Information Transfer in Multi-Task Learning via a Bias-Variance Decomposition in High Dimensions

August 10, 2020

Abstract

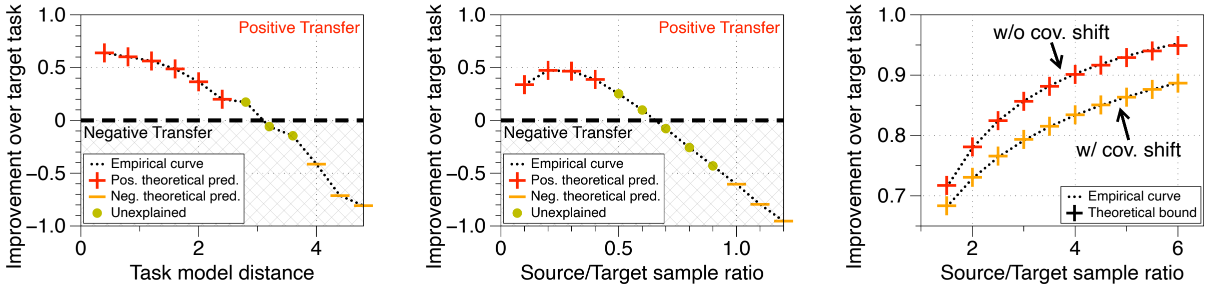
Multi-task learning is a powerful approach in many applications such as image and text classification. Yet, there is little rigorous understanding of when multi-task learning outperforms single-task learning. In this work, we provide a rigorous study to anwer the question in the high-dimensional linear regression setting. We show that a bias-variance tradeoff of multi-task learning determines the effect of information transfer and develop new concentration bounds to analyze the tradeoff. Our key observation is that three properties of task data, namely *task similarity*, *sample size*, and *covariate shift* can affect transfer in the high-dimensional linear regression setting. We relate each property to the bias and variance of multi-task learning and explain three negative effects with decreased task similarity, increased source sample size, and covariate shift under increased source sample size. We validate the three effects on text classification tasks. Inspired by our theory, we show two practical connections of interest. First, single-task results can help understand when multi-task learning gives gains. Second, incrementally adding training data can mitigate negative transfer and improve multi-task training efficiency.

1 Introduction

Multi-task learning is a powerful approach to improve performance for many tasks in computer vision (Rajpurkar et al., 2017; Zamir et al., 2018), natural language processing (Wang et al., 2018, 2019), and other areas (Zhang and Yang, 2017). In many settings, multiple source tasks are available to help predict a particular target task. The performance of multi-task learning depends on the relationship between the source and target tasks (Caruana, 1997). When the sources are relatively different from the target, multi-task learning (MTL) has often been observed to perform worse than single-task learning (STL) (Alonso and Plank, 2016; Bingel and Søgaard, 2017), which is referred to as *negative transfer* (Pan and Yang, 2009). While many empirical approaches have been proposed to mitigate negative transfer (Zhang and Yang, 2017), a precise understanding of when negative transfer occurs remains elusive in the literature (Ruder, 2017).

Understanding negative transfer requires developing generalization bounds that scale tightly with properties of each task data, such as its sample size. This presents a technical challenge in the multi-task setting because of the difference among task features, even for two tasks. For Rademacher complexity or VC-based techniques, the generalization error scales down as the sample sizes of all tasks increase, when applied to the multi-task setting (Baxter, 2000; Ando and Zhang, 2005; Maurer, 2006; Maurer et al., 2016; Wu et al., 2020). Without a tight lower bound for multi-task learning, comparing its performance to single-task learning results in vacuous bounds. From a practical standpoint, developing a better understanding of multi-task learning in terms of properties of task data can provide guidance for downstream applications (Ratner et al., 2019).

In this work, we study the bias and variance of multi-task learning in the high-dimensional linear regression setting (Hastie et al., 2019; Bartlett et al., 2020). Our key observation is that three properties of task data, including *task similarity*, *sample size*, and *covariate shift*, can affect whether multi-task learning outperforms single-task learning (which we refer to as *positive transfer*). As an example, we vary each property in Figure 1 for two linear regression tasks and measure the improvement of multi-task learning over single-task learning for a particular task. We observe that the effect of transfer can be positive or negative as we vary each property. These phenomena cannot be explained using previous techniques (Wu et al., 2020). The high-dimensional linear regression setting allows us to measure the three properties precisely. We define each property for the



(a) Task similarity

(b) Sample size

(c) Covariate shift

Figure 1: We observe a transition from positive to negative transfer as (a) *task model distance* increases and (b) source/target *sample ratio* increases. For the special case of having the same task model, we observe in (c) that as source/target *sample ratio* increases, having *covariate shift* worsens the performance of MTL. The y-axis measures the loss of STL minus MTL.

case of two tasks and our definition applies to general settings. We refer to the first task as the source task and the second as the target task.

- **Task similarity:** Assume that both tasks follow a linear model with parameters $\beta_1, \beta_2 \in \mathbb{R}^p$, respectively. We measure the distance between them by $\|\beta_1 - \beta_2\|$.
- **Sample size:** Let $n_1 = \rho_1 \cdot p, n_2 = \rho_2 \cdot p$ be the sample size of each task, where $\rho_1, \rho_2 > 1$ are both fixed values that do not grow with p. We measure the source/target sample ratio by ρ_1/ρ_2.
- **Covariate shift:** Assume that the task features are random vectors with positive semidefinite covariance matrix $\Sigma_1, \Sigma_2 \in \mathbb{R}^{p\times p}$, respectively. We measure covariate shift with matrix $\Sigma_1^{1/2}\Sigma_2^{-1/2}$.

We consider a multi-task estimator obtained using a shared linear layer for all tasks and a separate output layer for each task (Wu et al., 2020). This two-layer model is inspired by a commonly used idea of hard parameter sharing in multi-task learning (Ruder, 2017; Liu et al., 2019). We consider the bias and variance of the multi-task estimator for predicting a target task and compare its performance to single-task learning.

Main results. First, we develop tight bounds for the bias and variance of the multi-task estimator for two tasks by applying recent development in random matrix theory (Erdos and Yau, 2017; Bloemendal et al., 2014; Knowles and Yin, 2016). We observe that the variance of the multi-task estimator is *always smaller* than single-task learning, because of added source task samples. On the other hand, the bias of the multi-task estimator is *always larger* than single-task learning, because of model distances. Hence, the tradeoff between bias and variance determines whether the transfer is positive or negative. We provide a sharp analysis of the *variance* that scales with sample size and covariate shift. We extend the analysis to the bias, which *in addition* scales with task similarity. Combining both, we analyze the bias-variance tradeoff for two tasks in Theorem 3.2 and extend the analysis to many tasks with the same features in Theorem 3.3.

Second, we explain the phenomena in Figure 1 in isotropic and covariate shifted settings.

- We provide conditions to predict the effect of transfer as a parameter of model distance $\|\beta_1 - \beta_2\|$ (Section 4.1). As model distance increases, the bias becomes larger, resulting in negative transfer.
- We provide conditions to predict transfer as a parameter of sample ratio ρ_1/ρ_2 (Section 4.2). Adding source task samples helps initially by reducing variance, but hurts eventually due to bias.
- For a special case of $\beta_1 = \beta_2$, we show that MTL performs best when the singular values of $\Sigma_1^{1/2}\Sigma_2^{-1/2}$ are all equal (Section 4.3). Otherwise, the variance reduces less with covariate shift.

Along the way, we analyze the benefit of MTL for reducing labeled data to achieve comparable performance to STL, which has been empirically observed in Taskonomy by Zamir et al. (Zamir et al., 2018).

Our study also leads to several algorithmic consequences with practical interest. First, we show that single-task learning results can help predict positive or negative transfer for multi-task learning. We validate this observation on ChestX-ray14 (Rajpurkar et al., 2017) and sentiment analysis datasets (Lei et al., 2018a). Second, we propose a new multi-task training schedule by incrementally adding task data batches to the training procedure. This is inspired by our observation in Figure 1b where adding more source task data helps initially, but hurts eventually. Using our incremental training schedule, we reduce the computational cost by 65% compared to baseline multi-task training over six sentiment analysis datasets while keeping the accuracy the same. Third, we provide a fine-grained insight on a covariance alignment procedure proposed in (Wu et al., 2020). We show that the alignment procedure provides more significant improvement when the source/target sample ratio is large. Finally, we validate our three theoretical findings on sentiment analysis tasks.

Organizations. The appendix is organized as follows. In Section 2.4, we describe an extended background and a further reivew of related work. In Section 6.1, we present the analysis of the bias-variance tradeoff for the case of two tasks. In Section B, we illustrate the above analysis in simplified settings including the isotropic model and covariate shifted settings without model distance. In Section 6.2, we extend our analysis to many tasks with the same features. In Section A, we prove the bias and variance bounds used in the analysis. Finally in Section 5.5, we fill in the missing details of our experiments.

2 Problem Formulation and Related Work

We begin by defining our problem setup including the multi-task estimator we study. Then, we describe the bias-variance tradeoff of the multi-task estimator and connect the bias and variance of the estimator to *task similarity*, *sample size*, and *covariate shift*.

2.1 Problem Formulation

Suppose we have t datasets, where t is a fixed value that does not grow with the feature dimension p. In the high-dimensional linear regression setting (e.g. (Hastie et al., 2019; Bartlett et al., 2020)), the features of the k-th task, denoted by $X_k \in \mathbb{R}^{n_i \times p}$, consist of n_k feature vectors given by $x_1, x_2, \dots, x_{n_k}$. And each feature $x_i = \Sigma^{1/2} z_i$, where $z_i \in \mathbb{R}^p$ consists of i.i.d. entries with mean zero and unit variance. The sample size n_k equals $\rho_k \cdot p$ for a fixed value ρ_k. The labels $Y_k = X_k \beta_k + \varepsilon_k$, where β_k denotes the linear model parameters and ε_k denotes i.i.d. noise with mean zero and variance σ^2.

We focus on the commonly used hard parameter sharing model for multi-task learning (Ruder, 2017). When specialized to the linear regression setting, the model consists of a linear layer $B \in \mathbb{R}^{p \times r}$ that is shared by all tasks and t output layers $W_1, \dots, W_t$ that are in $\mathbb{R}^r$. The width of B, denoted by r, plays an important role in regularization. As observed in Proposition 1 of (Wu et al., 2020), if $r \geqslant t$, there is no regularization effect. Hence, we assume that $r < t$ in our study. For example, when there are only two tasks, $r = 1$ and B reduces to a vector whereas W_1, W_2 become scalars. We study the following procedure inspired by how hard parameter sharing models are trained in practice (e.g. (Liu et al., 2019)).

- Separate each dataset (X_i, Y_i) randomly into a training set (X_i^{tr}, Y_i^{tr}) and a validation set (X_i^{val}, Y_i^{val}). The size of each set is described below.
- Learn the shared layer B: minimize the training loss over B and $W_1, \dots, W_t$, leading to a closed form equation for $\hat{B}$ that depends on $W_1, \dots, W_k$.
$$f(B; W_1, \dots, W_t) = \sum_{k=1}^{t} \|X_k^{tr} B W_k - Y_k^{tr}\|^2. \tag{2.1}$$
- Tune the output layers W_i: set $B = \hat{B}$ and minimize the validation loss over $W_1, \dots, W_k$.
$$g(W_1, \dots, W_t) = \sum_{k=1}^{t} \|X_k^{val} \hat{B} W_k - Y_k^{val}\|^2. \tag{2.2}$$

We make several remarks. In general, the objective $f(\cdot)$ is non-convex in B and the W_k's. Therefore, we first minimize B in equation (2.1) and then minimize W_k given B in equation (2.2). For our purpose, a validation set of size $\rho_i \cdot p^{0.99}$ that is much larger than the number of output layer parameters $r \cdot t$ suffices. The size of the training set is then $\rho_i(p - p^{0.99})$. The advantage of tuning the output layers on the validation set is to reduce the effect of noise from $\hat{B}$.

Problem statement. We focus on predicting a particular task, say the t-th task, without loss of generality. Let $\hat{\beta}_t^{\mathrm{MTL}}$ denote the multi-task estimator obtained from the procedure above. Our goal is to compare the prediction loss of $\hat{\beta}_t^{\mathrm{MTL}}$, defined by

$$L(\hat{\beta}_t^{\mathrm{MTL}}) = \mathop{\mathbb{E}}_{\{\varepsilon_i\}_{i=1}^t} \mathop{\mathbb{E}}_{x = \Sigma_t^{1/2} z} \left[(x^\top \hat{\beta} - x^\top \beta_t)^2 \right] = \mathop{\mathbb{E}}_{\{\varepsilon_i\}_i^t} \left\| \Sigma_2^{1/2} (\hat{\beta}_t^{\mathrm{MTL}} - \beta_t) \right\|^2,$$

to the prediction loss $L(\hat{\beta}_t^{\mathrm{STL}})$ of the single-task estimator $\hat{\beta}_t^{\mathrm{STL}} = (X_t^\top X_t)^{-1} X_t^\top Y_t$. We say there is negative transfer if $L(\hat{\beta}_t^{\mathrm{MTL}}) > L(\hat{\beta}_t^{\mathrm{STL}})$ and positive transfer otherwise.

2.2 Bias and Variance

As an example, for the setting of two tasks, we can decompose $L(\hat{\beta}_t^{\mathrm{MTL}}) - L(\hat{\beta}_t^{\mathrm{STL}})$ into a bias term and a variance term as follows (derived in Appendix 2.4).

$$L(\hat{\beta}_t^{\mathrm{MTL}}) - L(\hat{\beta}_t^{\mathrm{STL}}) = \hat{v}^2 \left\| \Sigma_2^{1/2} (\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 (\beta_1 - \hat{v}\beta_2) \right\|^2 \tag{2.3}$$

$$+ \sigma^2 \left(\mathrm{Tr}\left[(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_2 \right] - \mathrm{Tr}\left[(X_2^\top X_2)^{-1} \Sigma_2 \right] \right). \tag{2.4}$$

In the above, $\hat{v} = W_1 / W_2$ where W_1, W_2 are obtained from solving equation (2.2) (recalling that W_1, W_2 are scalars for two tasks). The role of $\hat{v}$ is to scale the shared subspace B to fit each task.

Equation (2.3) corresponds to the bias of $\hat{\beta}_t^{\mathrm{MTL}}$. Hence, the bias term introduces a negative effect that depends on the *similarity* between β_1 and β_2. Equation (2.4) corresponds to the variance of $\hat{\beta}_t^{\mathrm{MTL}}$ minus the variance of $\hat{\beta}_t^{\mathrm{STL}}$, which is always negative. Intuitively, the more *samples* we have, the smaller the variance is. Meanwhile, *covariate shift* also affects how small the variance can be.

2.3 Related Work

We refer the interested readers to several excellent surveys on multi-task learning for a comprehensive survey (Pan and Yang, 2009; Ruder, 2017; Zhang and Yang, 2017; Vandenhende et al., 2020). Below, we describe several lines of work that are most related to this work.

Theoretical works. Some of the earliest works on multi-task learning are Baxter (Baxter, 2000), Ben-David and Schuller (Ben-David and Schuller, 2003). Mauer (Maurer, 2006) studies generalization bounds for linear separation settings of MTL. Ben-David et al. (Ben-David et al., 2010) provides uniform convergence bounds that combines source and target errors optimally. The benefit of learning multi-task representations has been studied for learning certain half-spaces (Maurer et al., 2016) and sparse regression (Lounici et al., 2009, 2011). Our work is closely related to Wu et al. (Wu et al., 2020). While Wu et al. provide generalization bounds to show that adding more labeled helps learn the target task more accurately, their techniques cannot be used to explain when MTL outperforms STL.

Methodological works. Ando and Zhang (Ando and Zhang, 2005) introduces an alternating minimization framework for learning multiple tasks. Argyriou et al. (Argyriou et al., 2008) present a convex algorithm which learns common sparse representations across a pool of related tasks. Evgeniou et al. (Evgeniou et al., 2005) develop a framework for multi-task learning in the context of kernel methods. The multi-task learning model that we have focused on uses the idea of hard parameter sharing (Caruana; Kumar and Daumé III, 2012; Ruder, 2017). We believe that our theoretical framework can apply to other approaches to multi-task learning.

Random matrix theory. The random matrix theory tool and related proof of our work fall into a paradigm of the so-called local law of random matrices (Erdos and Yau, 2017). For a sample covariance matrix $X^\top X$ with $\Sigma = \mathrm{Id}$, such a local law was proved in (Bloemendal et al., 2014). It was later extended to sample covariance matrices with non-identity Σ (Knowles and Yin, 2016), and separable covariance matrices (Yang, 2019). On the other hand, one may derive the asymptotic result in Theorem 3.1 with error $\mathrm{o}(1)$ using the free addition of two independent random matrices in free probability theory (Nica and Speicher, 2006). To the best of my knowledge, we do not find an *explicit result* for the sum of two sample covariance matrices with general covariates in the literature.

2.4 Extended Background and Related Work

We restate our linear regression setting, in particular the moment assumption required by our random matrix model. We then derive a closed-form solution of the multi-task estimator that is defined in Section 2. Finally, we describe further related work that is covered in the main text.

First, we give the basic assumption for our main objects—the random matrices X_i, $i = 1, 2$.

Assumption 2.1 (Moment assumptions)**.** *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 matrix, 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*

$$\mathbb{E}|z_{ij}|^k \leqslant C_k, \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant p. \tag{2.5}$$

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 (2.5) 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 A.8 in Section A.1.

Then we make the following assumptions on the data models.

Assumption 2.2 (Linear regression model)**.** *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 2.1 with $\rho_i := n_i/p > 1$ being fixed constants. $\varepsilon_i \in \mathbb{R}^{n_i}$ are random vectors with i.i.d. entries with mean zero, variance σ_i^2 and all moments as in (2.5).*

Throughout the appendix, we shall say an event Ξ holds with high probability (w.h.p.) 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 quantity that converges to 0 as $p \to \infty$.

