n^{-\phi}$ for some constant $\phi > 0$. Then we have the following bound:*

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

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

where $B_d := \max_i |B_{ii}|$ and $B_o := \max_{i\neq j} |B_{ij}|$. Moreover, if all the moments of $\sqrt{n}x_i$ and $\sqrt{n}y_j$ exist in the sense of (A.5)*, then we have stronger bounds*

$$\Big|\sum_i A_i x_i\Big| \prec \frac{1}{\sqrt{n}}\Big(\sum_i |A_i|^2\Big)^{1/2}, \quad \Big|\sum_{i,j} x_i B_{ij} y_j\Big| \prec \frac{1}{n}\Big(\sum_{i\neq j} |B_{ij}|^2\Big)^{1/2},$$

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

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

$$W = (\Lambda Z_1^\top, Z_2^\top). \tag{D.12}$$

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

$$\max_{\mathfrak{a},\mathfrak{b}} |(G_0)_{\mathfrak{a}\mathfrak{b}}(z) - \Pi_{\mathfrak{a}\mathfrak{b}}(z)| \prec n^{-1/2}. \tag{D.13}$$

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.12).

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 [51, Section 3] and [28, Section 5], we introduce the Z variables

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

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

$$\begin{aligned} Z_i = (\mathbb{E}_i - 1)\left(WG^{(i)}W^\top\right)_{ii} &= \sigma_i^2 \sum_{\mu,\nu\in\mathcal{I}_2} G^{(i)}_{\mu\nu}\left(\frac{1}{n}\delta_{\mu\nu} - z_{\mu i}z_{\nu i}\right) \\ &\quad + \sum_{\mu,\nu\in\mathcal{I}_3} G^{(i)}_{\mu\nu}\left(\frac{1}{n}\delta_{\mu\nu} - z_{\mu i}z_{\nu i}\right), \quad i \in \mathcal{I}_1, \end{aligned} \tag{D.14}$$

and

$$\begin{aligned} Z_\mu &= (\mathbb{E}_\mu - 1)\left(W^\top G^{(\mu)} W\right)_{\mu\mu} = \sum_{i,j\in\mathcal{I}_1} \sigma_i\sigma_j G^{(\mu)}_{ij}\left(\frac{1}{n}\delta_{ij} - z_{\mu i}z_{\mu j}\right), \quad \mu\in\mathcal{I}_2, \\ Z_\mu &= (\mathbb{E}_\mu - 1)\left(W^\top G^{(\mu)} W\right)_{\mu\mu} = \sum_{i,j\in\mathcal{I}_1} G^{(\mu)}_{ij}\left(\frac{1}{n}\delta_{ij} - z_{\mu i}z_{\mu j}\right), \quad \mu\in\mathcal{I}_3. \end{aligned} \tag{D.15}$$

For simplicity, we introduce the following random error

$$\Lambda_o := \max_{\mathfrak{a}\neq\mathfrak{b}} \left|G^{-1}_{\mathfrak{a}\mathfrak{a}} G_{\mathfrak{a}\mathfrak{b}}\right|. \tag{D.16}$$

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 $z\in\mathbf{D}$:*

$$\Lambda_o + \max_{\mathfrak{a}\in\mathcal{I}} |Z_\mathfrak{a}| \prec n^{-1/2}. \tag{D.17}$$

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.14) and (D.15), and using the a priori bound (D.4), we get that for any $i\in\mathcal{I}_1$,

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

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

$$\sum_{\nu\in\mathcal{I}_2\cup\mathcal{I}_3}\left|G^{(i)}_{\mu\nu}\right|^2 \leqslant \sum_{\mathfrak{a}\in\mathcal{I}}\left|G^{(i)}_{\mu\mathfrak{a}}\right|^2 = \left[G^{(i)}(G^{(i)})^*\right]_{\mu\mu} = \mathrm{O}(1) \tag{D.18}$$

by (D.4). Similarly, applying Lemma D.6 to Z_μ in (D.15) 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

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

and

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

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

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

Thus we obtain that $\Lambda_o\prec n^{-1/2}$, which concludes (D.17). □

Note that comibining (D.4) and (D.17), we immediately conclude (D.13) 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

$$\Xi(z) := \left\{|m_2(z) - m_{2c}(z)| + |m_3(z) - m_{3c}(z)| \leqslant (\log n)^{-1/2}\right\}. \tag{D.19}$$

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

$$|m_{2c}| \sim 1, \quad |m_{3c}| \sim 1, \quad |z+\sigma_i^2 r_1 m_{2c} + r_2 m_{3c}| \sim 1, \quad |1+\gamma_n m_c| \sim 1, \quad |1+\gamma_n m_{1c}| \sim 1, \tag{D.20}$$

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.20) into (C.18), we get

$$|\Pi_{\mathfrak{a}\mathfrak{a}}(z)| \sim 1 \text{ uniformly in } z \in \mathbf{D}, \ \mathfrak{a} \in \mathcal{I}. \tag{D.21}$$

Then we claim the following result.

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

$$\begin{aligned} \mathbf{1}(\Xi)\left|\frac{1}{m_2} + 1 - \frac{1}{n}\sum_{i=1}^p \frac{\sigma_i^2}{z+\sigma_i^2 r_1 m_2 + r_2 m_3}\right| &\prec n^{-1/2}, \\ \mathbf{1}(\Xi)\left|\frac{1}{m_3} + 1 - \frac{1}{n}\sum_{i=1}^p \frac{1}{z+\sigma_i^2 r_2 m_2 + r_2 m_3}\right| &\prec n^{-1/2}. \end{aligned} \tag{D.22}$$

Proof. By (D.8), (D.14) and (D.15), we obtain that for $i \in \mathcal{I}_1$, $\mu \in \mathcal{I}_2$ and $\nu \in \mathcal{I}_3$,

$$\frac{1}{G_{ii}} = -z - \frac{\sigma_i^2}{n}\sum_{\mu\in\mathcal{I}_2} G_{\mu\mu}^{(i)} - \frac{1}{n}\sum_{\mu\in\mathcal{I}_3} G_{\mu\mu}^{(i)} + Z_i = -z - \sigma_i^2 r_1 m_2 - r_2 m_3 + \varepsilon_i, \tag{D.23}$$

$$\frac{1}{G_{\mu\mu}} = -1 - \frac{1}{n}\sum_{i\in\mathcal{I}_1} \sigma_i^2 G_{ii}^{(\mu)} + Z_\mu = -1 - \gamma_n m_1 + \varepsilon_\mu, \tag{D.24}$$

$$\frac{1}{G_{\nu\nu}} = -1 - \frac{1}{n}\sum_{i\in\mathcal{I}_1} G_{ii}^{(\nu)} + Z_\nu = -1 - \gamma_n m + \varepsilon_\nu, \tag{D.25}$$

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.17) in the second step. Similarly, we can get that

$$|m - m^{(\mu)}| + |m_1 - m_1^{(\mu)}| + |m_2 - m_2^{(i)}| + |m_3 - m_3^{(i)}| \prec n^{-1} \tag{D.26}$$

for any $i \in \mathcal{I}_1$ and $\mu \in \mathcal{I}_2 \cup \mathcal{I}_3$. Together with (D.17), we obtain that for all i and μ,

$$|\varepsilon_i| + |\varepsilon_\mu| \prec n^{-1/2}. \tag{D.27}$$

With (D.20) 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.23), (D.27) and (D.17), we obtain that

$$\mathbf{1}(\Xi)G_{ii} = \mathbf{1}(\Xi)\left[-\frac{1}{z+\sigma_i^2 r_1 m_2 + r_2 m_3} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right]. \tag{D.28}$$

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

$$\mathbf{1}(\Xi)m = \mathbf{1}(\Xi)\left[-\frac{1}{p}\sum_{i\in\mathcal{I}_1}\frac{1}{z+\sigma_i^2 r_1 m_2 + r_2 m_3} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right], \tag{D.29}$$

$$\mathbf{1}(\Xi)m_1 = \mathbf{1}(\Xi)\left[-\frac{1}{p}\sum_{i\in\mathcal{I}_1}\frac{\sigma_i^2}{z+\sigma_i^2 r_1 m_2 + r_2 m_3} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right]. \tag{D.30}$$

As a byproduct, we obtain from the two estimates that

$$\mathbf{1}(\Xi)\left(|m-m_c|+|m_1-m_{1c}|\right)\lesssim (\log n)^{-1/2},\quad \text{with high probability.} \tag{D.31}$$

Together with (D.20), we get that

$$|1+\gamma_n m_1|\sim 1,\quad |1+\gamma_n m|\sim 1,\quad \text{with high probability on } \Xi. \tag{D.32}$$