Next, we derive a closed-form solution of the multi-task learning estimator for the case of two tasks. From (Wu et al., 2020), we know that we need to explicitly restrict the output dimension r of B so that there is transfer between the two tasks. Hence for the case of two tasks, we consider the setting where $r = 1$. For simplicity of notations, we shall denote (X_i^{tr}, Y_i^{tr}) and (X_i^{val}, Y_i^{val}) as (X_i, Y_i) and $(\widetilde{X}_i, \widetilde{Y}_i)$, respectively. Then equation (2.1) simplifies to

$$f(B; w_1, w_2) = \|X_1 B w_1 - Y_1\|^2 + \|X_2 B w_2 - Y_2\|^2, \tag{2.6}$$

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

$$\begin{aligned} \hat{B}(w_1, w_2) &= (w_1^2 X_1^\top X_1 + w_2^2 X_2^\top X_2)^{-1}(w_1 X_1^\top Y_1 + w_2 X_2^\top Y_2) \qquad (2.7) \\ &= \frac{1}{w_2}\left(\frac{w_1^2}{w_2^2} X_1^\top X_1 + X_2^\top X_2\right)^{-1}\left(\frac{w_1}{w_2} X_1^\top Y_1 + X_2^\top Y_2\right) \\ &= \frac{1}{w_2}\left[\beta_2 + \left(\frac{w_1^2}{w_2^2} X_1^\top X_1 + X_2^\top X_2\right)^{-1}\left(X_1^\top X_1\left(\frac{w_1}{w_2}\beta_1 - \frac{w_1^2}{w_2^2}\beta_2\right) + \left(\frac{w_1}{w_2} X_1^\top \varepsilon_1 + X_2^\top \varepsilon_2\right)\right)\right]. \end{aligned}$$

As a remark, when $w_1 = w_2 = 1$, we obtain linear regression. If β_1 is a scaling of β_2, then w_1, w_2 can be scaled accordingly to fix both tasks more accurately than linear regression.

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 $\widetilde{Y}_i \in \mathbb{R}^{N_i}$, $i = 1, 2$, whose row vectors are given by $\widetilde{x}_k^{(i)}$ and $\widetilde{y}_k^{(i)}$. We assume that N_1 and N_2 satisfy $N_1/N_2 = n_1/n_2$ and $N_i \geqslant n_i^{1-\varepsilon_0}$ for some constant $\varepsilon_0 > 0$. Then we write the validation loss in (2.2) as

$$g(w_1, w_2) = \left\|\widetilde{X}_1 \hat{B} w_1 - \widetilde{Y}_1\right\|^2 + \left\|\widetilde{X}_2 \hat{B} w_2 - \widetilde{Y}_2\right\|^2. \tag{2.8}$$

Inserting (2.7) into (2.8), one can see that the optimal solution of g only depends on the ratio $v := w_1/w_2$. Hence we overload the notation by writing $g(v)$ in the following discussion. The expectation of $g(v)$ can be written as follows.

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

Claim 2.3. *In the setting described above, we have that*

$$g(v)=\left[val(v)+(N_1\sigma_1^2+N_2\sigma_2^2)\right]\cdot\left(1+\mathrm{O}(p^{-(1-\varepsilon_0)/2+\varepsilon})\right). \tag{2.9}$$

Proof. We use the fact that our random vectors have i.i.d. entries. Recall that $Y_i = X_i\beta_i+\varepsilon_i$ and $\widetilde{Y}_i=\widetilde{X}_i\beta_i+\widetilde{\varepsilon}_i$, $i=1,2$, all satisfy Assumption 2.2. Then we rewrite (2.8) as

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

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 result, Lemma A.14, 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

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

where we also used $N_i\geqslant p^{-1+\varepsilon_0}$. Inserting (2.7) into the above expression and using again the concentration result, Lemma A.14, we get that

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

with high probability. Thus we conclude the proof. □

Hence to minimize $g(v)$, it suffices to minimize $val(v)$ over v. Let $\hat{v}=\hat{w}_1/\hat{w}_2$ be the global minimizer of $g(v)$. Now we can define the multi-task learning estimator for the target task as

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

The prediction loss of using $\hat{\beta}_2^{\mathrm{MTL}}$ for the target task is

$$\begin{aligned}
L(\hat{\beta}_2^{\mathrm{MTL}})=&\ \hat{v}^2\left\|\Sigma_2^{1/2}(\hat{v}^2X_1^\top X_1+X_2^\top X_2)^{-1}X_1^\top X_1(\beta_1-\hat{v}\beta_2)\right\|^2\\
&+\ \operatorname{Tr}\left[\Sigma_2(\hat{v}^2X_1^\top X_1+X_2^\top X_2)^{-2}\left(\sigma_1^2\cdot\hat{v}^2X_1^\top X_1+\sigma_2^2\cdot X_2^\top X_2\right)\right],
\end{aligned}\tag{2.10}$$

which only depends on $\hat{v}$, the sample covariance matrices, and β_1,β_2.

Extended related work. Our setting is closely related to domain adaptation (Daume III and Marcu, 2006; Ben-David et al., 2007; Blitzer et al., 2008; Daumé III, 2009; Mansour et al., 2009; Chen et al., 2011; Zhang et al., 2013; McNamara and Balcan, 2017; Zhao et al., 2019). The important distinction is that we focus on predicting the target task using a hard parameter sharing model. For such models, their output dimension plays an important role of regularization (Kumar and Daumé III, 2012). Linear models in multi-task learning have been studied in various settings, including representation learning (Bullins et al., 2019), online learning (Cavallanti et al., 2010; Denevi et al., 2018), and sparse regression (Lounici et al., 2011).

3 Main Results

3.1 Two Tasks

Lemma 3.1 (Variance bound)**.** *In the setting of two tasks, let $n_1 = \rho_1 \cdot p$ and $n_2 = \rho_2 \cdot$ be the sample size of the two tasks. Let $\lambda_1, \dots, \lambda_p$ be the singular values of the covariate shift matrix $\Sigma_1^{1/2}\Sigma_2^{-1/2}$ in decreasing order. With high probability, the variance of the multi-task estimator $\hat{\beta}_t^{MTL}$ equals*

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

where a_1, a_2 are solutions of the following equations:

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

Lemma 3.1 allows us to get a tight bound on equation (2.4), that only depends on *sample size*, *covariate shift* and the scalar $\hat{v}$. As a remark, the concentration error $\mathrm{O}(p^{-1/2+o(1)})$ of our result is nearly optimal. For the bias term of equation (2.3), a similar result that scales with task model distance in addition to sample size and covariate shift holds (cf. Lemma 6.3 in Appendix 6.1). Combining the two lemmas, we provide a sharp analysis of the bias-variance tradeoff of the multi-task estimator. For a matrix X, let $\lambda_{\min}(X)$ denote its smallest singular value and $\|X\|$ denote its spectral norm.

3.2 Multiple Tasks

A well-known result in the high-dimensional linear regression setting states that $\mathrm{Tr}[(X_2^\top X_2)^{-1}\Sigma_2]$ is concentrated around $1/(\rho_2 - 1)$ (e.g. Chapter 6 of (Serdobolskii, 2007)), which scales with the sample size of the target task. Our main technical contribution is to extend this result to two tasks. We show how the variance of the multi-task estimator scales with sample size and covariate shift in the following result.

Theorem 3.2 (Two tasks)**.** *For the setting of two tasks, let $\delta > 0$ be a fixed error margin, $\rho_2 > 1$ and $\rho_1 \gtrsim \delta^{-2} \cdot \lambda_{\min}(\Sigma_1^{1/2}\Sigma_2^{-1/2})^{-4}\|\Sigma_1\|\max(\|\beta_1\|^2, \|\beta_2\|^2)$. There exist 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 $L(\hat{\beta}_t^{MTL}) > L(\hat{\beta}_t^{STL})$.*

Theorem 3.2 applies to settings where large amounts of source task data are 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 these functions can be complex (as are previous generalization bounds for MTL), they admit interpretable forms for simplified settings.

The proof of Theorem 3.2 is presented in Appendix 6.1 and the proof of Lemma 3.1 is in Appendix A.

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 (Rajpurkar et al., 2017; Eyuboglu et al., 2020).

Theorem 3.3 (Many tasks)**.** *For the setting of t tasks where $X_i = X$, for all $1 \leqslant i \leqslant t$, let $B^\star := [\beta_1, \beta_2, \dots, \beta_t]$ and $U_r \in \mathbb{R}^{t\times r}$ denote the linear model parameters. Let $U_r U_r^\top$ denote the best rank-r subspace approximation of $(B^\star)^\top \Sigma B^\star$. Assume that $\lambda_{\min}({B^\star}^\top \Sigma B^\star) \gtrsim \sigma^2$. Let v_i denote the i-th row vector of U_r. There exists a value $\delta = \mathrm{o}\left(\|B^\star\|^2 + \sigma^2\right)$ such that*

- *If* $\left(1 - \|v_t\|^2\right) \frac{\sigma^2}{\rho-1} - \|\Sigma(B^\star U_r v_t - \beta_t)\|^2 > \delta$*, then w.h.p* $L(\hat{\beta}_t^{MTL}) < L(\hat{\beta}_t^{STL})$*.*
- *If* $\left(1 - \|v_t\|^2\right) \frac{\sigma^2}{\rho-1} - \|\Sigma(B^\star U_r v_t - \beta_t)\|^2 < -\delta$*, then w.h.p.* $L(\hat{\beta}_t^{MTL}) > L(\hat{\beta}_t^{STL})$*.*

Theorem 3.3 provides a sharp analysis of the bias-variance tradeoff beyond two tasks. Specifically, $(1 - \|v_t\|^2)\sigma^2/(\rho - 1)$ shows the amount of reduced variance and $\|\Sigma(B^\star U_r v_t - \beta_t)\|$ shows the bias of the multi-task estimator. The proof of 3.3 can be found in Appendix 6.2.

4 Theoretical Implication

We provide tight bounds on the bias and variance of the multi-task estimator for two tasks. We show theoretical implications for understanding the performance of multi-task learning. (a) *Task similarity*: we explain the phenomenon of negative transfer precisely as tasks become more different. (b) *Sample size*: we further explain a curious phenomenon where increasing the source sample size helps initially, but hurts eventually. (c) *Covariate shift*: as the source sample size increases, we show that the covariate shift worsens the performance of the multi-task estimator. Finally, we extend our results from two tasks to many tasks with the same features.

4.1 Task Similarity

It is well-known since the seminal work of Caruana (Caruana, 1997) that how well multi-task learning performs depends on task relatedness. We formalize this connection in the following simplified setting, where we can perform explicit calculations. 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.

The isotropic model. Consider two tasks with isotropic covariances $\Sigma_1 = \Sigma_2 = \mathrm{Id}$. Each task has sample size $n_1 = \rho_1 \cdot p$ and $n_2 = \rho_2 \cdot p$. Assume that for task two, 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 $Y_i = X_i\beta_i + \varepsilon_i$, where ε_i consists of i.i.d. entries with mean zero and variance σ^2. For our purpose, it is enough to think of the order of d being $1/\sqrt{p}$ and pd^2/σ^2 being constant.

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 4.1 (Task model distance)**.** *In the isotropic model, suppose that* ρ_1 *and* $\rho_2 > 1$*. Then*

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

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

Proposition 4.1 simplifies Theorem 3.2 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. Roughly speaking, the transition threshold scales as $\frac{pd^2}{\sigma^2} - \frac{1}{\rho_1} - \frac{1}{\rho_2}$. We apply Proposition 4.1 to the parameter setting of Figure 1a (the details are left to Appendix 5.6). We can see that our result is able to predict positive or negative transfer accurately and matches the empirical curve. There are several unexplained observations near the transition threshold 0, which are caused by the concentration error ν. The proof of Proposition 4.1 can be found in Appendix B.1. A key part of the analysis shows that $\hat{v} \approx 1$ in the isotropic model, thus simplifying the result of Theorem 3.2.

Algorithmic consequence. We can in fact extend the result to the cases where the noise variances are 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 Proposition B.3 in Appendix B.1. 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.

4.2 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 (Baxter, 2000; Maurer, 2006; Maurer et al., 2016). On the other hand, uneven sample sizes between different tasks (or even dominating tasks) have been empirically observed as a cause of negative transfer (Yu et al., 2020). For such settings, we have also observed that adding more labeled data from one task does not always help. In the isotropic model, we consider what happens if we vary the source task sample size. Our theory accurately predicts a curious phenomenon, where increasing the sample size of the source task results in negative transfer!

Proposition 4.2 (Source/target sample ratio)**.** *In the isotropic model, suppose that $\rho_1 > 40$ and $\rho_2 > 110$ are fixed constants, and $\Psi(\beta_1, \beta_2) > 2/(\rho_2 - 1)$. Then we have that*

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

Proposition 4.2 describes the bias-variance tradeoff in terms of the sample ratio ρ_1/ρ_2. We apply the result to the setting of Figure 1b (described in Appendix 5.6). There are several unexplained observations near $y = 0$ caused by ν. The proof of Proposition 4.2 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. (Zamir et al., 2018), 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 $x \cdot n_i$ datapoints from every task. The data efficiency ratio is defined 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 roughly

$$\frac{1}{\rho_1 + \rho_2} + \frac{2}{(\rho_1 + \rho_2)(\rho_1^{-1} + \rho_2^{-1} - \Theta(\Psi(\beta_1, \beta_2)))}.$$

Compared with Proposition 4.1, we see that when $\Psi(\beta_1, \beta_2)$ is smaller than $\rho_1^{-1} + \rho_2^{-1}$ (up to a constant multiple), the transfer is positive. Moreover, the data efficiency ratio quantifies how effective the positive transfer is using MTL. The result can be found in Proposition B.4 in Appendix B.2.

Algorithmic consequence. An interesting consequence of Proposition 4.2 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 5.5 describes the procedure in detail.

4.3 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. We measure covariate shift by $M = \Sigma_1^{1/2}\Sigma_2^{-1/2}$. Assume that $\Psi(\beta_1, \beta_2) = 0$ for simplicity. We compare two cases: (i) when $M = \mathrm{Id}$; (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 the covariate shift. Figure 1c shows a simulation of this setting by varying λ. We observe that as source/target sample ratio increases, the performance gap between the two cases increases.

We compare different choices of M that belong to the following bounded set. Let λ_i be the i-th singular value of M. Let $\mu_{\min} < \mu < \mu_{\max}$ be fixed values that do not grow with p.

$$\mathcal{S}_\mu := \left\{ M \,\middle|\, \prod_{i=1}^{p} \lambda_i \leqslant \mu^p, \mu_{\min} \leqslant \lambda_i \leqslant \mu_{\max}, \text{ for all } 1 \leqslant i \leqslant p \right\},$$

Proposition 4.3 (Covariate shift). *Assume that $\Psi(\beta_1, \beta_2) = 0$ and $\rho_1, \rho_2 > 1$. Let $g(M)$ denote the prediction loss of $\hat{\beta}_t^{MTL}$ when $M = \Sigma_1^{1/2}\Sigma_2^{-1/2} \in \mathcal{S}_\mu$. We have that*

$$g(\mu\,\mathrm{Id}) \leqslant (1 + \mathrm{O}(\rho_2/\rho_1)) \min_{M \in \mathcal{S}_\mu} g(M).$$

This proposition shows that when source/target sample ratio is large, then having no covariate shift is optimal. The proof of Proposition 4.3 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., 2020), 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.

5 Experiments

We validate our algorithmic insights and then our theory. In Section 5.4, we first show that single-task learning results can help predict positive or negative transfer. Second, our proposed incremental training schedule improves the training efficiency of standard multi-task training on sentiment analysis tasks. In Section 5.3, we validate our theoretical results. We further show that when the sample ratio is large, performing the alignment procedure of (Wu et al., 2020) provides more improvement for MTL.

5.1 Experimental Setup

We consider a text classification task and an image classification task as follows.

Sentiment Analysis. We consider six tasks: movie review sentiment (MR) (Pang and Lee, 2005), sentence subjectivity (SUBJ) (Pang and Lee, 2004), customer reviews (CR) (Hu and Liu, 2004), question type (TREC) (Li and Roth, 2002), opinion polarity (MPQA) (Wiebe et al., 2005), and the Stanford sentiment treebank (SST) tasks (Socher et al., 2013). We use an embedding layer with GloVe embeddings (Pennington et al., 2014) followed by an LSTM, MLP or CNN layer proposed by (Lei et al., 2018b).

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

Models	**Sentiment analysis**	
	all tasks	w/o TREC
MLP	31%	29%
LSTM	35%	34%
CNN	30%	28%

Table 2: Efficiency of incremental training compared to baseline MTL.

ChestX-ray14. This dataset contains 112,120 frontal-view X-ray images (Rajpurkar et al., 2017). There are 14 diseases (tasks) for every image that we would like to predict. We use densenet121 as the shared module (Huang et al., 2017).

For all models, we use a shared module for all tasks and assign a separate output layer on top of the shared module for each task. The baseline training schedule for MTL is the round-robin training schedule. We measure the test accuracy of predicting a target task. We measure computational cost by summing over all epochs the number of samples used in every epoch.

5.2 Simulation Studies

5.3 Validating the Theoretical Implication

We first validate our theoretical results in Section 4.1 and 4.2. In Figure 2a, we compare the performance training with a semantically similar task versus a dissimilar task with a target task. We select each task pair based on our domain knowledge. We observe that adding a similar task helps the target task whereas adding a dissimilar task hurts. In Figure 2b, we validate that adding more source samples does not always improve performance on the target task. Finally, we validate the algorithmic consequence of Section 4.3. In Figure 2c, we measure the performance gains from performing the alignment procedure proposed in (Wu et al., 2020) minus baseline MTL. We average the results over all 15 task pairs. The result shows that as the source samples increases, the alignment procedure shows a bigger improvement over MTL. The rest of experimental procedures are left to Appendix 5.7.

5.4 Validating the Algorithmic Consequences

Predicting transfer effect via STL results. We show that the single-task based metric proposed in Section 4.1 can predict positive or negative transfer in MTL. A common challenge in the study of MTL is that the results can be hard to understand. It is difficult to predict when MTL performs well without running extensive trials. Our insight is that we can use STL results to help understand MTL results. Table 1 shows the result on both the sentiment analysis and the ChestX-ray14 tasks. We find that using a threshold of $\tau = 0.1$, the STL results correctly predict positive or negative transfer with 75.6% precision and 38.8% recall among 30 times 5 (random seeds) task pairs! We observe similar results for 91 task pairs from the ChestX-ray14 dataset.

Mitigating negative transfer via incremental training. First, we show that our proposed incremental training schedule (Algorithm 1) can help mitigate negative transfer for predicting a particular target task. Over all 15 pairs from the sentiment analysis tasks, we find that Algorithm 1 requires only 45% of the computational cost to achieve similar performance on the target task, compared to the MTL baseline. Our insight is that since adding more samples from the source task does not always help, we can improve efficiency by adding source samples *incrementally* during training.