Now using (D.24), (D.25), (D.27), (D.17) and (D.32), we can obtain that with high probability,

$$\mathbf{1}(\Xi)G_{\mu\mu}=\mathbf{1}(\Xi)\left[-\frac{1}{1+\gamma_n m_1}+\mathrm{O}_{\prec}\left(n^{-1/2}\right)\right],\quad \mu\in\mathcal{I}_2, \tag{D.33}$$

$$\mathbf{1}(\Xi)G_{\nu\nu}=\mathbf{1}(\Xi)\left[-\frac{1}{1+\gamma_n m}+\mathrm{O}_{\prec}\left(n^{-1/2}\right)\right],\quad \nu\in\mathcal{I}_3. \tag{D.34}$$

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

$$\mathbf{1}(\Xi)m_2=\mathbf{1}(\Xi)\left[-\frac{1}{1+\gamma_n m_1}+\mathrm{O}_{\prec}\left(n^{-1/2}\right)\right],\quad \mathbf{1}(\Xi)m_3=\mathbf{1}(\Xi)\left[-\frac{1}{1+\gamma_n m}+\mathrm{O}_{\prec}\left(n^{-1/2}\right)\right], \tag{D.35}$$

which further implies

$$\mathbf{1}(\Xi)\left(\frac{1}{m_2}+1+\gamma_n m_1\right)\prec n^{-1/2},\quad \mathbf{1}(\Xi)\left(\frac{1}{m_3}+1+\gamma_n m\right)\prec n^{-1/2}. \tag{D.36}$$

Finally, plugging (D.29) and (D.30) into (D.36), we conclude (D.22). □

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.22) 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

$$|m_2(0)-m_{2c}(0)|+|m_3(0)-m_{3c}(0)|\prec n^{-1/2}. \tag{D.37}$$

Once we know (D.37), 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.37), we obtain that

$$|m_2(z)-m_{2c}(z)|+|m_3(z)-m_{3c}(z)|\lesssim(\log n)^{-1}\quad \text{with high probability.} \tag{D.38}$$

and

$$\sup_{z\in\mathbf{D}}\left(|m_2(z)-m_{2c}(0)|+|m_3(z)-m_{3c}(0)|\right)\lesssim(\log n)^{-1}\quad \text{with high probability.} \tag{D.39}$$

The condition (D.38) shows that Ξ holds with high probability, and the condition (D.39) verifies the condition (C.21) of Lemma C.6. Hence applying Lemma C.6 to (D.22), we obtain that

$$|m_2(z)-m_{2c}(z)|+|m_3(z)-m_{3c}(z)|\prec n^{-1/2} \tag{D.40}$$

for all $z\in\mathbf{D}$. Plugging (D.40) into (D.23)-(D.25), we get the diagonal estimate

$$\max_{\mathfrak{a}\in\mathcal{I}}|G_{\mathfrak{a}\mathfrak{a}}(z)-\Pi_{\mathfrak{a}\mathfrak{a}}(z)|\prec n^{-1/2}. \tag{D.41}$$

Together with the off-diagonal estimate in (D.17), we conclude (D.13).

Lemma D.10. *Under the assumptions in Proposition D.1, the estimate* (D.37) *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.24), (D.25) and (D.27), we obtain that (D.35) and (D.36) 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.23) and (D.27), we obtain that (D.29) and (D.30) hold at $z = 0$ even without the indicator function $\mathbf{1}(\Xi)$. Finally, plugging (D.29) and (D.30) into (D.36), we conclude (D.22) holds at $z = 0$, that is,

$$\begin{aligned} \left|\frac{1}{m_2(0)} + 1 - \frac{1}{n}\sum_{i=1}^p \frac{\sigma_i^2}{\sigma_i^2 r_1 m_2(0) + r_2 m_3(0)}\right| &\prec n^{-1/2}, \\ \left|\frac{1}{m_3(0)} + 1 - \frac{1}{n}\sum_{i=1}^p \frac{1}{\sigma_i^2 r_2 m_2(0) + r_2 m_3(0)}\right| &\prec n^{-1/2}. \end{aligned} \tag{D.42}$$

Denoting $\omega_2 = -m_{2c}(0)$ and $\omega_3 = -m_{2c}(0)$. By (D.36), 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

$$c \leqslant \omega_2 \leqslant 1, \quad c \leqslant \omega_3 \leqslant 1, \quad \text{with high probability.} \tag{D.43}$$

Moreover, one can verify from (D.42) that (ω_2, ω_3) satisfy approximately the same equations as in (C.19):

$$r_1\omega_2 + r_2\omega_3 = 1 - \gamma_n + \mathrm{O}_\prec(n^{-1/2}), \quad f(\omega_2) = 1 + \mathrm{O}_\prec(n^{-1/2}). \tag{D.44}$$

The first equation and (D.43) 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.44) gives that $|\omega_2 - b_2| \prec n^{-1/2}$. Together with the first equation in (D.44), we get $|\omega_3 - b_3| \prec n^{-1/2}$. This concludes (D.37). □

This lemma concludes (D.37), 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.13), one can repeat the polynomialization method in [33, 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 [28]. The proof is similar to the ones in Sections 7-8 of [28], 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 $\rho^0_{\mu i}$ and $\rho^1_{\mu i}$ be the laws of $Z^0_{\mu i}$ and $Z^1_{\mu i}$, respectively. For $\theta \in [0,1]$, we define the interpolated law*

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

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

$$\prod_{i \in \mathcal{I}_1} \prod_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} \rho^\theta_{\mu i}(\mathrm{d}Z^\theta_{\mu i}). \tag{D.45}$$

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

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

We also introduce the matrices

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

We shall prove (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.45) and fundamental calculus, we get the following basic interpolation formula: for $F : \mathbb{R}^{n\times p} \to \mathbb{C}$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}F(Z^\theta) = \sum_{i \in \mathcal{I}_1} \sum_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} \left[\mathbb{E}F\left(Z^{\theta,Z^1_{\mu i}}_{(\mu i)}\right) - \mathbb{E}F\left(Z^{\theta,Z^0_{\mu i}}_{(\mu i)}\right)\right] \tag{D.46}$$

provided all the expectations exist.

We shall apply (D.46) 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

$$F_{\mathbf{u}\,\mathbf{v}}(Z,z) := \left|G_{\mathbf{u}\mathbf{v}}(Z,z) - \Pi_{\mathbf{u}\mathbf{v}}(z)\right|. \tag{D.47}$$

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.46) for any fixed $p \in 2\mathbb{N}$ and constant $\varepsilon > 0$:

$$\sum_{i \in \mathcal{I}_1} \sum_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} \left[\mathbb{E}F^p_{\mathbf{u}\,\mathbf{v}}\left(Z^{\theta,Z^1_{\mu i}}_{(\mu i)}, z\right) - \mathbb{E}F^p_{\mathbf{u}\,\mathbf{v}}\left(Z^{\theta,Z^0_{\mu i}}_{(\mu i)}, z\right)\right] = \mathrm{O}\left((n^\varepsilon q)^p + \mathbb{E}F^p_{\mathbf{u}\,\mathbf{v}}\left(Z^\theta, z\right)\right) \tag{D.48}$$

for all $\theta \in [0,1]$. If (D.48) holds, then combining (D.46) 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.48), 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

$$\sum_{i \in \mathcal{I}_1} \sum_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} \left[\mathbb{E}F^p_{\mathbf{u}\,\mathbf{v}}\left(Z^{\theta,Z^a_{\mu i}}_{(\mu i)}, z\right) - \mathbb{E}F^p_{\mathbf{v}}\left(Z^{\theta,0}_{(\mu i)}, z\right)\right] = \mathrm{O}\left((n^\varepsilon q)^p + \mathbb{E}F^p_{\mathbf{u}\,\mathbf{v}}\left(Z^\theta, z\right)\right) \tag{D.49}$$

for all $a \in \{0,1\}$ and $\theta \in [0,1]$. Underlying the proof of (D.49) is an expansion approach. We define the $\mathcal{I} \times \mathcal{I}$ matrix $\Delta^\lambda_{(\mu i)}$ as

$$\Delta^\lambda_{(\mu i)} := \lambda \begin{pmatrix} 0 & \mathbf{u}^{(\mu)}_i \mathbf{e}^\top_\mu \\ \mathbf{e}_\mu (\mathbf{u}^{(\mu)}_i)^\top & 0 \end{pmatrix}, \tag{D.50}$$