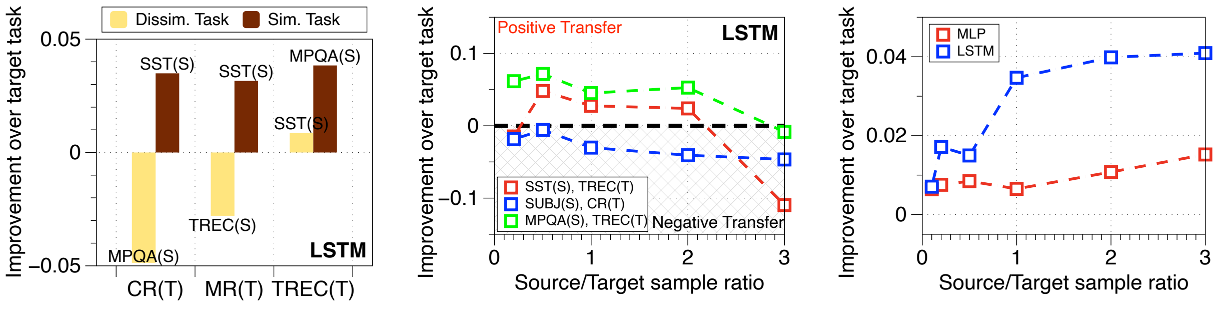
Our next result shows the incremental training schedule applies to multiple tasks as well. In Table 2, we find that over all six sentiment analysis tasks, incremental training requires less than 35% of the computational cost compared to baseline MTL training, while achieving the same accuracy averaged over all six tasks. As a further validation, excluding TREC, we observe similar comparative results.

5.5 Missing Details from the Experiments

5.6 Synthetic Settings

We describe the parameter settings used in the synthetic experiments (Figure 1). For all three synthetic experiments, we set the input feature dimension as $p = 200$.

- Task similarity: We use $\rho_1 = 90, \rho_2 = 30$ for sample sizes. The noise level is $\sigma = 5$. We set $\kappa = 1$ and vary model distance d from 0.01 to 0.2. The x-axis measures the function $\Phi(\beta_1, \beta_2)$ described in Section 4.1. The y-axis measures the improvement of prediction loss using MTL compared to STL.
- Sample size: We set $\kappa = 1$ and $d = 0.02$ for model distance. We use $\rho_2 = 500$ and vary ρ_1 from 50 to 800 for sample sizes.



(a) Task similarity

(b) Sample size

(c) Covariate shift

Figure 2: Validating the three results of Section 4 on sentiment analysis tasks. (a) Adding a semantically similar source task in MTL performs better than adding a dissimilar task. (b) As source/target sample ratio increases, we observe a transition from positive to negative transfer. (c) As source/target sample ratio increases, aligning task covariances (Wu et al., 2020) improves more over the baseline. Note: (S) denotes the source task and (T) denotes the target task.

- Covariate shift: We set $\kappa = 1$ and $d = 0$. We set $\rho_2 = 4$ and vary ρ_1 from 5 to 25 for sample sizes. We use the scale parameter $\lambda = 1$ for the curve without covariate shift and $\lambda = 2$ for the curve with covariate shift (cf. Section 4.3). Without loss of generality, we have rescaled the x and y axis for the ease of presentation. This can be achieved similarly by rescaling the task data.

5.7 Image and Text Classification Settings

For the text classification experiment, we encode each word using the GLoVe word embeddings.[1] We evaluate three model choices. For multi-layer perception (MLP), we apply an average pooling layer over the word embeddings. For LSTM and CNN, we add a shared feature representation layer on top of the word embeddings (Lei et al., 2018b).

Predicting transfer effect via STL results. We simplify the setup of this experiment by fixing the sample sizes of every task pair. For sentiment analysis tasks, the sample size of the source task ranges in $500, 1000, 1500$. We randomly sample these many data points from the task. The sample size of the target task is 1000. For image classification tasks, the training sample size is $10,000$ for every task.

Mitigating negative transfer via incremental training. For the case of two tasks, we compare the incremental training scheduled described in Algorithm 1 to a round-robin training schedule baseline. We add 20% of source task samples for every $T = 2$ epochs. The total number of epochs is 20. We set the threshold τ as the validation accuracy of the target task using the baseline training schedule.

For the case of six tasks, we extend Algorithm 1 accordingly. Initially, we use 5% of samples from SST and 50% of samples from the other tasks. We add 19% of samples from SST and 5% of samples from the other tasks for every 2 epochs. We run 20 epochs in total.

Validating the Theoretical Results. We fill in the details of the experimental procedure used for the results in Figure 2.

- Task similarity: We select a similar and a dissimilar source task compared to the target task using domain knowledge. First pair: the customer review dataset (CR) , which predicts whether a review is positive or negative, is more similar to SST (sentiment treebank) than MPQA (question type). Second pair: SST is more similar to MR since they both concern about positive or negative opinions expressed the text. TREC is less similar to MR because the task is about question types. Third pair: MPQA (opinion polarity) is more similar to TREC (question type)
- Sample size: We vary the sample size of the source task from 100 to 3000.
- Covariate shift: We implement the covariance alignment procedure following (Wu et al., 2020). We fix the sample size of the target task as 1000.

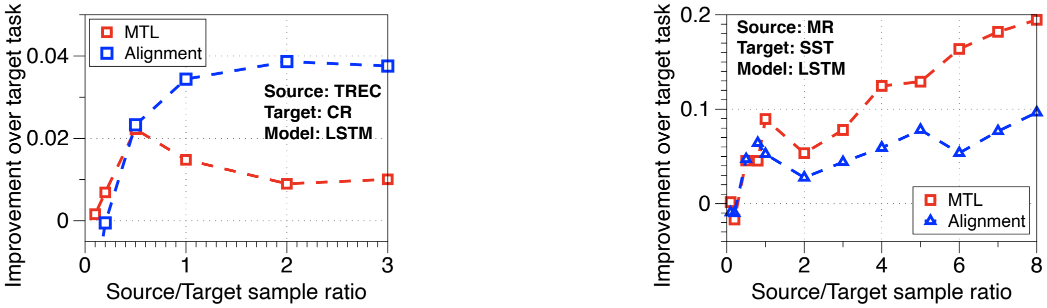
Further results of the covariance alignment procedure. Our results in Figure 2c are averaged over all the task pairs. In Figure 3, we show two task pairs as examples. In Figure 3a, we observe that for the