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

$$G_{(i\mu)}^{\theta,\lambda'} = G_{(\mu i)}^{\theta,\lambda} + \sum_{k=1}^{K} G_{(\mu i)}^{\theta,\lambda} \left(\Delta_{(\mu i)}^{\lambda-\lambda'} G_{(\mu i)}^{\theta,\lambda} \right)^k + G_{(\mu i)}^{\theta,\lambda'} \left(\Delta_{(\mu i)}^{\lambda-\lambda'} G_{(\mu i)}^{\theta,\lambda} \right)^{K+1}. \tag{D.51}$$

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

$$G_{(\mu i)}^{\theta,y} = \mathrm{O}(1), \quad i \in \mathcal{I}_1, \ \mu \in \mathcal{I}_2 \cup \mathcal{I}_3, \tag{D.52}$$

with high probability.

In the following proof, for simplicity of notations, we introduce $f_{(\mu i)}(\lambda) := F_{\mathbf{v}}^p(Z_{(\mu i)}^{\theta,\lambda})$. We use $f_{(\mu i)}^{(r)}$ to denote the r-th derivative of $f_{(\mu i)}$. By (D.52), it is easy to see that for any fixed $r \in \mathbb{N}$, $f_{(\mu i)}^{(r)}(y) = \mathrm{O}(1)$ with high probability for any random variable y satisfying $|y| \prec q$. Then the Taylor expansion of $f_{(\mu i)}$ gives

$$f_{(\mu i)}(y) = \sum_{r=0}^{p+4} \frac{y^r}{r!} f_{(\mu i)}^{(r)}(0) + \mathrm{O}_{\prec}\left(q^{p+4}\right), \tag{D.53}$$

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

$$\begin{aligned} &\mathbb{E} F_{\mathbf{v}}^p \left(Z_{(\mu i)}^{\theta, Z_{\mu i}^a} \right) - \mathbb{E} F_{\mathbf{v}}^p \left(Z_{(\mu i)}^{\theta,0} \right) = \mathbb{E} \left[f_{(\mu i)} \left(Z_{i\mu}^a \right) - f_{(\mu i)}(0) \right] \\ &= \mathbb{E}\, f_{(\mu i)}(0) + \frac{1}{2n} \mathbb{E}\, f_{(\mu i)}^{(2)}(0) + \sum_{r=4}^{p+4} \frac{1}{r!} \mathbb{E}\, f_{(\mu i)}^{(r)}(0)\, \mathbb{E} \left(Z_{i\mu}^a \right)^r + \mathrm{O}_{\prec}(q^{p+4}). \end{aligned} \tag{D.54}$$

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

$$\mathbb{E}(Z_{\mu i}^1)^3 = 0, \quad 1 \leqslant \mu \leqslant n, \ \ 1 \leqslant i \leqslant p. \tag{D.55}$$

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

$$\left| \mathbb{E} \left(Z_{i\mu}^a \right)^r \right| \prec n^{-2} q^{r-4}, \quad r \geqslant 4. \tag{D.56}$$

Thus to show (D.49), we only need to prove for $r = 4, 5, ..., p + 4$,

$$n^{-2} q^{r-4} \sum_{i \in \mathcal{I}_1} \sum_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} \left| \mathbb{E}\, f_{(\mu i)}^{(r)}(0) \right| = \mathrm{O}\left((n^{\varepsilon} q)^p + \mathbb{E} F_{\mathbf{u}\,\mathbf{v}}^p(Z^{\theta}, z) \right). \tag{D.57}$$

In order to get a self-consistent estimate in terms of the matrix Z^{θ} on the right-hand side of (D.57), 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*

$$n^{-2} q^{r-4} \sum_{i \in \mathcal{I}_1} \sum_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} \left| \mathbb{E}\, f_{(\mu i)}^{(r)}(Z_{i\mu}^{\theta}) \right| = \mathrm{O}\left((n^{\varepsilon} q)^p + \mathbb{E} F_{\mathbf{v}}^p(X^{\theta}, z) \right) \tag{D.58}$$

holds for $r = 4, ..., 4p + 4$. Then (D.57) holds for $r = 4, ..., 4p + 4$.