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

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: $(x^{(1)}, y^{(1)}), \dots, (x^{(S)}, y^{(S)})$.
for $i = 1, \dots, S$ do
    for $j = 1, \dots, T$ 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
```

particular task pair, covariance alignment provides more significant gains when the sample ratio is large. In Figure 3b, we observe that covariance alignment does not always improve over the baseline multi-task learning model. One explanation is that MR and SST are similar tasks, hence adding the alignment module is unnecessary. An interesting question is to understand when adding the alignment module benefits the multi-task learning model. We leave this question for future work.



(a) Task pair TREC and CR

(b) Task pair MR and SST

Figure 3: (a) For the task pair TREC and CR, adding the covariance alignment procedure provides more improvement when the source/target sample ratio is large. (b) For the task pair MR and SST, adding the covariance alignment procedure hurts performance. One explanation is that MR and SST are similar tasks, hence adding the alignment module is unnecessary.

6 Proof of Main Results

6.1 Two Tasks

We now state several helper lemmas to get estimates on $L(\hat{\beta}_t^{\mathrm{STL}})$ and $L(\hat{\beta}_t^{\mathrm{MTL}})$ for $t = 2$. 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 (Anderson, 2003). For general non-Gaussian random matrices, it can be obtained using Stieltjes transform method; see e.g., Lemma 3.11 of (Bai and Silverstein, 2010). Here we shall state a result obtained from Theorem 2.4 in (Bloemendal et al., 2014), which gives an almost sharp error bound.

Lemma 6.1. *Suppose X satisfies assumption 2.1. Let A be any $p \times p$ matrix that is independent of X. We have that for any constant $\varepsilon > 0$,*

$$\operatorname{Tr}\left[(X^\top X)^{-1} A\right] = \frac{1}{\rho - 1}\frac{1}{p}\operatorname{Tr}(\Sigma^{-1}A) + \mathrm{O}\left(\|A\| p^{-1/2+\varepsilon}\right) \tag{6.1}$$

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 6.1 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 6.2 (Variance bound: Lemma 3.1 restated)**.** *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 2.1 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 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 $\varepsilon > 0$,*

$$\operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1} A\right] = \frac{1}{\rho_1 + \rho_2}\frac{1}{p}\operatorname{Tr}\left[(a_1\Sigma_1 + a_2\Sigma_2)^{-1}A\right] + \mathrm{O}\left(\|A\| p^{-1/2+\varepsilon}\right) \tag{6.2}$$

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

$$a_1 + a_2 = 1 - \frac{1}{\rho_1+\rho_2}, \quad a_1 + \frac{1}{\rho_1+\rho_2}\cdot\frac{1}{p}\sum_{i=1}^p \frac{\lambda_i^2 a_1}{\lambda_i^2 a_1 + a_2} = \frac{\rho_1}{\rho_1+\rho_2}. \tag{6.3}$$

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 (2.10).

Lemma 6.3 (Bias bound)**.** *In the setting of Lemma 6.2, let $\beta \in \mathbb{R}^p$ be any vector that is independent of X_1 and X_2. We have that for any constant $\varepsilon > 0$,*

$$\begin{aligned}&(n_1+n_2)^2\left\|\Sigma_2^{1/2}(X_1^\top X_1 + X_2^\top X_2)^{-1}\beta\right\|^2 \\ &= \beta^\top\Sigma_2^{-1/2}\frac{(1+a_3)\operatorname{Id} + a_4 M^\top M}{(a_1 M^\top M + a_2)^2}\Sigma_2^{-1/2}\beta + \mathrm{O}(p^{-1/2+\varepsilon}\|\beta\|^2),\end{aligned} \tag{6.4}$$

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

$$\left(\rho_2 a_2^{-2} - b_0\right)\cdot a_3 - b_1\cdot a_4 = b_0, \quad \left(\rho_1 a_1^{-2} - b_2\right)\cdot a_4 - b_1 \cdot a_3 = b_1. \tag{6.5}$$

Here b_0, b_1 and b_2 are defined as

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

Proof overview. The proofs of Lemma 6.2 and Lemma 6.3 are based on the Stieltjes transform method (or the resolvent method) in random matrix theory (Bai and Silverstein, 2010; Tao, 2012; Erdos and Yau, 2017). 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 (Knowles and Yin, 2016; Yang, 2019), 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 $z = 0$, $\operatorname{Tr}[\Sigma_2^{-1/2}A\Sigma_2^{-1/2}R_\infty(0)]$ gives the limit in (6.2). On the other hand, we can write

$$\left\|\Sigma_2^{1/2}(X_1^\top X_1 + X_2^\top X_2)^{-1}\beta\right\|^2 = \beta^\top\Sigma_2^{-1/2}R'(0)\Sigma_2^{-1/2}\beta.$$

Hence its limit can be calculated through $R'_\infty(z)$, which gives the expression in (6.4). The details can be found in Appendix A.

We remark that one can probably derive the same asymptotic result using free probability theory (see e.g. (Nica and Speicher, 2006)), but our results (6.2) and (6.4) also give an almost sharp error bound $\mathrm{O}\left(p^{-1/2+\varepsilon}\right)$.

For the rest of this section, we shall state and prove the formal version of Theorem 3.2. First, we introduce several quantities that will be used in our statement, and they are also related to the quantities in Lemma 6.2 and Lemma 6.3. Given the optimal ratio $\hat{v}$, let $(\hat{a}_1, \hat{a}_2)$ be the solution to the following system of deterministic equations,

$$\hat{a}_1 + \hat{a}_2 = 1 - \frac{1}{\rho_1 + \rho_2}, \quad \hat{a}_1 + \frac{1}{\rho_1 + \rho_2} \cdot \frac{1}{p} \sum_{i=1}^{p} \frac{\hat{v}^2 \lambda_i^2 \hat{a}_1}{\hat{v}^2 \lambda_i^2 \hat{a}_1 + \hat{a}_2} = \frac{\rho_1}{\rho_1 + \rho_2}. \tag{6.6}$$

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

$$\left(\rho_2 \hat{a}_2^{-2} - \hat{b}_0\right) \cdot \hat{a}_3 - \hat{b}_1 \cdot \hat{a}_4 = \hat{b}_0, \quad \left(\rho_1 \hat{a}_1^{-2} - \hat{b}_2\right) \cdot \hat{a}_4 - \hat{b}_1 \cdot \hat{a}_3 = \hat{b}_1. \tag{6.7}$$

where we denoted

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

Then we introduce the following matrix

$$\Pi \equiv \Pi(\hat{v}) = \frac{\rho_1^2}{(\rho_1 + \rho_2)^2} \cdot \hat{v}^2 M \frac{(1 + \hat{a}_3) \operatorname{Id} + \hat{a}_4 \hat{v}^2 M^\top M}{(\hat{a}_1 \hat{v}^2 M^\top M + \hat{a}_2)^2} M^\top. \tag{6.8}$$

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 (Marčenko and Pastur, 1967). 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

$$\delta \equiv \delta(\hat{v}) := \frac{\alpha_+^2(\rho_1) - 1}{\alpha_-^2(\rho_1) \hat{v}^2 \lambda_{\min}^2(M)} \cdot \|\Sigma_1^{1/2}(\beta_1 - \hat{v}\beta_2)\|^2. \tag{6.9}$$

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 6.4 (Theorem 3.2 restated)**.** *Consider two data models $Y_i = X_i \beta_i + \varepsilon_i$, $i = 1, 2$, that satisfy Assumption 2.2 with $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Then with high probability, we have*

$$L(\hat{\beta}_t^{MTL}) \leqslant L(\hat{\beta}_t^{STL}) \quad \textit{when:} \quad \Delta_{var} - \Delta_{bias} \geqslant \delta(\hat{v}) \tag{6.10}$$

$$L(\hat{\beta}_t^{MTL}) \geqslant L(\hat{\beta}_t^{STL}) \quad \textit{when:} \quad \Delta_{var} - \Delta_{bias} \leqslant -\delta(\hat{v}), \tag{6.11}$$

where

$$\Delta_{var} := \sigma^2 \left(\frac{1}{\rho_2 - 1} - \frac{1}{\rho_1 + \rho_2} \cdot \frac{1}{p} \operatorname{Tr}\left[(\hat{a}_1 \hat{v}^2 M^\top M + \hat{a}_2 \operatorname{Id})^{-1} \right] \right) \tag{6.12}$$

$$\Delta_{bias} := (\beta_1 - \hat{v}\beta_2)^\top \Sigma_1^{1/2} \Pi(\hat{v}) \Sigma_1^{1/2} (\beta_1 - \hat{v}\beta_2). \tag{6.13}$$

The proof of Theorem 6.4 is based on Lemma 6.2, Lemma 6.3, and the following bound on the singular values of Z_1: for any fixed $\varepsilon > 0$, we have

$$\alpha_-(\rho_1) - \mathrm{O}(p^{-1/2+e}) \preceq n_1^{-1} Z_1^T Z_1 \preceq \alpha_+(\rho_1) + \mathrm{O}(p^{-1/2+e}) \quad \text{w.h.p.} \tag{6.14}$$

In fact, $n_1^{-1} Z_1^T Z_1$ is a standard sample covariance matrix, and it is well-known that its eigenvalues are inside the support of the Marchenko-Pastur law $[\alpha_-(\rho_1) - \mathrm{o}(1), \alpha_+(\rho_1) + \mathrm{o}(1)]$ with probability $1 - \mathrm{o}(1)$ (Bai and

Silverstein, 1998). The estimate (6.14) provides tight bounds for the concentration error. The result is formally stated in Lemma A.4 below.

The main error δ of Theorem 6.4 comes from approximating $n_1^{-1}Z_1^T Z_1$ by Id using (6.14); see the estimate (6.18) below. In order to improve this estimate and obtain an exact asymptotic result, one needs to study the singular value distribution of the following random matrix:

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

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. (Zheng et al., 2017). 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 $\mathcal{X} + \hat{v}^2$ will be even harder. We leave this problem to future study.

Proof of Theorem 6.4. Note that

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

We introduce the notation $\hat{M} \equiv \hat{M}(v) = v\Sigma_1^{1/2}\Sigma_2^{-1/2}$ for $v \in \mathbb{R}$. Then the proof is divided into the following four steps.

(i) We first consider $\hat{M}(v)$ for a fixed $v \in \mathbb{R}$. Then we use Lemma 6.1 and Lemma 6.2 to calculate the variance reduction $\delta_{\text{var}}(v)$, which will lead to the Δ_{var} term.

(ii) Using the approximate isometry property of Z_1 in (6.14), we will bound the bias term $\delta_{\text{bias}}(v)$ through

$$\widetilde{\delta}_{\text{bias}}(v) := v^2 n_1^2\left\|\Sigma_2^{1/2}(v^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\Sigma_1(\beta_1 - v\beta_2)\right\|^2. \tag{6.15}$$

(iii) We use Lemma 6.3 to calculate (6.15), which will lead to the Δ_{bias} term.

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

Step I: Variance reduction. Consider $\hat{M}(v)$ for any fixed $v \in \mathbb{R}$. Using Lemma 6.2, we can obtain that for any constant $\varepsilon > 0$,

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

and

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

with high probability, where $\hat{a}_1$ and $\hat{a}_2$ satisfy (6.6) with $\hat{v}$ replaced by v. Combining them, we get

$$\delta_{\text{var}}(v) = \Delta_{\text{var}}(v) + \mathrm{O}(\sigma^2 p^{-1/2+e}) \quad \text{w.h.p.}, \tag{6.16}$$

where $\Delta_{\text{var}}(v)$ is defined as in (6.12) but with $\hat{v}$ replaced by v.

Step II: Bounding the bias term. Next we use (6.14) to approximate $\delta_{\text{bias}}(v)$ with $\widetilde{\delta}_{\text{bias}}(v)$.

Claim 6.5. *In the setting of Theorem 6.4, we denote by $K = (v^2 X_1^\top X_1 + X_2^\top X_1)^{-1}$, and*

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

Then we have w.h.p.

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

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

$$\begin{aligned}\delta_{\mathrm{bias}}(v) - \widetilde{\delta}_{\mathrm{bias}}(v) &= 2v^2 n_1(\beta_1 - v\beta_2)^\top \Sigma_1^{1/2}\mathcal{E}\left(\Sigma_1^{1/2}K\Sigma_2K\Sigma_1^{1/2}\right)\Sigma_1^{1/2}(\beta_1 - v\beta_2) \\ &\quad + v^2\left\|\Sigma_2^{1/2}K\Sigma_1^{1/2}\mathcal{E}\Sigma_1^{1/2}(\beta_1 - v\beta_2)\right\|^2. \end{aligned} \tag{6.17}$$

Using (6.14), we can bound

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

Thus we can estimate that

$$\begin{aligned}|\delta_{\mathrm{bias}}(v) - \widetilde{\delta}_{\mathrm{bias}}(v)| &\leqslant v^2\left(2n_1\|\mathcal{E}\| + \|\mathcal{E}\|^2\right)\left\|\Sigma_1^{1/2}K\Sigma_2K\Sigma_1^{1/2}\right\|\left\|\Sigma_1^{1/2}(\beta_1 - v\beta_2)\right\|^2 \\ &= v^2\left[(n_1 + \|\mathcal{E}\|)^2 - n_1^2\right]\left\|\Sigma_1^{1/2}K\Sigma_2K\Sigma_1^{1/2}\right\|\left\|\Sigma_1^{1/2}(\beta_1 - v\beta_2)\right\|^2 \\ &\leqslant v^2n_1^2\left[\alpha_+^2(\rho_1) + \mathrm{O}(p^{-1/2+\varepsilon}) - 1\right]\left\|\Sigma_1^{1/2}K\Sigma_2K\Sigma_1^{1/2}\right\|\left\|\Sigma_1^{1/2}(\beta_1 - v\beta_2)\right\|^2,\end{aligned}$$

which concludes the proof by the definition of δ_ε. □

Note by (6.14), we have with high probability,

$$\begin{aligned}&v^2n_1^2\Sigma_1^{1/2}K\Sigma_2K\Sigma_1^{1/2} = n_1^2\hat{M}(\hat{M}^\top Z_1^\top Z_1\hat{M} + Z_2^\top Z_2)^{-2}\hat{M}^\top \\ &\preceq n_1^2\hat{M}\left[n_1\alpha_-(\rho_1)\hat{M}^\top\hat{M} + n_2\alpha_-(\rho_2) + \mathrm{O}(p^{1/2+\varepsilon})\right]^{-2}\hat{M}^\top \\ &\preceq \left[\alpha_-^2(\rho_1)\hat{M}\hat{M}^\top + 2\frac{\rho_2}{\rho_1}\alpha_-(\rho_1)\alpha_-(\rho_2) + 2\left(\frac{\rho_2}{\rho_1}\right)^2\alpha_-^2(\rho_2)(\hat{M}\hat{M}^\top)^{-1}\right]^{-1} + \mathrm{O}(p^{-1/2+\varepsilon}) \\ &\preceq [\alpha_-^2(\rho_1)\lambda_{\min}^2(\hat{M})]^{-1}\cdot(1-c)\end{aligned}$$

for some small enough constant $c > 0$. Together with Claim 6.5, we get with high probability,

$$\left|\delta_{\mathrm{bias}}(v) - \widetilde{\delta}_{\mathrm{bias}}(v)\right| \leqslant (1-c)\delta(v) \tag{6.18}$$

for some small constant $c > 0$, where we recall $\delta(v)$ defined in (6.9).

Step III: The limit of $\widetilde{\delta}_{\mathrm{bias}}(v)$. Using Lemma 6.3 with Σ_1 and M replaced by $v^2\Sigma_1$ and $\hat{M}$, we obtain that

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

with high probability. Together with and (6.16) and (6.18), we obtain that w.h.p.,

$$\begin{cases}\delta_{\mathrm{var}}(v) > \delta_{\mathrm{bias}}(v), & \text{if} \quad \Delta_{\mathrm{var}}(v) - \Delta_{\mathrm{bias}}(v) \geqslant \delta(v), \\ \delta_{\mathrm{var}}(v) < \delta_{\mathrm{bias}}(v), & \text{if} \quad \Delta_{\mathrm{var}}(v) - \Delta_{\mathrm{bias}}(v) \leqslant -\delta(v).\end{cases} \tag{6.19}$$

Step IV: An ε-net argument. Finally, it remains to extend the above result (6.19) 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 (6.19) holds uniformly for all $v \in [-C_0, C_0]$. In fact, we consider v belonging to a discrete set

$$V := \{v_k = kp^{-1} : -(C_0p+1) \leqslant k \leqslant C_0p + 1\}.$$

Then using the arguments for the first three steps and a simple union bound, we get that (6.19) holds simultaneously for all $v \in V$ with high probability. On the other hand, by (6.14) 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:

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

Then a simple application of triangle inequality gives that the event

$$\Xi_2 = \{(6.19) \text{ holds simultaneously for all } -C_0 \leqslant v \leqslant C_0\}$$

holds with high probability. On the other hand, on Ξ_1 one can see that for any small constant $\varepsilon > 0$, there exists a large enough constant $C_0 > 0$ depending on ε such that

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

for all $v \geqslant C_0$. Similar estimates hold for $v \leqslant -C_0$ if we replace C_0 with $-C_0$ in the above estimates. Together with the estimate at $\pm C_0$, we get that (6.19) holds simultaneously for all $v \in \mathbb{R}$ on the high probability event $\Xi_1 \cap \Xi_2$. This concludes the proof since v must be one of the real values. □

For the isotropic model in Section 4.1, we actually have an easier and sharper bound than Theorem 6.4 as follows.

Lemma 6.6 (The isotropic model)**.** *In the setting of Theorem 6.4, 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 $\Delta^\star_{bias} := \left((1-\hat{v})^2\kappa^2 + d^2\right)\mathrm{Tr}\left[\Pi(\hat{v})\right]$. Then we have with high probability,*

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

Proof. The proof of Lemma 6.6 is similar to that for Theorem 6.4, except that we can replace (6.18) with a tighter bound. We only describe the main difference.

For any fixed $v \in \mathbb{R}$, $\beta_1 - v\beta_2$ 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 A.14, we get that for any constant $\varepsilon > 0$,

$$\begin{aligned}
&\left|\delta_{\mathrm{bias}}(v) - [(1-v)^2\kappa^2 + d^2]\,\mathrm{Tr}(\mathcal{K}^\top\mathcal{K})\right| \\
&= \left|(\beta_1 - v\beta_2)^\top\mathcal{K}^\top\mathcal{K}(\beta_1 - v\beta_2) - [(1-v)^2\kappa^2 + d^2]\,\mathrm{Tr}(\mathcal{K}^\top\mathcal{K})\right| \\
&\leqslant p^\varepsilon[(1-v)^2\kappa^2 + d^2]\left\{\mathrm{Tr}\left[(\mathcal{K}^\top\mathcal{K})^2\right]\right\}^{1/2} \lesssim p^{1/2+\varepsilon}[(1-v)^2\kappa^2 + d^2],
\end{aligned} \tag{6.20}$$

where we denoted $\mathcal{K} := v\Sigma_2^{1/2}(v^2X_1^\top X_1 + X_2^\top X_2)^{-1}X_1^\top X_1$, and in the last step we used $\|\mathcal{K}\| = \mathrm{O}(1)$ by (6.14). Now for $\mathrm{Tr}(\mathcal{K}^\top\mathcal{K})$, we rewrite it as

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

Bounding $(X_1^\top X_1)^2 = (Z_1^\top Z_1)^2$ using (6.14) again, we obtain that

$$\delta^\star_{\mathrm{bias}}(v) \cdot (\alpha_-^2(\rho_1) - \mathrm{O}(p^{-1/2+\varepsilon})) \leqslant [(1-v)^2\kappa^2 + d^2]\,\mathrm{Tr}(\mathcal{K}^\top\mathcal{K}) \leqslant \delta^\star_{\mathrm{bias}}(v) \cdot (\alpha_+^2(\rho_1) + \mathrm{O}(p^{-1/2+\varepsilon})), \tag{6.21}$$

where

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

Note that $\delta^{\star}_{\text{bias}}(v) \sim p$, hence combining (6.20) and (6.21) we get

$$\delta^{\star}_{\text{bias}}(v) \cdot (\alpha_-^2(\rho_1) - \mathrm{O}(p^{-1/2+\varepsilon})) \leqslant \delta_{\text{bias}}(v) \leqslant \delta^{\star}_{\text{bias}}(v) \cdot (\alpha_+^2(\rho_1) + \mathrm{O}(p^{-1/2+\varepsilon})). \tag{6.22}$$

Now we can replace the estimate (6.18) with this stronger estimate, and repeat all the other parts of the proof of Theorem 6.4 to conclude Lemma 6.6. In particular, one can calculate $\delta^{\star}_{\text{bias}}(v)$ using Lemma 6.3 and get the $\Delta^{\star}_{\text{bias}}(v)$ term, We omit the details. □

6.2 Multiple Tasks

Proof of Theorem 3.3. In this setting, we need to study the following loss function:

$$f(B; W_1, \dots, W_t) = \sum_{i=1}^{t} \|XBW_i - Y_i\|^2. \tag{6.23}$$

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

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

where we denote $\mathcal{W} \in \mathbb{R}^{r \times t}$ as $\mathcal{W} = [W_1, W_2, \dots, W_t]$. Then as in (2.9), 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

$$g(\mathcal{W}) = N \left[val(\mathcal{W}) + t\sigma^2 \right] \cdot \left(1 + \mathrm{O}(p^{-(1-\varepsilon_0)/2+\varepsilon}) \right). \tag{6.24}$$