Proof. The proof is the same as the one for [28, Lemma 7.16]. □

What remains now is to prove (D.58). For simplicity of notations, we shall abbreviate $Z^{\theta} \equiv Z$. For any $k \in \mathbb{N}$, we denote

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

The derivative on the right-hand side can be calculated using the expansion (D.51). In particular, it is easy to verify that it satisfies the following bound

$$|A_{\mu i}(k)| \prec \begin{cases} (\mathcal{R}_i^{(\mu)})^2 + \mathcal{R}_\mu^2, & \text{if } k \geqslant 2 \\ \mathcal{R}_i^{(\mu)} \mathcal{R}_\mu, & \text{if } k = 1 \end{cases}, \tag{D.59}$$

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

$$\mathcal{R}_i^{(\mu)} := |G_{\mathbf{u}\,\mathbf{u}_i^{(\mu)}}| + |G_{\mathbf{v}\,\mathbf{u}_i^{(\mu)}}|, \quad \mathcal{R}_\mu := |G_{\mathbf{u}\mu}| + |G_{\mathbf{v}\mu}|. \tag{D.60}$$

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.58), it suffices to show that

$$n^{-2}q^{r-4} \sum_{i\in\mathcal{I}_1} \sum_{\mu\in\mathcal{I}_2\cup\mathcal{I}_3} \left| \mathbb{E} \prod_{t=1}^{p/2} A_{\mu i}(k_t)\overline{A_{\mu i}(k_{t+p/2})} \right| = \mathrm{O}\left((n^\varepsilon q)^p + \mathbb{E}F_{\mathbf{u}\,\mathbf{v}}^p(Z,z)\right) \tag{D.61}$$

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

$$n^{-2}q^{r-4} \sum_{i\in\mathcal{I}_1} \sum_{\mu\in\mathcal{I}_2\cup\mathcal{I}_3} \mathbb{E}\,|A_{\mu i}(0)|^{p-l} \prod_{t=1}^{l} |A_{\mu i}(k_t)| = \mathrm{O}\left((n^\varepsilon q)^p + \mathbb{E}F_{\mathbf{u}\,\mathbf{v}}^p(Z,z)\right) \tag{D.62}$$

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.59) we have

$$\prod_{t=1}^{l} |A_{\mu i}(k_t)| \prec \mathbf{1}(r \geqslant 2l-1)\left[(\mathcal{R}_i^{(\mu)})^2 + \mathcal{R}_\mu^2\right] + \mathbf{1}(r \leqslant 2l-2)(\mathcal{R}_i^{(\mu)})^2\mathcal{R}_\mu^2. \tag{D.63}$$

Using (D.4) and a similar argument as in (D.18), we get that

$$\sum_{i\in\mathcal{I}_1} (\mathcal{R}_i^{(\mu)})^2 = \mathrm{O}(1), \quad \sum_{\mu\in\mathcal{I}_2\cup\mathcal{I}_3} \mathcal{R}_\mu^2 = \mathrm{O}(1), \quad \text{with high probability.} \tag{D.64}$$

Using (D.64) 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.62) follows from Hölder's inequality. This concludes the proof of (D.58), and hence of (D.49), and hence of (C.26).

Finally, if the condition (D.55) does not hold, then there is also an $r = 3$ term in the Taylor expansion (D.54):

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

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.49) 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 [28, 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:

$$r_1 m_{2c} = -(1-\gamma_n) - r_2 m_{3c} - z\left(\frac{1}{m_{3c}} + 1\right), \quad g_z(m_{3c}(z)) = 1, \tag{D.65}$$

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.65) 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

$$g_z'(m_{3c}(0))\varepsilon(z) = -\left[g_z(m_{3c}(0)) - g_0(m_{3c}(0))\right] - \left[g_z(m_{3c}(0)+\varepsilon(z)) - g_z(m_{3c}(0)) - g_z'(m_{3c}(0))\varepsilon(z)\right]. \tag{D.66}$$

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

$$\varepsilon^{(k+1)} = -\frac{g_z(m_{3c}(0)) - g_0(m_{3c}(0))}{g_z'(m_{3c}(0))} - \frac{g_z(m_{3c}(0)+\varepsilon^{(k)}) - g_z(m_{3c}(0)) - g_z'(m_{3c}(0))\varepsilon^{(k)}}{g_z'(m_{3c}(0))}. \tag{D.67}$$

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

With direct calculation, one can get the derivative

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

Using (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

$$|[g_z'(m_{3c}(0))]^{-1}| \leqslant \widetilde{C}, \quad \left|\frac{g_z(m_{3c}(0)+\varepsilon_1) - g_z(m_{3c}(0)+\varepsilon_2) - g_z'(m_{3c}(0))(\varepsilon_1-\varepsilon_2)}{g_z'(m_{3c}(0))}\right| \leqslant \widetilde{C}|\varepsilon_1 - \varepsilon_2|^2, \tag{D.68}$$

and

$$\left|\frac{g_z(m_{3c}(0)) - g_0(m_{3c}(0))}{g_z'(m_{3c}(0))}\right| \leqslant \widetilde{C}|z|, \tag{D.69}$$

for all $|z| \leqslant \widetilde{c}$ and $|\varepsilon_1| \leqslant \widetilde{c}$, $|\varepsilon_2| \leqslant \widetilde{c}$. Then with (D.68) and (D.69), 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.65) subject to the condition $\|\varepsilon\|_\infty \leqslant r$. From the iteration relation (D.67), using (D.68) one can readily check that

$$\varepsilon^{(k+1)} - \varepsilon^{(k)} = h(\varepsilon^{(k)}) - h(\varepsilon^{(k-1)}) \leqslant \widetilde{C}|\varepsilon^{(k)} - \varepsilon^{(k-1)}|^2. \tag{D.70}$$

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.65).

Note that with (D.69) and $\varepsilon^{(0)} = \mathbf{0}$, we get from (D.67) that

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

With the contraction mapping, we have the bound

$$|\varepsilon| \leqslant \sum_{k=0}^{\infty} \|\varepsilon^{(k+1)} - \varepsilon^{(k)}\|_\infty \leqslant 2\widetilde{C}|z|. \tag{D.71}$$

This gives the bound (C.22) for $m_{3c}(z)$. Using the first equation in (D.65), 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.65) and (D.71), 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.65) approximately up to some small error

$$r_1 m_{2c} = -(1 - \gamma_n) - r_2 m_{3c} - z\left(\frac{1}{m_{3c}} + 1\right) + \mathcal{E}_2'(z), \quad g_z(m_{3c}(z)) = 1 + \mathcal{E}_3'(z), \tag{D.72}$$

with $|\mathcal{E}_2'(z)| + |\mathcal{E}_3'(z)| = \mathrm{O}(\delta(z))$. Then we subtract the equations (D.65) from (D.72), 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. □

E Supplementary Materials for the Experiments

E.1 Synthetic Settings

In Figure 1 (c), we plot the test error of the target task for $n_2 = 4p$ and n_1 ranging from p to $20p$.

E.2 Image and Text Classification Settings

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 similarity. We validate that MTL performs better when the source task is more similar to the target task. We show the result on the sentiment analysis tasks. For a target task, we manually select a similar task and a dissimilar task based on prior knowledge. Figure 2a confirms the result. Recall that Section 3.2 shows that increasing the data size of the source task does not always improve the performance of MTL for the target task. In Figure 2b, we show that for source task MR and target task SST, there is a transition from positive to negative transfer as we increase the data size of the source task. When the source task data size is particularly large compared to the target task, we show that applying the covariance alignment algorithm results in more significant gains. In Figure 2c, we observe that the benefit from aligning task covariances becomes more significant for LSTM and MLP as we increase the number of datapoints of the source task.

Algorithm 1 An incremental training schedule for efficient multi-task learning with two tasks

Input: Two tasks (X_1, Y_1) and (X_2, Y_2).
Parameter: A shared module B, output layers W_1, W_2 as in the hard parameter sharing architecture.
Require: # batches S, epochs T, task 2's validation accuracy $\hat{g}(B; W_2)$, a threshold $\tau \in (0, 1)$.
Output: The trained modules B, W_2 optimized for task 2.

Divide (X_1, Y_1) randomly into S batches: $(x^{(1)}, y^{(1)}), \ldots, (x^{(S)}, y^{(S)})$.
for $i = 1, \ldots, S$ **do**
    for $j = 1, \ldots, T$ **do**
        Update B, W_1, W_2 using the training data $\{x^{(k)}, x^{(k)}\}_{k=1}^{i}$ and (X_2, Y_2).
    end for
    Let $a_i = \hat{g}(B; W_2)$ be the validation accuracy.
    if $a_i < a_{i-1}$ or $a_i > \tau$ **then**
        break
    end if
end for

(a) Task pair TREC and CR

(b) Task pair MR and SST

Figure 3: 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.