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

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

It suffices to minimize $\delta_{\text{bias}}(\mathcal{W})$ over $\mathcal{W}$, since both $t\sigma^2$ and δ_{var} do not depend on $\mathcal{W}$.

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 $Q_0 \equiv \mathcal{W}_0^\top (\mathcal{W}_0 \mathcal{W}_0^\top)^{-1} \mathcal{W}_0 = U_r U_r^\top$. On the other hand, let $\hat{\mathcal{W}}$ be the minimizer of g, 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

$$\|Q_0^{-1} \hat{Q} - \mathrm{Id}\,\| = \mathrm{o}(1) \quad \text{w.h.p.} \tag{6.25}$$

In fact, if (6.25) 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 (6.14), we obtain that

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

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

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

with high probability, where we used Lemma 6.1 in the last step. Similar, by Lemma 6.1 we have

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

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

7 Conclusions and Open Problems

In this work, we analyzed the bias and variance of multi-task learning versus single-task learning. We provided tight concentration bounds for the bias and the variance. Based on these bounds, we analyzed the impact of three properties, including task similarity, sample size, and covariate shift on the bias and variance, to derive conditions for transfer. We validated our theoretical results. Based on the theory, we proposed to train multi-task models by incrementally adding labeled data and showed encouraging results inspired by our theory. We describe several open questions for future work. First, our bound on the bias term (cf. Lemma 6.3) involves an error term that scales down with ρ_1. Tightening this error bound might cover the unexplained observations in Figure 1. Second, it would be interesting to extend our results to non-linear settings. We remark that this likely requires addressing significant technical challenges to deal with non-linearity.

References

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

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

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

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.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

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.

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

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

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

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

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

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

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

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pages 129–136, 2008.

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.

Brian Bullins, Elad Hazan, Adam Kalai, and Roi Livni. Generalize across tasks: Efficient algorithms for linear representation learning. In *Algorithmic Learning Theory*, pages 235–246, 2019.

Rich Caruana. Multitask learning: A knowledge-based source of inductive bias. *Proceedings of the Tenth International Conference on Machine Learning.*

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

Giovanni Cavallanti, Nicolo Cesa-Bianchi, and Claudio Gentile. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11(Oct):2901–2934, 2010.

Minmin Chen, Kilian Q Weinberger, and John Blitzer. Co-training for domain adaptation. In *Advances in neural information processing systems*, pages 2456–2464, 2011.

Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Incremental learning-to-learn with statistical guarantees. *arXiv preprint arXiv:1803.08089*, 2018.

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.

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

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, 2013b.

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

Theodoros Evgeniou, Charles A Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of machine learning research*, 6(Apr):615–637, 2005.

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. URL `http://stanford.edu/~gangus/res/petct.pdf`.

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

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

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

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.

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, 2018a.

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, 2018b.

Xin Li and Dan Roth. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

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

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.

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.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in neural information processing systems*, pages 1041–1048, 2009.

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

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

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.

Daniel McNamara and Maria-Florina Balcan. Risk bounds for transferring representations with and without fine-tuning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2373–2381. JMLR. org, 2017.

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

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

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

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

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.

Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4763–4771, 2019.

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

Vadim Ivanovich Serdobolskii. *Multiparametric statistics*. Elsevier, 2007.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

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

Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Revisiting multi-task learning in the deep learning era. *arXiv preprint arXiv:2004.13379*, 2020.

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.

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.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

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

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.

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

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.

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.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

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

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532, 2019.

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.

A Proofs of the Bias and Variance Bounds

The main goal of this section is to prove Lemma 6.2 and Lemma 6.3. In random matrix theory, it is much more convenient to rescale the matrices Z_1 and Z_2 such that their entries have variance n^{-1}, where $n := n_1 + n_2$. The advantage of this scaling is that the singular eigenvalues of Z_1 and Z_2 all lie in a bounded support that does grow with n. For reader's convenience, we now restate the setting with rescaled Z_1 and Z_2, and introduce a couple of other notations.

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 matrices with i.i.d. entries satisfying

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

Moreover, we assume that the fourth moments exist:

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

for some constant $C > 0$. Let $0 < \tau < 1$ be a small constant. We assume that the aspect ratios $\rho_1 = n_1/p$ and $\rho_2 = n_2/p$ satisfy that

$$0 \leqslant \rho_1 \leqslant \tau^{-1}, \quad 1 + \tau \leqslant \rho_2 \leqslant \tau^{-1}. \tag{A.3}$$

Here the lower bound $1 + \tau \leqslant \rho_2$ is to ensure that the sample covariance matrix $X_2^\top X_2$ is non-singular with high probability; see Lemma A.4 below. We assume that Σ_1 and Σ_2 have eigendecompositions

$$\Sigma_1 = O_1 \Lambda_1 O_1^\top, \;\; \Sigma_2 = O_2 \Lambda_2 O_2^\top, \;\; \Lambda_1 = \mathrm{diag}(\sigma_1^{(1)}, \ldots, \sigma_p^{(1)}), \;\; \Lambda_2 = \mathrm{diag}(\sigma_1^{(2)}, \ldots, \sigma_p^{(2)}), \tag{A.4}$$

where the eigenvalues satisfy that

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

such that Σ_1 and Σ_2 are both well-conditioned. We assume that $M = \Sigma_1^{1/2} \Sigma_2^{-1/2}$ has a singular value decomposition

$$M = U \Lambda V^\top, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p), \tag{A.6}$$

where by (A.5) we have

$$\tau \leqslant \lambda_p \leqslant \lambda_1 \leqslant \tau^{-1}. \tag{A.7}$$

We summarize our basic assumptions here for future reference. Note that this assumption is in accordance with the assumptions of Lemma 6.2, except that we rescale the entries of Z_1 and Z_2 and allow ρ_1 to be smaller than 1.

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

For simplicity of notations, we will use the following notion of stochastic domination, which was first introduced in (Erdős et al., 2013a) and subsequently used in many works on random matrix theory. It greatly 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 A.2 (Stochastic domination)**.** *Let $\xi \equiv \xi^{(n)}$ and $\zeta \equiv \zeta^{(n)}$ be two n-dependent random variables.*

(i) We say ξ is stochastically dominated by ζ, denoted by $\xi \prec \zeta$ or $\xi = \mathrm{O}_\prec(\zeta)$, if for any (small) constant $\varepsilon > 0$ and (large) constant $D > 0$,

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

for large enough $n \geqslant n_0(\varepsilon, D)$. Moreover, if ξ and ζ depend on a parameter u, then we say ξ is stochastically dominated by ζ uniformly in $u \in \mathcal{U}$, if for any constants $\varepsilon, D > 0$,

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

for large enough $n \geqslant n_0(\varepsilon, D)$.

(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.

Then we introduce the following bounded support condition.

Definition A.3. *We say a random matrix Z satisfies the bounded support condition with q, if*

$$\max_{i,j} |Z_{ij}| \prec q. \tag{A.8}$$

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 (A.8) holds, we say that X has support q.

Note that if the entries of $\sqrt{n}Z$ have finite moments up to any order as in (2.5), then using Markov's inequality one can show that Z has bounded support $n^{-1/2}$. More generally, if the entries of $\sqrt{n}Z$ have finite a-th moment for some $a > 4$, then by Markov's inequality and a simple union bound, it is easy to see that Z has bounded support $q = n^{2/a-1/2}$ on a event with probability $1 - \mathrm{o}(1)$. Hence the bounded support condition allows us to relax the moment conditions using a simple cut-off argument; see Corollary A.8 below.

For Z_1 and Z_2 with bounded support q, we have the following estimates on their singular values. We have used it in our previous proofs; see (6.14).

Lemma A.4. *Suppose Assumption A.1 holds, and Z_1, Z_2 satisfy the bounded support condition* (A.8) *for some deterministic parameter $q \equiv q(n)$ satisfying $n^{-1/2} \leqslant q \leqslant n^{-\phi}$ for a constant $\phi > 0$. Then for any constant $\varepsilon > 0$, we have that with high probability,*

$$\lambda_1(Z_1^\top Z_1) \leqslant \frac{(\sqrt{n_1} + \sqrt{p})^2}{n} + n^\varepsilon q, \tag{A.9}$$

and

$$\frac{(\sqrt{n_2} - \sqrt{p})^2}{n} - n^\varepsilon q \leqslant \lambda_p(Z_2^\top Z_2) \leqslant \lambda_1(Z_2^\top Z_2) \leqslant \frac{(\sqrt{n_2} + \sqrt{p})^2}{n} + n^\varepsilon q. \tag{A.10}$$

where $\lambda_i(Z_k^\top Z_k)$, $k = 1, 2$ and $i = 1, \cdots, p$, is the i-th largest eigenvalue of $Z_k^\top Z_k$.

Proof. This lemma essentially follows from (Bloemendal et al., 2014, Theorem 2.10), although the authors considered the case with $q \prec n^{-1/2}$ only. The results for larger q follows from (Ding and Yang, 2018, Lemma 3.12), but only the bounds for 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 (A.3), the lower bound for the smallest eigenvalue follows from the same arguments as in (Ding and Yang, 2018). Hence we omit the details. □

The rest of the proof is organized as follows. In Section A.1, we introduce the concept of resolvents, and give an almost optimal convergent estimate on it—Theorem A.7. This estimate is conventionally called *local law* in random matrix theory literature. Based on Theorem A.7, we then complete the proof of Lemma 6.2 and Lemma 6.3. The proof of Theorem A.7 is presented in Section A.2.

A.1 Resolvent and Local Law

Our main goal is to study the matrix inverse $(X_1^\top X_1 + X_2^\top X_2)^{-1}$ for $X_1 = \sqrt{n} Z_1 \Sigma_1^{1/2}$ and $X_2 = \sqrt{n} Z_2 \Sigma_2^{1/2}$. Using (A.6), we can rewrite it as

$$(X_1^\top X_1 + X_2^\top X_2)^{-1} = n^{-1} \Sigma_2^{-1/2} V \left(\Lambda U^\top Z_1^\top Z_1 U \Lambda + V^\top Z_2^\top Z_2 V \right)^{-1} V^\top \Sigma_2^{-1/2}. \tag{A.11}$$

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

$$\mathcal{G}(z) := \left(\Lambda U^\top Z_1^\top Z_1 U \Lambda + V^\top Z_2^\top Z_2 V - z \right)^{-1}, \quad z \in \mathbb{C}_+, \tag{A.12}$$

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

Next we introduce a convenient self-adjoint linearization trick. It has been proved to be useful in studying the local laws of random matrices of the Gram type (Knowles and Yin, 2016; Alt et al., 2017; Xi et al., 2017). We define the following $(p+n) \times (p+n)$ self-adjoint block matrix, which is a linear function of Z_1 and Z_2:

$$H \equiv H(Z_1, Z_2) := \begin{pmatrix} 0 & \Lambda U^\top Z_1^\top & V^\top Z_2^\top \\ Z_1 U \Lambda & 0 & 0 \\ Z_2 V & 0 & 0 \end{pmatrix}. \tag{A.13}$$

Then we define its resolvent as

$$G \equiv G(Z_1, Z_2, z) := \left[H(Z_1, Z_2) - \begin{pmatrix} z \, \mathrm{Id}_p & 0 & 0 \\ 0 & \mathrm{Id}_{n_1} & 0 \\ 0 & 0 & \mathrm{Id}_{n_2} \end{pmatrix} \right]^{-1}, \quad z \in \mathbb{C}_+. \tag{A.14}$$

For simplicity of notations, we define the index sets

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

We will consistently use the latin letters $i, j \in \mathcal{I}_1$, greek letters $\mu, \nu \in \mathcal{I}_2 \cup \mathcal{I}_3$, and $\mathfrak{a}, \mathfrak{b} \in \mathcal{I}$. Correspondingly, the indices of the matrices Z_1 and Z_2 are labelled as

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

For simplicity, we abbreviate $W := (\Lambda^\top U^\top Z_1^\top, V^\top Z_2^\top)$. By Schur complement formula, we can obtain that

$$G(z) := \begin{pmatrix} \mathcal{G}(z) & \mathcal{G}(z) W \\ W^\top \mathcal{G}(z) & z(W^\top W - z)^{-1} \end{pmatrix}^{-1}. \tag{A.15}$$

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

$$\begin{aligned} m(z) &:= \frac{1}{p} \sum_{i \in \mathcal{I}_1} G_{ii}(z), \quad m_1(z) := \frac{1}{p} \sum_{i \in \mathcal{I}_1} \lambda_i^2 G_{ii}(z), \\ m_2(z) &:= \frac{1}{n_1} \sum_{\mu \in \mathcal{I}_2} G_{\mu\mu}(z), \quad m_3(z) := \frac{1}{n_2} \sum_{\mu \in \mathcal{I}_3} G_{\mu\mu}(z). \end{aligned} \tag{A.16}$$

Our proof will use the spectral decomposition of G. Let

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

be a singular value decomposition of W, where $\{\xi_k\}_{k=1}^p$ are the left-singular vectors, and $\{\zeta_k\}_{k=1}^n$ are the right-singular vectors. Then using (A.15), we get that for $i, j \in \mathcal{I}_1$ and $\mu, \nu \in \mathcal{I}_2 \cup \mathcal{I}_3$,

$$G_{ij} = \sum_{k=1}^{p} \frac{\xi_k(i) \xi_k^\top(j)}{\mu_k - z}, \quad G_{\mu\nu} = z \sum_{k=1}^{n} \frac{\zeta_k(\mu) \zeta_k^\top(\nu)}{\mu_k - z}, \quad G_{i\mu} = G_{\mu i} = \sum_{k=1}^{p} \frac{\sqrt{\mu_k} \xi_k(i) \zeta_k^\top(\mu)}{\mu_k - z}. \tag{A.17}$$

We now describe the asymptotic limit of $\mathcal{G}(z)$ as $n \to \infty$. First we define the deterministic limits of $(m_2(z), m_3(z))$, denoted by $(m_{2c}(z), m_{3c}(z))$, as the unique solution to the following system of equations

$$\frac{1}{m_{2c}} = \frac{\gamma_n}{p} \sum_{i=1}^{p} \frac{\lambda_i^2}{z + \lambda_i^2 r_1 m_{2c} + r_2 m_{3c}} - 1, \quad \frac{1}{m_{3c}} = \frac{\gamma_n}{p} \sum_{i=1}^{p} \frac{1}{z + \lambda_i^2 r_1 m_{2c} + r_2 m_{3c}} - 1, \tag{A.18}$$

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

$$\gamma_n := \frac{p}{n}, \quad r_1 \equiv r_1(n) := \frac{n_1}{n}, \quad r_2 \equiv r_2(n) := \frac{n_2}{n}. \tag{A.19}$$

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

$$\Pi(z) := \begin{pmatrix} -(z + r_1 m_{2c}(z)\Lambda^2 + r_2 m_{3c}(z))^{-1} & 0 & 0 \\ 0 & m_{2c}(z)\,\mathrm{Id}_{n_1} & 0 \\ 0 & 0 & m_{3c}(z)\,\mathrm{Id}_{n_2} \end{pmatrix}. \tag{A.20}$$

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 (A.18) in are reduced to

$$r_1 x_2 + r_2 x_3 = 1 - \gamma_n, \quad x_2 + \frac{1}{n} \sum_{i=1}^{p} \frac{\lambda_i^2 x_2}{\lambda_i^2 r_1 x_2 + (1 - \gamma_n - r_1 x_2)} = 1. \tag{A.21}$$

where $x_2 := -m_{2c}(0)$ and $x_3 := -m_{3c}(0)$. Note that the function

$$f(x_2) := x_2 + \frac{1}{n} \sum_{i=1}^{p} \frac{\lambda_i^2 x_2}{\lambda_i^2 x_2 + (1 - \gamma_n - r_1 x_2)}$$

is a strictly increasing function on $[0, r_1^{-1}(1-\gamma_n)]$. Moreover, we have $f(0) = 0 < 1$ and $f(r_1^{-1}(1-\gamma_n)) = r_1^{-1} > 1$. Hence by mean value theorem, there exists a unique solution $x_2 \in (0, r_1^{-1}(1-\gamma_n))$. 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

$$\tau \leqslant x_2 \leqslant \min\{r_1^{-1}(1-\gamma_n) - \tau, 1 - \tau\}, \quad \tau < r_2 x_3 \leqslant 1 - \gamma_n - r_1 \tau. \tag{A.22}$$

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.

Lemma A.5. *There exist constants $c_0, C_0 > 0$ depending only on τ in* (A.3), (A.5), (A.7) *and* (A.22) *such that the following statements hold. There exists a unique solution to* (A.18) *under the conditions*

$$|z| \leqslant c_0, \quad |m_{2c}(z) - m_{2c}(0)| + |m_{3c}(z) - m_{3c}(0)| \leqslant c_0. \tag{A.23}$$

Moreover, the solution satisfies

$$|m_{2c}(z) - m_{2c}(0)| + |m_{3c}(z) - m_{3c}(0)| \leqslant C_0 |z|. \tag{A.24}$$

The proof is a standard application of the contraction principle. For reader's convenience, we will include its proof in Appendix A.2.4. As a byproduct of the contraction mapping argument there, we also obtain the following stability result that will be used in the proof of Theorem A.7.

Lemma A.6. *There exist constants $c_0, C_0 > 0$ depending only on τ in* (A.3), (A.5), (A.7) *and* (A.22) *so that the self-consistent equations in* (A.18) *are stable in the following sense. Suppose $|z| \leqslant c_0$ and $m_k : \mathbb{C}_+ \mapsto \mathbb{C}_+$, $k = 2, 3$, 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

$$\frac{1}{m_2}+1-\frac{\gamma_n}{p}\sum_{i=1}^p\frac{\lambda_i^2}{z+\lambda_i^2 r_1 m_2+r_2 m_3}=\mathcal{E}_2,\quad \frac{1}{m_3}+1-\frac{\gamma_n}{p}\sum_{i=1}^p\frac{1}{z+\lambda_i^2 r_1 m_2+r_2 m_3}=\mathcal{E}_3, \tag{A.25}$$

for some (random) errors satisfying $|\mathcal{E}_2|+|\mathcal{E}_3|\leqslant\delta(z)$, *where* $\delta(z)$ *is a deterministic* z*-dependent function satisfying* $\delta(z)\leqslant(\log n)^{-1}$. *Then we have*

$$|m_2(z)-m_{2c}(z)|+|m_3(z)-m_{3c}(z)|\leqslant C_0\delta(z). \tag{A.26}$$

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

$$\mathbf{D}:=\left\{z=E+\mathrm{i}\eta\in\mathbb{C}_+:|z|\leqslant(\log n)^{-1}\right\}. \tag{A.27}$$

The following theorem gives an almost optimal estimate on the resolvent G, which is conventionally called the *anisotropic local law.*

Theorem A.7. *Suppose Assumption A.1 holds, and* Z_1 *and* Z_2 *satisfy the bounded support condition* (A.8) *for a deterministic parameter* q *satisfying* $n^{-1/2}\leqslant q\leqslant n^{-\phi}$ *for some 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* $\mathbf{u},\mathbf{v}\in\mathbb{R}^{p+n_1+n_2}$, *we have*

$$\left|\mathbf{u}^\top(G(z)-\Pi(z))\mathbf{v}\right|\prec q. \tag{A.28}$$

The proof of this theorem will be given in Section A.2. We now use it to complete the proof of Lemma 6.2 and Lemma 6.3.

Proof of Lemma 6.2. In the setting of Lemma 6.2, using (A.11) and (A.12) we can write

$$\mathcal{R}:=(X_1^\top X_1+X_2^\top X_2)^{-1}=n^{-1}\Sigma_2^{-1/2}V\mathcal{G}(0)V^\top\Sigma_2^{-1/2}.$$

If the entries of $\sqrt{n}Z_1$ and $\sqrt{n}Z_2$ have arbitrarily high moments as in (2.5), then Z_1 and Z_2 have bounded support $q=n^{-1/2}$. Using Theorem A.7, we obtain that for any small constant $\varepsilon>0$,

$$\max_{1\leqslant i\leqslant p}|(A\mathcal{R}-n^{-1}A\Sigma_2^{-1/2}V\Pi(0)V^\top\Sigma_2^{-1/2})_{ii}|\prec n^{-3/2}\|A\|, \tag{A.29}$$

where by (A.20),

$$\Pi(0)=-(r_1m_{2c}(0)\Lambda^2+r_2m_{3c}(0))^{-1}=(r_1x_2V^\top M^\top MV+r_2x_3)^{-1},$$

with (x_2,x_3) satisfying (A.21). Thus from (A.29) we get that

$$\operatorname{Tr}(A\mathcal{R})=n^{-1}\operatorname{Tr}(r_1x_2M^\top M+r_2x_3)^{-1}+\mathrm{O}_\prec(n^{-1/2}\|A\|).$$

This concludes (6.2) if we rename (r_1x_2,r_2x_3) to (a_1,a_2).

Note that if we set $n_1=0$ and $n_2=n$, then $a_1=0$ and $a_2=(n_2-p)/n_2$ is the solution to (6.3). This gives (6.1) using (6.2). □

Proof of Lemma 6.3. In the setting of Lemma 6.3, we can write

$$\Delta:=n^2\left\|\Sigma_2^{1/2}(X_1^\top X_1+X_2^\top X_2)^{-1}\beta\right\|^2=\beta^\top\Sigma_2^{-1/2}\left(M^\top Z_1^\top Z_1M+Z_2^\top Z_2\right)^{-2}\Sigma_2^{-1/2}\beta,$$

where in the second step the n^2 factor disappeared due to the choice of scaling in (A.1). With (A.6), we can write the above expression as

$$\Delta=\mathbf{v}^\top(\mathcal{G}^2)(0)\,\mathbf{v},\quad \mathbf{v}:=V^\top\Sigma_2^{-1/2}\beta.$$

Note that $\mathcal{G}^2(0) = \partial_z \mathcal{G}|_{z=0}$. Now using Cauchy's integral formula and Theorem A.7, we get that

$$\begin{aligned} \mathbf{v}^\top \mathcal{G}^2(0)\,\mathbf{v} = \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{\mathbf{v}^\top \mathcal{G}(z)\,\mathbf{v}}{z^2} \mathrm{d}z &= \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{\mathbf{v}^\top \Pi(z)\,\mathbf{v}}{z^2} \mathrm{d}z + \mathrm{O}_{\prec}(n^{-1/2}\|\beta\|^2) \\ &= \mathbf{v}^\top \Pi'(0)\,\mathbf{v} + \mathrm{O}_{\prec}(n^{-1/2}\|\beta\|^2), \end{aligned} \tag{A.30}$$

where $\mathcal{C}$ is the contour $\{z \in \mathbb{C} : |z| = (\log n)^{-1}\}$. Hence it remains to study the derivatives

$$\mathbf{v}^\top \Pi'(0)\,\mathbf{v} = \mathbf{v}\, \frac{1 + r_1 m_{2c}'(0)\Lambda^2 + r_2 m_{3c}'(0)}{(r_1 m_{2c}(0)\Lambda^2 + r_2 m_{3c}(0))^2}\,\mathbf{v}, \tag{A.31}$$

where we need to calculate the derivatives $m_{2c}'(0)$ and $m_{3c}'(0)$.

Taking implicit differentiation of (A.18), we obtain that

$$\frac{m_{2c}'(0)}{m_{2c}^2(0)} = \frac{1}{n}\sum_{i=1}^p \frac{\lambda_i^2\left(1 + \lambda_i^2 r_1 m_{2c}'(0) + r_2 m_{3c}'(0)\right)}{(\lambda_i^2 r_1 m_{2c}(0) + r_2 m_{3c}(0))^2}, \quad \frac{m_{3c}'(0)}{m_{3c}^2(0)} = \frac{1}{n}\sum_{i=1}^p \frac{1 + \lambda_i^2 r_1 m_{2c}'(0) + r_2 m_{3c}'(0)}{(\lambda_i^2 r_1 m_{2c}(0) + r_2 m_{3c}(0))^2}.$$

If we rename $(-r_1 m_{2c}(0), -r_2 m_{3c}(0))$ to (a_1, a_2) and $(r_2 m_{3c}'(0), r_1 m_{2c}'(0))$ to (a_3, a_4), then these equations become

$$\begin{aligned} &\left(\frac{r_2}{a_2^2} - \frac{1}{n}\sum_{i=1}^p \frac{1}{(\lambda_i^2 a_1 + a_2)^2}\right) a_3 - \left(\frac{1}{n}\sum_{i=1}^p \frac{\lambda_i^2}{(\lambda_i^2 a_1 + a_2)^2}\right) a_4 = \frac{1}{n}\sum_{i=1}^p \frac{1}{(\lambda_i^2 a_1 + a_2)^2}, \\ &\left(\frac{r_1}{a_1^2} - \frac{1}{n}\sum_{i=1}^p \frac{\lambda_i^4}{(\lambda_i^2 a_1 + a_2)^2}\right) a_4 - \left(\frac{1}{n}\sum_{i=1}^p \frac{\lambda_i^2}{(\lambda_i^2 a_1 + a_2)^2}\right) a_3 = \frac{1}{n}\sum_{i=1}^p \frac{\lambda_i^2}{(\lambda_i^2 a_1 + a_2)^2}, \end{aligned} \tag{A.32}$$

which are equivalent to (6.5). Then by (A.30) and (A.31), we get

$$\Delta = \beta^\top \Sigma_2^{-1/2} V \frac{1 + a_3 + a_4\Lambda^2}{(a_1\Lambda^2 + a_2)^2} V^\top \Sigma_2^{-1/2}\beta = \beta^\top \Sigma_2^{-1/2} \frac{1 + a_3 + a_4 M^\top M}{(a_1 M^\top M + a_2)^2} \Sigma_2^{-1/2}\beta,$$

where we used $M^\top M = V\Lambda^2 V^\top$ in the second step. This concludes Lemma 6.3. □

Using a simple cutoff argument, it is easy to obtain from Theorem A.7 the following corollary under weaker moment assumptions.

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

$$\max_{i,j} \mathbb{E}|\sqrt{n} z_{ij}^{(k)}|^a = \mathrm{O}(1), \quad k = 1, 2, \tag{A.33}$$

for some fixed $a > 4$. Then (A.28) *holds for $q = n^{2/a - 1/2}$ on an event with probability* $1 - \mathrm{o}(1)$.

Proof. Fix any sufficiently small constant $\varepsilon > 0$. We 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

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

By the moment conditions (A.33) and a simple union bound, we have

$$\mathbb{P}(\widetilde{Z}_1 = Z_1, \widetilde{Z}_2 = Z_2) = 1 - \mathrm{O}(n^{-a\varepsilon}). \tag{A.34}$$

Using (A.33) and integration by parts, it is easy to verify that

$$|\mathbb{E}\widetilde{z}_{ij}^{(k)}| = \mathrm{O}(n^{-2-\varepsilon}), \quad \mathbb{E}|\widetilde{z}_{ij}^{(k)}|^2 = n^{-1} + \mathrm{O}(n^{-2-\varepsilon}), \quad k = 1, 2. \tag{A.35}$$

Then we can centralize and rescale $\widetilde{Z}_1$ and $\widetilde{Z}_2$ as $\widehat{Z}_k := (\widetilde{Z}_k - \mathbb{E}\widetilde{Z}_k)/(\mathbb{E}|\widetilde{z}_{11}^{(k)}|^2)^{1/2}$, $k = 1, 2$. Now $\widehat{Z}_1$ and $\widehat{Z}_2$ satisfy the assumptions in Theorem A.7 with $q = n^{-c_a+\varepsilon}$, and (A.28) gives that

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

Then using (A.35) and (A.37) below, it is obtain 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 $\|\mathbb{E}\widetilde{Z}_k\| = \mathrm{O}(n^{-1-\varepsilon})$, $k = 1, 2$, by (A.35). This shows that (A.28) also holds for $G(\widetilde{Z}_1, \widetilde{Z}_2, z)$ with $q = n^{-c_a+\varepsilon}$, which concludes the proof by (A.34) and the fact that ε can be chosen arbitrarily small. □

With this corollary, we can easily extend Lemma 6.2 and Lemma 6.3 to the case with weaker moment assumptions. Considering the length of this paper, we will not go into further details here.

A.2 Proof of the Anisotropic Local Law

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

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

where "$\overset{d}{=}$" means "equal in distribution". 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 problem can be handled using the standard resolvent methods as in e.g. (Bloemendal et al., 2014; Yang, 2019; Pillai and Yin, 2014). To go from the Gaussian case to the general X case, we will adopt a continuous self-consistent comparison argument developed in (Knowles and Yin, 2016).

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

$$G_0(z) := \begin{pmatrix} -z \,\mathrm{Id}_p & \Lambda Z_1^\top & Z_2^\top \\ Z_1 \Lambda & -\mathrm{Id}_{n_1} & 0 \\ Z_2 & 0 & -\mathrm{Id}_{n_2} \end{pmatrix}^{-1}, \quad z \in \mathbb{C}_+. \tag{A.36}$$

We will prove the following local law on G_0.

Proposition A.9. *Suppose Assumption A.1 holds, and Z_1 and Z_2 satisfy the bounded support condition* (A.8) *with $q = n^{-1/2}$. Suppose U and V are identity. Then the estimate* (A.28) *holds for $G_0(z)$.*

This section is organized as follows. In Section A.2.1, we collect some basic estimates and resolvent identities that will be used in the proof of Theorem A.7 and Proposition A.9. Then in Section A.2.2 we give the proof of Proposition A.9, which concludes Theorem A.7 when Z_1 and Z_2 have i.i.d. Gaussian entries. In Section A.2.3, we describe how to extend the result in Theorem A.7 from the Gaussian case to the case where the entries of Z_1 and Z_2 are generally distributed. Finally, in Section A.2.4, we give the proof of Lemma A.5 and Lemma A.6. In the proof, we always denote the spectral parameter by $z = E + \mathrm{i}\eta$.

A.2.1 Basic Estimates

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

First with Lemma A.4, we can obtain the following a priori estimate on the resolvent $G(z)$ for $z \in \mathbf{D}$.

Lemma A.10. *Suppose the assumptions of Lemma A.4 holds. Then there exists a constant $C > 0$ such that the following estimates hold uniformly in $z, z' \in \mathbf{D}$ with high probability:*

$$\|G(z)\| \leqslant C, \tag{A.37}$$

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

$$\left|\mathbf{u}^\top \left[G(z) - G(z')\right] \mathbf{v}\right| \leqslant C|z - z'|. \tag{A.38}$$

Proof. As in (A.17), we let $\{\lambda_k\}_{1\leqslant k\leqslant p}$ be the eigenvalues of $WW^\top$. By Lemma A.4 and the assumption (A.3), we obtain that $\lambda_p \geqslant \lambda_p(Z_2^\top Z_2) \gtrsim 1$, which further implies the estimate that $\inf_{z\in\mathbf{D}}\min_{1\leqslant k\leqslant p}|\lambda_k - z| \gtrsim 1$. Together with (A.17), we obtain (A.37) and (A.38). □

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

Lemma A.11 (Lemma 3.2 in (Bloemendal et al., 2014))**.** *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 $\mathbb{E}\xi(u)^2\leqslant n^C$. If $\xi(u)\prec\Psi(u)$ uniformly in u, then we also have $\mathbb{E}\xi(u)\prec\Psi(u)$ uniformly in u.

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

Definition A.12 (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* (A.15)*)*

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

and the partial traces $m^{(\mathbb{T})}$, $m_1^{(\mathbb{T})}$, $m_2^{(\mathbb{T})}$ and $m_3^{(\mathbb{T})}$ by replacing G with $G^{(\mathbb{T})}$ in (A.16)*. 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 $(\{a,b\})\equiv(ab)$.*

The following resolvent identities and the concentration bounds are the main tools for our proof.

Lemma A.13. *We have the following resolvent identities.*

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

$$\frac{1}{G_{ii}} = -z - \left(WG^{(i)}W^\top\right)_{ii}, \quad \frac{1}{G_{\mu\mu}} = -1 - \left(W^\top G^{(\mu)}W\right)_{\mu\mu}. \tag{A.39}$$

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

$$G_{i\mathfrak{a}} = -G_{ii}\left(WG^{(i)}\right)_{i\mathfrak{a}}, \quad G_{\mu\mathfrak{b}} = -G_{\mu\mu}\left(W^\top G^{(\mu)}\right)_{\mu\mathfrak{b}}. \tag{A.40}$$

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

$$G_{\mathfrak{b}\mathfrak{c}}^{(\mathfrak{a})} = G_{\mathfrak{b}\mathfrak{c}} - \frac{G_{\mathfrak{b}\mathfrak{a}}G_{\mathfrak{a}\mathfrak{c}}}{G_{\mathfrak{a}\mathfrak{a}}}, \quad \frac{1}{G_{\mathfrak{b}\mathfrak{b}}} = \frac{1}{G_{\mathfrak{b}\mathfrak{b}}^{(\mathfrak{a})}} - \frac{G_{\mathfrak{b}\mathfrak{a}}G_{\mathfrak{a}\mathfrak{b}}}{G_{\mathfrak{b}\mathfrak{b}}G_{\mathfrak{b}\mathfrak{b}}^{(\mathfrak{a})}G_{\mathfrak{a}\mathfrak{a}}}. \tag{A.41}$$

Proof. All these identities can be proved directly using Schur's complement formula. The reader can also refer to, for example, (Knowles and Yin, 2016, Lemma 4.4). □

Lemma A.14 (Lemma 3.8 of (Erdős et al., 2013b))**.** *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 variances at most n^{-1} and satisfy the bounded support condition (A.8) with $q \leqslant n^{-\phi}$ for some constant $\phi > 0$. Then we have the following bounds:*

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

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

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 (2.5), then we have stronger bounds

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

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

A.2.2 Entrywise Local Law

The main goal of this subsection is to prove the following entrywise local law. The anisotropic local law (A.28) then follows from the entrywise local law combined with a polynomialization method as we will explain later. Recall that in the setting of Proposition A.9, we have $q = n^{-1/2}$ and

$$W = (\Lambda Z_1^\top, Z_2^\top). \tag{A.42}$$

Lemma A.15. *Suppose the assumptions in Proposition A.9 hold. Then the following estimate holds uniformly for $z \in \mathbf{D}$:*

$$\max_{\mathfrak{a},\mathfrak{b}\in\mathcal{I}} |(G_0)_{\mathfrak{a}\mathfrak{b}}(z) - \Pi_{\mathfrak{a}\mathfrak{b}}(z)| \prec n^{-1/2}. \tag{A.43}$$

Proof. The proof of Lemma A.15 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 (A.42).

Step 1: Large deviations estimates. In this step, we prove some (almost) optimal large deviation estimates on the off-diagonal entries of G, and on the following Z variables. In analogy to (Erdős et al., 2013b, Section 3) and (Knowles and Yin, 2016, Section 5), we introduce the Z variables

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

where $\mathbb{E}_{\mathfrak{a}}[\cdot] := \mathbb{E}[\cdot \mid H^{(\mathfrak{a})}]$, i.e. it is the partial expectation over the randomness of the $\mathfrak{a}$-th row and column of H. Using (A.39), we get that for $i \in \mathcal{I}_1$, $\mu \in \mathcal{I}_2$ and $\nu \in \mathcal{I}_3$,

$$Z_i = \lambda_i^2 \sum_{\mu,\nu\in\mathcal{I}_2} G_{\mu\nu}^{(i)} \left(\frac{1}{n}\delta_{\mu\nu} - z_{\mu i} z_{\nu i}\right) + \sum_{\mu,\nu\in\mathcal{I}_3} G_{\mu\nu}^{(i)} \left(\frac{1}{n}\delta_{\mu\nu} - z_{\mu i} z_{\nu i}\right), \tag{A.44}$$

$$Z_\mu = \sum_{i,j\in\mathcal{I}_1} \sigma_i \sigma_j G_{ij}^{(\mu)} \left(\frac{1}{n}\delta_{ij} - z_{\mu i} z_{\mu j}\right), \quad Z_\nu = \sum_{i,j\in\mathcal{I}_1} G_{ij}^{(\nu)} \left(\frac{1}{n}\delta_{ij} - z_{\nu i} z_{\nu j}\right). \tag{A.45}$$

For simplicity, we introduce the random error $\Lambda_o := \max_{\mathfrak{a}\neq\mathfrak{b}} \left|G_{\mathfrak{a}\mathfrak{a}}^{-1} G_{\mathfrak{a}\mathfrak{b}}\right|$. The following lemma gives the desired large deviations estimates on Λ_o and the Z variables.

Lemma A.16. *Suppose the assumptions in Proposition A.9 hold. Then the following estimates hold uniformly for all $z \in \mathbf{D}$:*

$$\Lambda_o + \max_{\mathfrak{a}\in\mathcal{I}} |Z_{\mathfrak{a}}| \prec n^{-1/2}. \tag{A.46}$$

Proof. Note that for any $\mathfrak{a} \in \mathcal{I}$, $H^{(\mathfrak{a})}$ and $G^{(\mathfrak{a})}$ also satisfies the assumptions for Lemma A.10. Hence (A.37) and (A.38) also hold for $G^{(\mathfrak{a})}$. Now applying Lemma A.14 to (A.44) and (A.45), and using the a priori bound (A.37), we get that for any $i \in \mathcal{I}_1$,

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

where in the last step we used (A.37) to get that for any μ,

$$\sum_{\nu\in\mathcal{I}_2\cup\mathcal{I}_3} \left|G^{(i)}_{\mu\nu}\right|^2 \leqslant \sum_{\mathfrak{a}\in\mathcal{I}} \left|G^{(i)}_{\mu\mathfrak{a}}\right|^2 = \left[G^{(i)}(G^{(i)})^*\right]_{\mu\mu} = \mathrm{O}(1). \tag{A.47}$$

Similarly, applying Lemma A.14 to Z_μ and Z_ν in (A.45) and using (A.37), we obtain the same bound.

we have

$$G_{i\mathfrak{a}} = -G_{ii}\left(WG^{(i)}\right)_{i\mathfrak{a}}, \quad G_{\mu\mathfrak{b}} = -G_{\mu\mu}\left(W^\top G^{(\mu)}\right)_{\mu\mathfrak{b}}. \tag{A.48}$$

Then we prove the off-diagonal estimate on Λ_o. For $i \in \mathcal{I}_1$ and $\mathfrak{a} \in \mathcal{I}\setminus\{i\}$, using (A.40), Lemma A.14 and (A.37), we obtain that

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

We have a similar estimate for $\left|G_{\mu\mu}^{-1}G_{\mu\mathfrak{b}}\right|$ with $\mu \in \mathcal{I}_2\cup\mathcal{I}_3$ and $\mathfrak{b}\in\mathcal{I}\setminus\{\mu\}$. Thus we obtain that $\Lambda_o \prec n^{-1/2}$, which concludes (A.46). □

Note that comibining (A.37) and (A.46), we immediately conclude (A.43) for $\mathfrak{a} \neq \mathfrak{b}$.

Step 2: Self-consistent equations. This is the key step of the proof for Proposition A.15, 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 (A.25) for some small error $|\mathcal{E}_{2,3}| \prec n^{-1/2}$. Then in Step 3 we will apply Lemma A.6 to show that $(m_2(z), m_3(z))$ is close to $(m_{2c}(z), m_{3c}(z))$.

We define the following z-dependent event

$$\Xi(z) := \Big\{|m_2(z) - m_{2c}(z)| + |m_3(z) - m_{3c}(z)| \leqslant (\log n)^{-1/2}\Big\}. \tag{A.49}$$

Note that by (A.24), we have $|m_{2c}+x_2| \lesssim (\log n)^{-1}$ and $|m_{3c}+x_3| \lesssim (\log n)^{-1}$. Together with (A.18), (A.22) and (A.7), we obtain the following basic estimates

$$|m_{2c}| \sim |m_{3c}| \sim 1, \quad |z + \lambda_i^2 r_1 m_{2c} + r_2 m_{3c}| \sim 1, \quad |1+\gamma_n m_c| \sim |1+\gamma_n m_{1c}| \sim 1, \tag{A.50}$$

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

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

Plugging (A.50) into (A.20), we get

$$|\Pi_{\mathfrak{a}\mathfrak{a}}(z)| \sim 1 \ \text{ uniformly in } z\in\mathbf{D},\ \mathfrak{a}\in\mathcal{I}. \tag{A.51}$$

Then we claim the following result.

Lemma A.17. *Suppose the assumptions in Proposition A.9 hold. Then the following estimates hold uniformly in $z \in \mathbf{D}$:*

$$\begin{aligned} \mathbf{1}(\Xi)\left|\frac{1}{m_2} + 1 - \frac{1}{n}\sum_{i=1}^{p} \frac{\lambda_i^2}{z+\lambda_i^2 r_1 m_2 + r_2 m_3}\right| &\prec n^{-1/2}, \\ \mathbf{1}(\Xi)\left|\frac{1}{m_3} + 1 - \frac{1}{n}\sum_{i=1}^{p} \frac{1}{z+\lambda_i^2 r_1 m_2 + r_2 m_3}\right| &\prec n^{-1/2}. \end{aligned} \tag{A.52}$$

Proof. By (A.39), (A.44) and (A.45), we obtain that for $i \in \mathcal{I}_1$, $\mu \in \mathcal{I}_2$ and $\nu \in \mathcal{I}_3$,

$$\frac{1}{G_{ii}} = -z - \frac{\lambda_i^2}{n}\sum_{\mu\in\mathcal{I}_2} G^{(i)}_{\mu\mu} - \frac{1}{n}\sum_{\mu\in\mathcal{I}_3} G^{(i)}_{\mu\mu} + Z_i = -z - \lambda_i^2 r_1 m_2 - r_2 m_3 + \varepsilon_i, \tag{A.53}$$

$$\frac{1}{G_{\mu\mu}} = -1 - \frac{1}{n}\sum_{i\in\mathcal{I}_1} \lambda_i^2 G^{(\mu)}_{ii} + Z_\mu = -1 - \gamma_n m_1 + \varepsilon_\mu, \tag{A.54}$$

$$\frac{1}{G_{\nu\nu}} = -1 - \frac{1}{n}\sum_{i\in\mathcal{I}_1} G^{(\nu)}_{ii} + Z_\nu = -1 - \gamma_n m + \varepsilon_\nu, \tag{A.55}$$

where we recall Definition A.12, and

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

By (A.41) 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 (A.46) in the second step. Similarly, we can get that

$$|m - m^{(\mu)}| + |m_1 - m_1^{(\mu)}| + |m_2 - m_2^{(i)}| + |m_3 - m_3^{(i)}| \prec n^{-1} \tag{A.56}$$

for any $i \in \mathcal{I}_1$ and $\mu \in \mathcal{I}_2 \cup \mathcal{I}_3$. Together with (A.46), we obtain that for all i and μ,

$$|\varepsilon_i| + |\varepsilon_\mu| \prec n^{-1/2}. \tag{A.57}$$

With (A.50) and the definition of Ξ, we get that $\mathbf{1}(\Xi)|z + \lambda_i^2 r_1 m_2 + r_2 m_3| \sim 1$. Hence using (A.53), (A.57) and (A.46), we obtain that

$$\mathbf{1}(\Xi) G_{ii} = \mathbf{1}(\Xi)\left[-\frac{1}{z + \lambda_i^2 r_1 m_2 + r_2 m_3} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right]. \tag{A.58}$$

Plugging it into the definitions of m and m_1 in (A.16), we get

$$\mathbf{1}(\Xi) m = \mathbf{1}(\Xi)\left[-\frac{1}{p}\sum_{i\in\mathcal{I}_1}\frac{1}{z + \lambda_i^2 r_1 m_2 + r_2 m_3} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right], \tag{A.59}$$

$$\mathbf{1}(\Xi) m_1 = \mathbf{1}(\Xi)\left[-\frac{1}{p}\sum_{i\in\mathcal{I}_1}\frac{\lambda_i^2}{z + \lambda_i^2 r_1 m_2 + r_2 m_3} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right]. \tag{A.60}$$

As a byproduct, we obtain from these two estimates that

$$\mathbf{1}(\Xi)\left(|m - m_c| + |m_1 - m_{1c}|\right) \lesssim (\log n)^{-1/2}, \quad \text{with high probability.} \tag{A.61}$$

Together with (A.50), we get that

$$|1 + \gamma_n m_1| \sim 1, \quad |1 + \gamma_n m| \sim 1, \quad \text{with high probability on } \Xi. \tag{A.62}$$

Now using (A.54), (A.55), (A.57), (A.46) and (A.62), we obtain that for $\mu \in \mathcal{I}_2$ and $\nu \in \mathcal{I}_3$,

$$\mathbf{1}(\Xi)\left(G_{\mu\mu} + \frac{1}{1 + \gamma_n m_1}\right) = \mathrm{O}_\prec(n^{-1/2}), \quad \mathbf{1}(\Xi)\left(G_{\nu\nu} + \frac{1}{1 + \gamma_n m}\right) = \mathrm{O}_\prec(n^{-1/2}). \tag{A.63}$$

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

$$\mathbf{1}(\Xi)\left(m_2 + \frac{1}{1 + \gamma_n m_1}\right) = \mathrm{O}_\prec\left(n^{-1/2}\right), \quad \mathbf{1}(\Xi)\left(m_3 + \frac{1}{1 + \gamma_n m}\right) = \mathrm{O}_\prec\left(n^{-1/2}\right). \tag{A.64}$$

Finally, plugging (A.59) and (A.60) into (A.64), we conclude (A.52). □

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 A.6 to (A.52) immediately shows that $(m_2(z), m_3(z))$ is equal to $(m_{2c}(z), m_{3c}(z))$ up to an error of order $n^{-1/2}$.

We claim that it suffices to show

$$|m_2(0) - m_{2c}(0)| + |m_3(0) - m_{3c}(0)| \prec n^{-1/2}. \tag{A.65}$$

Once we know (A.65), then by (A.24) and (A.38), we get $\max_{\alpha=2}^{3} |m_{\alpha c}(z) - m_{\alpha c}(0)| = \mathrm{O}((\log n)^{-1})$ and $\max_{\alpha=2}^{3} |m_{\alpha}(z) - m_{\alpha}(0)| = \mathrm{O}((\log n)^{-1})$ with high probability for all $z \in \mathbf{D}$. Together with (A.65), we obtain that

$$\sup_{z\in\mathbf{D}} \left(|m_2(z) - m_{2c}(z)| + |m_3(z) - m_{3c}(z)|\right) \lesssim (\log n)^{-1} \quad \text{with high probability,} \tag{A.66}$$

and

$$\sup_{z\in\mathbf{D}} \left(|m_2(z) - m_{2c}(0)| + |m_3(z) - m_{3c}(0)|\right) \lesssim (\log n)^{-1} \quad \text{with high probability.} \tag{A.67}$$

The condition (A.66) shows that Ξ holds with high probability, and the condition (A.67) verifies the condition (A.23) of Lemma A.6. Then applying Lemma A.6 to (A.52), we obtain that

$$|m_2(z) - m_{2c}(z)| + |m_3(z) - m_{3c}(z)| \prec n^{-1/2} \tag{A.68}$$

for all $z \in \mathbf{D}$. Plugging (A.68) into (A.53)-(A.55), we get the diagonal estimate

$$\max_{\mathfrak{a}\in\mathcal{I}} |G_{\mathfrak{a}\mathfrak{a}}(z) - \Pi_{\mathfrak{a}\mathfrak{a}}(z)| \prec n^{-1/2}. \tag{A.69}$$

Together with the off-diagonal estimate in (A.46), we conclude (A.43). □

Now we give the proof of (A.65).

Proof of (A.65). By (A.17), 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 (A.54), (A.55) and (A.57), we obtain that (A.64) holds at $z = 0$ even without the indicator function $\mathbf{1}(\Xi)$. Furthermore, it gives that

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

with high probability. Then using (A.53) and (A.57), we obtain that (A.59) and (A.60) hold at $z = 0$ even without the indicator function $\mathbf{1}(\Xi)$. Finally, plugging (A.59) and (A.60) into (A.64), we conclude (A.52) holds at $z = 0$, that is,

$$\begin{aligned} \left|\frac{1}{m_2(0)} + 1 - \frac{1}{n}\sum_{i=1}^{p} \frac{\lambda_i^2}{\lambda_i^2 r_1 m_2(0) + r_2 m_3(0)}\right| &\prec n^{-1/2}, \\ \left|\frac{1}{m_3(0)} + 1 - \frac{1}{n}\sum_{i=1}^{p} \frac{1}{\lambda_i^2 r_2 m_2(0) + r_2 m_3(0)}\right| &\prec n^{-1/2}. \end{aligned} \tag{A.70}$$

Denoting $\omega_2 = -m_{2c}(0)$ and $\omega_3 = -m_{2c}(0)$. By (A.64), 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

$$c \leqslant \omega_2 \leqslant 1, \quad c \leqslant \omega_3 \leqslant 1, \quad \text{with high probability.} \tag{A.71}$$

Also one can verify from (A.70) that (ω_2, ω_3) satisfy approximately the same equations as (A.21):

$$r_1\omega_2 + r_2\omega_3 = 1 - \gamma_n + \mathrm{O}_\prec(n^{-1/2}), \quad f(\omega_2) = 1 + \mathrm{O}_\prec(n^{-1/2}). \tag{A.72}$$

The first equation and (A.71) 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 (A.72) gives that $|\omega_2 - x_2| \prec n^{-1/2}$. Together with the first equation in (A.72), we get $|\omega_3 - x_3| \prec n^{-1/2}$. This concludes (A.65). □

With Lemma A.15, we can complete the proof of Proposition A.9.

Proof of Proposition A.9. With (A.43), one can use the polynomialization method in (Bloemendal et al., 2014, Section 5) to get the anisotropic local law (A.28) for G_0 with $q = n^{-1/2}$. The proof is exactly the same, except for some minor differences in notations, so we omit the details. □

A.2.3 Anisotropic Local Law

In this subsection, we finish the proof of Theorem A.7 for a general X satisfying the bounded support condition (A.8) with $q \leqslant n^{-\phi}$ for some constant $\phi > 0$. Proposition A.9 implies that (A.28) holds for Gaussian Z_1^{Gauss} and Z_2^{Gauss} as discussed before. Thus the basic idea is to prove that for Z_1 and Z_2 satisfying the assumptions in Theorem A.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 (Knowles and Yin, 2016). The proof is similar to the ones in Sections 7-8 of (Knowles and Yin, 2016), so we only give a rough description of the basic idea, without writing down all the details.

Definition A.18 (Interpolation)**.** *We denote $Z^0 := Z^{Gauss}$ and $Z^1 := Z$. Let $\rho_{\mu i}^0$ and $\rho_{\mu i}^1$ be the laws of $Z_{\mu i}^0$ and $Z_{\mu i}^1$, respectively. For $\theta \in [0,1]$, we define the interpolated law $\rho_{\mu i}^\theta := (1-\theta)\rho_{\mu i}^0 + \theta\rho_{\mu i}^1$. We shall work on the probability space consisting of triples (Z^0, Z^θ, Z^1) of independent $n \times p$ random matrices, where the matrix $Z^\theta = (Z_{\mu i}^\theta)$ has law*

$$\prod_{i \in \mathcal{I}_1} \prod_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} \rho_{\mu i}^\theta (\mathrm{d} Z_{\mu i}^\theta). \tag{A.73}$$

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

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

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

We shall prove (A.28) through interpolation matrices Z^θ between Z^0 and Z^1. We have see that (A.28) holds for Z^0 by Proposition A.9. Using (A.73) and fundamental calculus, we get the following basic interpolation formula: for $F : \mathbb{R}^{n \times p} \to \mathbb{C}$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}F(Z^\theta) = \sum_{i \in \mathcal{I}_1} \sum_{\mu \in \mathcal{I}_2 \cup \mathcal{I}_3} \left[\mathbb{E}F\left(Z_{(\mu i)}^{\theta, Z_{\mu i}^1}\right) - \mathbb{E}F\left(Z_{(\mu i)}^{\theta, Z_{\mu i}^0}\right)\right] \tag{A.74}$$

provided all the expectations exist. We shall apply (A.74) to $F(Z) := F^s_{\mathbf{u}\,\mathbf{v}}(Z,z)$ for (large) $s \in 2\mathbb{N}$ and $F_{\mathbf{u}\mathbf{v}}(Z,z)$ defined as

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

The main part of the proof is to show the following self-consistent estimate for the right-hand side of (A.74) for any fixed $s \in 2\mathbb{N}$ and constant $\varepsilon > 0$:

$$\sum_{i\in\mathcal{I}_1}\sum_{\mu\in\mathcal{I}_2\cup\mathcal{I}_3}\left[\mathbb{E}F^s_{\mathbf{u}\,\mathbf{v}}\left(Z^{\theta,Z^1_{\mu i}}_{(\mu i)},z\right) - \mathbb{E}F^s_{\mathbf{u}\,\mathbf{v}}\left(Z^{\theta,Z^0_{\mu i}}_{(\mu i)},z\right)\right] = \mathrm{O}\left((n^\varepsilon q)^s + \mathbb{E}F^s_{\mathbf{u}\,\mathbf{v}}\left(Z^\theta,z\right)\right) \tag{A.75}$$

for all $\theta \in [0,1]$. If (A.75) holds, then combining (A.74) with a Grönwall's argument we obtain that for any fixed $s \in 2\mathbb{N}$ 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 (A.28). Underlying the proof of (A.75) is an expansion approach, which is very similar to the ones for Lemma 7.10 of (Knowles and Yin, 2016) and Lemma 6.11 of (Yang, 2019). So we omit the details.

A.2.4 Proofs of the Limiting Equations

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

Proof of Lemma A.5. One can check that the equations in (A.18) are equivalent to the following ones:

$$r_1 m_{2c} = -(1-\gamma_n) - r_2 m_{3c} - z\left(m_{3c}^{-1}+1\right), \quad g_z(m_{3c}(z)) = 1, \tag{A.76}$$

where

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

We first show that there exists a unique solution $m_{3c}(z)$ to the equation $g_z(m_{3c}(z)) = 1$ under the conditions in (A.23), and the solution satisfies (A.24). Now we abbreviate $\varepsilon(z) := m_{3c}(z) - m_{3c}(0)$, and from (A.76) we 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

$$\varepsilon(z) = -\frac{g_z(m_{3c}(0)) - g_0(m_{3c}(0))}{g_z'(m_{3c}(0))} - \frac{g_z(m_{3c}(0)+\varepsilon(z)) - g_z(m_{3c}(0)) - g_z'(m_{3c}(0))\varepsilon(z)}{g_z'(m_{3c}(0))}.$$

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

$$\varepsilon^{(k+1)} = -\frac{g_z(m_{3c}(0)) - g_0(m_{3c}(0))}{g_z'(m_{3c}(0))} - \frac{g_z(m_{3c}(0)+\varepsilon^{(k)}) - g_z(m_{3c}(0)) - g_z'(m_{3c}(0))\varepsilon^{(k)}}{g_z'(m_{3c}(0))}. \tag{A.77}$$

Then (A.77) 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

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

Then it is easy to check that there exist constants $\widetilde{c}, \widetilde{C} > 0$ depending only on τ in (A.7) and (A.22) such that

$$\left|[g_z'(m_{3c}(0))]^{-1}\right| \leqslant \widetilde{C}, \quad \left|\frac{g_z(m_{3c}(0)) - g_0(m_{3c}(0))}{g_z'(m_{3c}(0))}\right| \leqslant \widetilde{C}|z|, \tag{A.78}$$

and

$$\left|\frac{g_z(m_{3c}(0)+\varepsilon_1)-g_z(m_{3c}(0)+\varepsilon_2)-g_z'(m_{3c}(0))(\varepsilon_1-\varepsilon_2)}{g_z'(m_{3c}(0))}\right| \leqslant \widetilde{C}|\varepsilon_1-\varepsilon_2|^2, \tag{A.79}$$

for all $|z|\leqslant \widetilde{c}$ and $|\varepsilon_1|\leqslant \widetilde{c}$, $|\varepsilon_2|\leqslant \widetilde{c}$. Then with (A.78) and (A.79), 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 (A.76) subject to the condition $\|\varepsilon\|_\infty\leqslant r$.

From the iteration relation (A.77), using (A.78) one can readily check that

$$\varepsilon^{(k+1)}-\varepsilon^{(k)} = h(\varepsilon^{(k)})-h(\varepsilon^{(k-1)}) \leqslant \widetilde{C}|\varepsilon^{(k)}-\varepsilon^{(k-1)}|^2. \tag{A.80}$$

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 (A.23) as $c_0=\min\{c_\delta, r\}$. After obtaining $m_{3c}(z)$, we can then find $m_{2c}(z)$ using the first equation in (A.76).

Note that with (A.79) and $\varepsilon^{(0)}=0$, we get from (A.77) that $|\varepsilon^{(1)}|\leqslant\widetilde{C}|z|$. With the contraction mapping, we have the bound

$$|\varepsilon| \leqslant \sum_{k=0}^{\infty}|\varepsilon^{(k+1)}-\varepsilon^{(k)}| \leqslant 2\widetilde{C}|z|. \tag{A.81}$$

This gives the bound (A.24) for $m_{3c}(z)$. Using the first equation in (A.76), we immediately obtain the bound $r_1|m_{2c}(z)-m_{2c}(0)|\leqslant C|z|$. This gives (A.24) 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 (A.18) and treat $m_{2c}(z)$ as the solution to the following equation:

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

Then with similar arguments as above between (A.76) and (A.81), we can conclude (A.24) for $m_{2c}(z)$. This concludes the proof of Lemma A.5. □

Proof of Lemma A.6. Under (A.23), we can obtain equation (A.76) approximately up to some small error

$$r_1 m_{2c} = -(1-\gamma_n) - r_2 m_{3c} - z\left(m_{3c}^{-1}+1\right) + \mathcal{E}_2'(z), \quad g_z(m_{3c}(z)) = 1+\mathcal{E}_3'(z), \tag{A.82}$$

with $|\mathcal{E}_2'(z)|+|\mathcal{E}_3'(z)| = \mathrm{O}(\delta(z))$. Then we subtract the equations (A.76) from (A.82), and consider the contraction principle for the functions $\varepsilon(z):=m_3(z)-m_{3c}(z)$. The rest of the proof is exactly the same as the one for Lemma A.5, so we omit the details. □

B Proofs for Isotropic and Covariate Shifted Settings

We show that for the isotropic model, the optimum of $\hat{v}$ is very close to 1. This allows us to simplify the expression of $L(\hat{\beta}_2^{\mathrm{MTL}})$ significantly. More precisely, we have the following lemma. Later on, we combine the following lemma with Lemma 6.6 together to analyze the bias-variance tradeoff.

Lemma B.1. *In the isotropic model, for some constant $c_0>0$, suppose that*

$$\frac{pd^2}{\sigma^2}\leqslant \mathrm{O}(1), \quad p^{-1+c_0}\leqslant\frac{\kappa^2}{\sigma^2}\leqslant p^{-\varepsilon_0-c_0}. \tag{B.1}$$

Then we have

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

w.h.p. for some constant $\varepsilon>0$.

Proof. Using Lemma A.14, we obtain that $val(v) = N_2 \cdot h(v)\left(1 + \mathrm{O}(p^{-1/2+\varepsilon})\right)$. We define the function

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

Hence the validation loss in (2.9) reduces to

$$g(v) = \left[N_2 \cdot h(v) + (N_1\sigma_1^2 + N_2\sigma_2^2)\right] \cdot \left(1 + \mathrm{O}(p^{-(1-\varepsilon_0)/2+\varepsilon})\right) \tag{B.3}$$

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

Let $\hat{w}$ the minimizer of $h(v)$. The proof consists of two steps.

- First, we show that $\hat{w}$ is close to 1.
- Second, with (B.3) we show that $\hat{v}$ is close to $\hat{w}$. Then we plug $\hat{v}$ into $L(\hat{\beta}_2^{\mathrm{MTL}})$ to get (B.2).

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

Claim B.2. *There exists a constant $C > 0$ such that*

$$|\hat{w} - 1| \leqslant C\left(\frac{d^2}{\kappa^2} + \frac{\sigma^2}{p\kappa^2}\right) \quad w.h.p. \tag{B.4}$$

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 $h(w) < h(-w)$ for $w > 0$. Hence it suffices to assume that $w \geqslant 0$. We first consider the case $w \geqslant 1$. We write

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

Notice that

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

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

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

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 (6.14), we get that $\mathcal{A}$ is lower bounded as

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

as long as

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

Hence $h'(w) > 0$ on (w_1, ∞), i.e. $h(w)$ is strictly increasing for $w > w_1$. This gives $\hat{w} \leqslant w_1$.

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

$$\begin{aligned} h'(w) \leqslant & \frac{\rho_1}{\rho_2}\left[2\left(w-1\right)\kappa^2\right] \cdot \mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-2}(X_2^\top X_2)^2\right] \\ & + \left[2wd^2 + 2w\left(w-1\right)\left(2w-1\right)\kappa^2\right] \cdot \mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-2}(X_1^\top X_1)^2\right] \\ & + \frac{\rho_1}{\rho_2}(2w\sigma^2) \cdot \mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-2}X_2^\top X_2\right] \\ = & \frac{\rho_1}{\rho_2}\,\mathrm{Tr}\left[(w^2X_1^\top X_1 + X_2^\top X_2)^{-1}\mathcal{B}\right], \end{aligned} \tag{B.5}$$

where the matrix $\mathcal{B}$ is

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

Using the estimate (6.14), 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 $h'(w) < 0$ on $[0, w_2)$, i.e. $h(w)$ is strictly decreasing for $w < w_2$. This gives $\hat{w} \geqslant w_2$.

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. □

Next we prove the following estimate on the optimizer $\hat{v}$: with high probability,

$$|\hat{v} - 1| = \mathrm{O}\left(\mathcal{E}\right), \quad \mathcal{E} := \frac{d^2}{\kappa^2} + \frac{\sigma^2}{p\kappa^2} + p^{-1/2+\varepsilon_0/2+2\varepsilon}. \tag{B.6}$$

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

$$h(w) = \mathrm{O}(pd^2), \quad \text{for} \quad |w - 1| \leqslant 2C\mathcal{E}.$$

Thus we obtain that for $|w - 1| \geqslant 2C\mathcal{E}$,

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

which leads to $g(w) > g(\hat{w})$ w.h.p. by (B.3). Thus w cannot be a minimizer of $g(v)$, and we must have $|\hat{v} - 1| \leqslant 2C\mathcal{E}$.

Inserting (B.6) into (2.10) and applying Lemma A.14 to $(\beta_1 - \hat{v}\beta_s)$ again, we get w.h.p.,

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

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 (6.14), we see that

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

Hence we need to choose that $pd^2 \sim \sigma^2$. 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.1). Moreover, under (B.1) we can simplify (B.7) to (B.2). □

B.1 Missing Proofs regarding Task Similarity

With (B.2) and Lemma 6.6, we can prove Proposition 4.1, which gives a transition threshold with respect to the ratio between the model bias and the noise level. With slight abuse of notations, we shall write $\hat{a}_i$ and $\hat{b}_k$ as a_i and b_k throughout the proof.

Proof of Proposition 4.1. In the setting of Proposition 4.1, we have $M = \Sigma_1^{1/2}\Sigma_2^{-1/2} = \mathrm{Id}$. Then solving equations (6.3) and (6.5) with $\lambda_i = 1$, 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 6.1 and Lemma 6.2, we can track the reduction of variance from $\hat{\beta}_2^{\mathrm{MTL}}$ to $\hat{\beta}_2^{\mathrm{STL}}$ as

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

with high probability, where

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

Next for the model shift bias

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

we can get from Lemma 6.6 (or rather its proof) that

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

where

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

Note that

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

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 $\delta_{\mathrm{var}} > \delta_{\mathrm{bias}}$, which implies $L(\hat{\beta}_2^{\mathrm{MTL}}) < L(\hat{\beta}_2^{\mathrm{STL}})$. 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 $\Psi(\beta_1,\beta_2) = pd^2/\sigma^2$ and $\nu \geqslant \alpha_+^2(\rho_1)+\mathrm{o}(1)$, it gives the first statement of Proposition 4.1.

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 $\delta_{\mathrm{var}} < \delta_{\mathrm{bias}}$, which implies $L(\hat{\beta}_2^{\mathrm{MTL}}) > L(\hat{\beta}_2^{\mathrm{STL}})$. 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 4.1. □

Next we consider the case where the two tasks have different noise variances $\sigma_1^2 \neq \sigma_2^2$. In particular, 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. *In the isotropic model, assume that* $\rho_1 > 40$ *and* $\mathbb{E}\left[\|\beta_1 - \beta_2\|^2\right] < \frac{1}{2}\sigma_2^2 \cdot \Phi(\rho_1, \rho_2)$. *Then we have the following transition with respect to* σ_1^2*:*

- *If* $\sigma_1^2 < -\nu^{1/2}\rho_1 \cdot pd^2 + \left(1 + \nu^{-1/2}\rho_1 \Phi(\rho_1, \rho_2)\right) \cdot \sigma_2^2$, *then w.h.p.* $L(\hat{\beta}_2^{MTL}) < L(\hat{\beta}_2^{STL})$.
- *If* $\sigma_1^2 > -\nu^{-1/2}\rho_1 \cdot pd^2 + \left(1 + \nu^{1/2}\rho_1 \Phi(\rho_1, \rho_2)\right) \cdot \sigma_2^2$, *then w.h.p.* $L(\hat{\beta}_2^{MTL}) > L(\hat{\beta}_2^{STL})$.

As a corollary, if $\sigma_1^2 \leqslant \sigma_2^2$, then we always get positive transfer.

Proof. In the setting of Proposition B.3, the test loss is given by (2.10). In the isotropic model, using again the concentration result, Lemma A.14, we can rewrite $L(\hat{\beta}_2^{\mathrm{MTL}})$ as

$$\begin{aligned} L(\hat{\beta}_2^{\mathrm{MTL}}) = \ & \hat{v}^2 \left[d^2 + (\hat{v}-1)^2 \kappa^2\right] \operatorname{Tr}\left[(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-2} (X_1^\top X_1)^2\right] \cdot \left(1 + \mathrm{O}(p^{-1/2+\varepsilon})\right) \\ & + \sigma_2^2 \cdot \operatorname{Tr}\left[(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1}\right] + (\sigma_1^2 - \sigma_2^2) \cdot \operatorname{Tr}\left[(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-2} \hat{v}^2 X_1^\top X_1\right] \end{aligned}$$

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

In the current setting, we can also show that (B.6) holds for $\hat{v}$. Since the proof is almost the same as the one for Lemma B.2, we omit the details. Thus under the choice of parameters in (B.1), $L(\hat{\beta}_2^{\mathrm{MTL}})$ can be simplified as in (B.2):

$$\begin{aligned} L(\hat{\beta}_2^{\mathrm{MTL}}) &= (1 + \mathrm{O}(n^{-\varepsilon})) \cdot d^2 \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2} (X_1^\top X_1)^2\right] \\ &\quad + (1 + \mathrm{O}(n^{-\varepsilon})) \cdot \sigma_2^2 \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-1}\right] \\ &\quad + (1 + \mathrm{O}(n^{-\varepsilon})) \cdot (\sigma_1^2 - \sigma_2^2) \operatorname{Tr}\left[(X_1^\top X_1 + X_2^\top X_2)^{-2} X_1^\top X_1\right]. \end{aligned} \tag{B.17}$$

Then we have

$$L(\hat{\beta}_2^{\mathrm{STL}}) - L(\hat{\beta}_2^{\mathrm{MTL}}) = \delta_{\mathrm{var}} - \delta_{\mathrm{bias}} - \delta_{\mathrm{var}}^{(2)},$$

where

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

satisfies (B.10) with σ^2 replaced by σ_2^2,

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

satisfies (B.11), and $\delta_{\mathrm{var}}^{(2)}$ is defined as

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

To estimate this new term $\delta_{\mathrm{var}}^{(2)}$, we use the same arguments as in the proof of Lemma 6.6: we first replace $X_1^\top X_1$ with $n_1 \operatorname{Id}$ up to some error using (6.14), and then apply Lemma 6.3 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 $\delta_{\mathrm{var}}^{(2)}$:

$$\alpha_-(\rho_1) - \mathrm{o}(1) \leqslant \delta_{\mathrm{var}}^{(2)} / \Delta_{\mathrm{var}}^{(2)} \leqslant \alpha_+(\rho_1) + \mathrm{o}(1), \tag{B.18}$$

where

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

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

Note that the condition $\mathbb{E}\left[\|\beta_1 - \beta_2\|^2\right] < \frac{1}{2}\sigma_2^2 \cdot \Phi(\rho_1, \rho_2)$ for $\rho_1 > 40$ gives that $\delta_{\mathrm{var}} > \delta_{\mathrm{bias}}$ by Proposition 4.1. Hence if $\sigma_1^2 \leqslant \sigma_2^2$, then $\delta_{\mathrm{var}}^{(2)} < 0$ and we always have $\delta_{\mathrm{var}} > \delta_{\mathrm{bias}} + \delta_{\mathrm{var}}^{(2)}$, which gives $L(\hat{\beta}_2^{\mathrm{MTL}}) < L(\hat{\beta}_2^{\mathrm{STL}})$. It remains to consider the case $\sigma_1^2 \geqslant \sigma_2^2$.

Positive transfer. By (B.10), (B.11) and (B.18), if the following inequality holds,

$$\begin{aligned}&\sigma_2^2 \cdot \frac{\rho_1}{(\rho_2-1)(\rho_1+\rho_2-1)} \cdot (1-\mathrm{o}(1)) \\ &> pd^2 \cdot \frac{\rho_1^2(\rho_1+\rho_2)}{(\rho_1+\rho_2-1)^3}\alpha_+^2(\rho_1) + (\sigma_1^2-\sigma_2^2)\cdot \frac{\rho_1(\rho_1+\rho_2)}{(\rho_1+\rho_2-1)^3}\alpha_+(\rho_1),\end{aligned} \tag{B.19}$$

then we have $\delta_{\mathrm{var}} > \delta_{\mathrm{bias}} + \delta_{\mathrm{var}}^{(2)}$ w.h.p., which gives $L(\hat{\beta}_t^{\mathrm{MTL}}) < L(\hat{\beta}_t^{\mathrm{STL}})$. We can solve (B.19) to get

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

This proves the first claim of Proposition B.3 using $\nu \geqslant \alpha_+^2(\rho_1) + \mathrm{o}(1)$.
Negative transfer. On the other hand, if the following inequality holds,

$$\begin{aligned}&\sigma_2^2 \cdot \frac{\rho_1}{(\rho_2-1)(\rho_1+\rho_2-1)} \cdot (1+\mathrm{o}(1)) \\ &< pd^2 \cdot \frac{\rho_1^2(\rho_1+\rho_2)}{(\rho_1+\rho_2-1)^3}\alpha_-^2(\rho_1) + (\sigma_1^2-\sigma_2^2)\cdot \frac{\rho_1(\rho_1+\rho_2)}{(\rho_1+\rho_2-1)^3}\alpha_-(\rho_1),\end{aligned} \tag{B.20}$$

then we have $\delta_{\mathrm{var}} < \delta_{\mathrm{bias}} + \delta_{\mathrm{var}}^{(2)}$ w.h.p., which gives $L(\hat{\beta}_t^{\mathrm{MTL}}) > L(\hat{\beta}_t^{\mathrm{STL}})$. We can solve (B.20) to get

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

This proves the second claim of Proposition B.3. □

B.2 Missing Proofs regarding Sample Size

We first prove Proposition 4.2, which describes the effect of source/task data ratio on the information transfer.

Proof of Proposition 4.2. Following the proof of Proposition 4.1, we see that $L(\hat{\beta}_2^{\mathrm{MTL}}) < L(\hat{\beta}_2^{\mathrm{STL}})$ w.h.p. if (B.14) holds, while $L(\hat{\beta}_2^{\mathrm{MTL}}) > L(\hat{\beta}_2^{\mathrm{STL}})$ w.h.p. if (B.16) holds.

We first explain the meaning of the condition

$$\Psi(\beta_1,\beta_2) > 2/(\rho_2-1). \tag{B.21}$$

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.21) 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 4.2.

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

$$\frac{(\rho_2-1)\rho_1}{\rho_1+\rho_2-2}\cdot\left(1-\frac{1}{(\rho_1+\rho_2-2)^2}\right) \leqslant \Phi(\rho_1,\rho_2) \leqslant \frac{(\rho_2-1)\rho_1}{\rho_1+\rho_2-2}. \tag{B.22}$$

Positive transfer. With (B.22), we see that (B.14) is implied by the following inequality:

$$\Psi(\beta_1,\beta_2)\cdot\frac{(\rho_2-1)\rho_1}{\rho_1+\rho_2-2} < \left(\alpha_+^2(\rho_1)+\mathrm{o}(1)\right)^{-1}. \tag{B.23}$$

Then we can solve (B.23) to get

$$\rho_1 < \frac{\rho_2-2}{\Psi(\beta_1,\beta_2)\cdot(\rho_2-1)\left(\alpha_+^2(\rho_1)+\mathrm{o}(1)\right)-1}. \tag{B.24}$$

This gives the first statement of Proposition 4.2 using $\nu \geqslant \alpha_+^2(\rho_1) + \mathrm{o}(1)$.

Note that if we require the RHS of (B.24) to be larger than 40, that is, (B.24) is not a null condition. Then together with (B.21), we get

$$\rho_2 - 2 > \left[2\left(\alpha_+^2(\rho_1) + \mathrm{o}(1)\right) - 1\right]\rho_1.$$

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

Negative transfer. With (B.22), we see that (B.16) is implied by the following inequality:

$$\Psi(\beta_1, \beta_2)\frac{(\rho_2 - 1)\rho_1}{\rho_1 + \rho_2 - 2}\left(1 - \frac{1}{(\rho_1 + \rho_2 - 2)^2}\right) > \Psi(\beta_1, \beta_2)\frac{(\rho_2 - 1.5)\rho_1}{\rho_1 + \rho_2 - 2} > \left(\alpha_-^2(\rho_1) - \mathrm{o}(1)\right)^{-1}. \tag{B.25}$$

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.25) to get

$$\rho_1 > \frac{(\rho_2 - 2)\sigma^2}{\Psi(\beta_1, \beta_2) \cdot (\rho_2 - 1.5)\left(\alpha_-^2(\rho_1) - \mathrm{o}(1)\right) - 1}, \tag{B.26}$$

which gives the second statement of Proposition 4.2. We remark that condition (B.21) implies $\Psi(\beta_1, \beta_2) \cdot (\rho_2 - 1.5)\left(\alpha_-^2(\rho_1) - \mathrm{o}(1)\right) > 1$, so (B.26) does not give a trivial bound. □

Next we state Proposition B.4, which gives precise upper and lower bounds on the data efficiency ratio for Taskonomy.

Proposition B.4. *In the isotropic model, assume that $\rho_1, \rho_2 \geqslant 9$ and $\Psi(\beta_1, \beta_2) < (5(\rho_1 - 1))^{-1} + (5(\rho_2 - 1))^{-1}$. Then the data efficiency ratio $x^\star$ satisfies*

$$x_l \leqslant x^\star \leqslant \frac{1}{\rho_1 + \rho_2}\left(\frac{2}{(\rho_1 - 1)^{-1} + (\rho_2 - 1)^{-1} - 5\Psi(\beta_1, \beta_2)} + 1\right), \tag{B.27}$$

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. 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 4.1 still hold, except that we need to replace (ρ_1, ρ_2) with $(x\rho_1, x\rho_2)$. More precisely, we have

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

Moreover, with high probability,

$$L_i(\hat{\beta}_i^{\mathrm{MTL}}(x)) = \frac{\sigma^2}{x(\rho_1 + \rho_2) - 1}\left(1 + \mathrm{o}(1)\right) + \delta_{\mathrm{bias}}^{(i)}, \quad i = 1, 2. \tag{B.28}$$

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 6.1 we have w.h.p.,

$$L_i(\hat{\beta}_i^{\mathrm{STL}}) = \frac{\sigma^2}{\rho_i - 1}\left(1 + \mathrm{o}(1)\right), \quad i = 1, 2. \tag{B.29}$$

Comparing (B.28) and (B.29), we immediately obtain the lower bound $x^\star \geqslant x_l$. 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, $L_1(\hat{\beta}_1^{\mathrm{MTL}}(x)) + L_2(\hat{\beta}_2^{\mathrm{MTL}}(x))$ 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 $x^\star$. In the following discussions, we only consider x such that $x > x_l$. In particular, we have $x\rho > x_l\rho \geqslant \min(\rho_1,\rho_2)$, where we abbreviated $\rho := \rho_1+\rho_2$.

The upper bound. From (B.28) and (B.29), we see that $x^\star \leqslant x$ if x satisfies

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

We rewrite the inequality as

$$(1+\mathrm{o}(1))\cdot\frac{\Psi(\beta_1,\beta_2)}{[1-(x\rho)^{-1}]^3}\sum_{i=1}^{2}\left(\sqrt{\frac{\rho_i}{\rho}}+\sqrt{\frac{1}{x\rho}}\right)^4 \leqslant \frac{1}{\rho_1-1}+\frac{1}{\rho_2-1}-\frac{2}{x\rho-1}. \tag{B.30}$$

With $x\rho \geqslant \min(\rho_1,\rho_2)\geqslant 9$, we can get the simple bound

$$\frac{1+\mathrm{o}(1)}{[1-(x\rho)^{-1}]^3}\sum_{i=1}^{2}\left(\sqrt{\frac{\rho_i}{\rho}}+\sqrt{\frac{1}{x\rho}}\right)^4 < 5.$$

Inserting it into (B.30), we can solve for the upper bound in (B.27).

We can get better bounds if the values of ρ_1 and ρ_2 increase. For example, if we consider the case $\min(\rho_1,\rho_2)\geqslant 100$, then with some basic calculations, one can show that in this case

$$\frac{1}{[1-(x\rho)^{-1}]^3}\sum_{i=1}^{2}\left(\sqrt{\frac{\rho_i}{\rho}}+\sqrt{\frac{1}{x\rho}}\right)^4 < \frac{\rho_1^2+\rho_2^2}{\rho^2}+0.52.$$

Thus the following inequality implies (B.30):

$$\left(\frac{\rho_1^2+\rho_2^2}{\rho^2}+0.52\right)\Psi(\beta_1,\beta_2) < \frac{1}{\rho_1-1}+\frac{1}{\rho_2-1}-\frac{2}{x\rho-1},$$

from which we can solve for the following upper bound on $x^\star$:

$$x^\star < \frac{1}{\rho}\frac{2}{\frac{1}{\rho_1-1}+\frac{1}{\rho_2-1}-\left(\frac{\rho_1^2+\rho_2^2}{\rho^2}+0.52\right)\Psi(\beta_1,\beta_2)}+\frac{1}{\rho}.$$

Similarly, we can get a better lower bound. From (B.28) and (B.29), we see that $x^\star \geqslant x$ if x satisfies

$$(1-\mathrm{o}(1))\cdot\frac{\Psi(\beta_1,\beta_2)}{[1-(x\rho)^{-1}]^3}\sum_{i=1}^{2}\left(\sqrt{\frac{\rho_i}{\rho}}-\sqrt{\frac{1}{x\rho}}\right)^4 \geqslant \frac{1}{\rho_1-1}+\frac{1}{\rho_2-1}-\frac{2}{x\rho-1}. \tag{B.31}$$

Then in the case $\min(\rho_1,\rho_2)\geqslant 100$, with some basic calculations, one can show that the sum on the left-hand side of (B.31) satisfies

$$\frac{1}{[1-(x\rho)^{-1}]^3}\sum_{i=1}^{2}\left(\sqrt{\frac{\rho_i}{\rho}}-\sqrt{\frac{1}{x\rho}}\right)^4 > \frac{\rho_1^2+\rho_2^2}{\rho^2}-0.33.$$

Thus the following inequality implies (B.31):

$$\left(\frac{\rho_1^2+\rho_2^2}{\rho^2}-0.33\right)pd^2 > \frac{\sigma^2}{\rho_1-1}+\frac{\sigma^2}{\rho_2-1}-\frac{2\sigma^2}{x\rho-1}, \tag{B.32}$$

from which we can solve for the following lower bound on $x^\star$:

$$x^\star > \frac{1}{\rho}\frac{2}{\frac{1}{\rho_1-1}+\frac{1}{\rho_2-1}-\left(\frac{\rho_1^2+\rho_2^2}{\rho^2}-0.33\right)\Psi(\beta_1,\beta_2)}+\frac{1}{\rho}.$$

This gives a lower bound above x_l. □

B.3 Missing Proofs regarding Covariate Shift

We now prove Proposition 4.3, which shows that $L(\hat{\beta}^{\mathrm{MTL}})$ is minimized approximately when M is a scalar matrix where there is enough source data.

Proof of Proposition 4.3. Let

$$M_0 := \operatorname*{arg\,min}_{M\in\mathcal{S}_\mu} g(M).$$

We now calculate $g(M_0)$. With the same arguments as in Lemma B.2, we can show that (B.6) holds. Moreover, if the parameters are chosen such that $p^{-1+c_0}\sigma^2 \leqslant \kappa^2 \leqslant p^{-\varepsilon_0-c_0}\sigma^2$ as in (B.1), 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 some constant $\varepsilon>0$. In fact, Lemma B.2 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 $\mu_{\min}\leqslant\lambda_p(M)\leqslant\lambda_1(M)\leqslant\mu_{\max}$. We omit the details here.

Now using Lemma 6.2, we obtain that with high probability,

$$g(M_0) = \frac{\sigma^2}{\rho_1+\rho_2}\cdot\frac{1}{p}\operatorname{Tr}\left(\frac{1}{a_1(M_0)\cdot M_0^\top M_0+a_2(M_0)}\right)\cdot\left(1+\mathrm{O}(p^{-\varepsilon})\right). \tag{B.33}$$

From equation (6.3), it is easy to obtain the following estimates on $a_1(M)$ and $a_2(M)$ for any $M\in\mathcal{S}_\mu$:

$$\frac{\rho_1-1}{\rho_1+\rho_2} < a_1(M) < \frac{\rho_1+\rho_2-1}{\rho_1+\rho_2},\quad a_2(M) < \frac{\rho_2}{\rho_1+\rho_2}. \tag{B.34}$$

Inserting (B.34) into (B.33) and using $M_0^\top M_0 \succeq \mu_{\min}^2$, we obtain that with high probability,

$$\left(1+\frac{\rho_2}{(\rho_1-1)\mu_{\min}^2}\right)^{-1} h(M_0)\cdot\left(1-\mathrm{O}(p^{-\varepsilon})\right)\leqslant g(M_0)\leqslant h(M_0)\cdot\left(1+\mathrm{O}(p^{-\varepsilon})\right), \tag{B.35}$$

where

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

By AM-GM inequality, we observe that

$$\operatorname{Tr}\left(\frac{1}{M^\top M}\right) = \sum_{i=1}^p\frac{1}{\lambda_i^2}$$

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

$$h(M_0)\leqslant\frac{\sigma^2}{\mu^2(\rho_1+\rho_2)a_1(M_0)}. \tag{B.36}$$

On the other hand, when $M=\mu\,\mathrm{Id}$, applying Lemma 6.2 we obtain that with high probability,

$$\begin{aligned} g(\mu\,\mathrm{Id}) &= \frac{\sigma^2}{\rho_1+\rho_2}\cdot\frac{1}{p}\operatorname{Tr}\left(\frac{1}{\mu^2a_1(\mu\,\mathrm{Id})+a_2(\mu\,\mathrm{Id})}\right)\cdot\left(1+\mathrm{O}(p^{-\varepsilon})\right)\\ &\leqslant\frac{\sigma^2}{\mu^2(\rho_1+\rho_2)a_1(\mu\,\mathrm{Id})}.\end{aligned} \tag{B.37}$$

Combining (B.34), (B.35), (B.36) and (B.37), we conclude the proof. □