Generalization Effects of Transferring Across High-dimensional Data in Multi-Task Learning

12/09/2020 at 11:18pm

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.

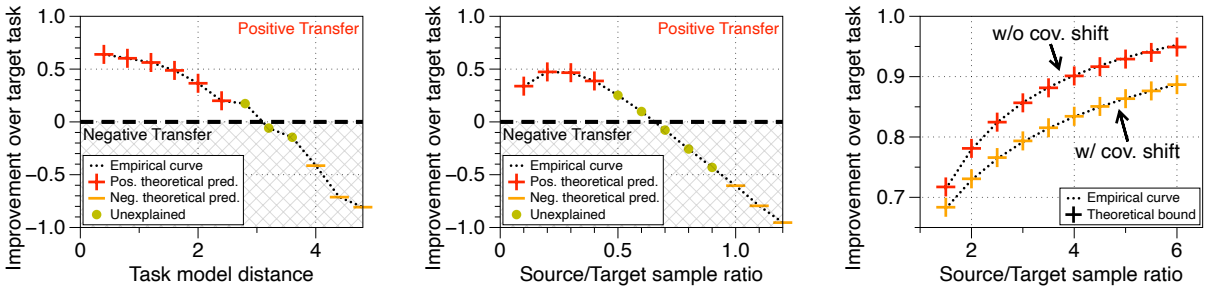
1 Introduction

Multi-task learning is a powerful approach to improve performance for many tasks in computer vision **??**, natural language processing **??**, and other areas **?**. 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 **?**. 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) **??**, which is referred to as *negative transfer* **?**. While many empirical approaches have been proposed to mitigate negative transfer **?**, a precise understanding of when negative transfer occurs remains elusive in the literature **?**.

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 **?????**. 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 **?**.

In this work, we study the bias and variance of multi-task learning in the high-dimensional linear regression setting **??**. 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 **??** 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 **?**. The high-dimensional linear regression setting allows us to measure the three properties precisely. We define each property for the 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 $\|\beta_1 - \beta_2\|$.



(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.

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 **?**. This two-layer model is inspired by a commonly used idea of hard parameter sharing in multi-task learning **??**. 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 **???**. 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 **??** and extend the analysis to many tasks with the same features in Theorem **??**.

Second, we explain the phenomena in Figure **??** 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 **??**). 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 **??**). 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 **??**). 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. **?**.

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 **?** and sentiment analysis datasets **?**. 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 **??** 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 **?**. 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.

2 Problem Formulation for Multi-Task Learning

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*.

Problem setup. 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. **??**), 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 $x_1, x_2, \ldots, x_{n_k}$. 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 **?**. 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 $W_1, \ldots, W_t$ 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 **?**, 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. **?**).

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 $W_1, \ldots, W_t$, leading to a closed form equation for $\hat{B}$ that depends on $W_1, \ldots, W_k$.

$$f(B; W_1, \ldots, W_t) = \sum_{k=1}^{t} \|X_k^{tr} B W_k - Y_k^{tr}\|^2. \tag{2.1}$$

Tune the output layers W_i: set $B = \hat{B}$ and minimize the validation loss over $W_1, \ldots, W_k$.

$$g(W_1, \ldots, W_t) = \sum_{k=1}^{t} \|X_k^{val} \hat{B} W_k - Y_k^{val}\|^2. \tag{2.2}$$

We make several remarks. In general, the objective $g(\cdot)$ is non-convex in B and the W_k's. Therefore, we first minimize B in equation (**??**) and then minimize W_k's in equation (**??**). For W_k's, 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.

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 **??**).

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

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

In the above, $\hat{v} = W_1/W_2$ where W_1, W_2 are obtained from solving equation (**??**) (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 (**??**) 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 (**??**) 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.

3 Comparing Multi-Task Learning to Single-Task Learning

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.

3.1 Analyzing the Bias-Variance Tradeoff using Random Matrix Theory

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 **?**), 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.

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 $\lambda_1, \ldots, \lambda_p$ 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 **??** allows us to get a tight bound on equation (**??**), 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 (**??**), a similar result that scales with task model distance in addition to sample size and covariate shift holds (cf. Lemma **??** in Appendix **??**). 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.

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 **??** 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 $\Delta_{\mathrm{bias}} - \Delta_{\mathrm{var}}$. 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 **??** is presented in Appendix **??** and the proof of Lemma **??** is in Appendix **??**.

3.2 Task Similarity

It is well-known since the seminal work of Caruana **?** 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 3.3 (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 **??** simplifies Theorem **??** 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 **??** to the parameter setting of Figure **??** (the details are left to Appendix **??**). 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 **??** can be found in Appendix **??**. A key part of the analysis shows that $\hat{v} \approx 1$ in the isotropic model, thus simplifying the result of Theorem **??**.

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 **??** in Appendix **??**. Inspired by the observation, we propose a single-task based metric to help understand MTL results using STL results.

For each task, we train a single-task model. Let z_s and z_t be the prediction accuracy of each task, respectively. Let $\tau \in (0, 1)$ be a fixed threshold.

If $z_s - z_t > \tau$, then we predict that there will be positive transfer when combining the two tasks using MTL. If $z_s - z_t < -\tau$, then we predict negative transfer.

3.3 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 **???**. On the other hand, uneven sample sizes between different tasks (or even dominating tasks) have been empirically observed as a cause of negative transfer **?**. 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 3.4 (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 **??** describes the bias-variance tradeoff in terms of the sample ratio ρ_1/ρ_2. We apply the result to the setting of Figure **??** (described in Appendix **??**). There are several unexplained observations near $y = 0$ caused by ν. The proof of Proposition **??** can be found in Appendix **??**.

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

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

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 **??**, 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 **??** in Appendix **??**.

Algorithmic consequence. An interesting consequence of Proposition **??** is that $L(\hat{\beta}_t^{\mathrm{MTL}})$ is not monotone in ρ_1. In particular, Figure **??** (and our analysis) shows that $L(\hat{\beta}_t^{\mathrm{MTL}})$ behaves as a quadratic function over ρ_1. More generally, depending on how large $\Psi(\beta_1, \beta_2)$ is, $L(\hat{\beta}_t^{\mathrm{MTL}})$ 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 **??** in Appendix **??** describes the procedure in detail.

3.4 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 **??** 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 3.5 (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 **??** is left to Appendix **??**.

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 **?**, 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.

3.5 Extensions

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 **??**.

Theorem 3.6 (Many tasks)**.** *For the setting of* t *tasks where* $X_i = X$*, for all* $1 \leqslant i \leqslant t$*, let* $B^\star := [\beta_1, \beta_2, \ldots, \beta_t]$ *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 **??** 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 **??** can be found in Appendix **??**.

4 Related Work

We refer the interested readers to several excellent surveys on multi-task learning for a comprehensive survey **????**. 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 **?**, Ben-David and Schuller **?**. Mauer **?** studies generalization bounds for linear separation settings of MTL. Ben-David et al. **?** 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 **?** and sparse regression **??**. Our work is closely related to Wu et al. **?**. 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 **?** introduces an alternating minimization framework for learning multiple tasks. Argyriou et al. **?** present a convex algorithm which learns common sparse representations across a pool of related tasks. Evgeniou et al. **?** 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 **???**. 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 **?**. For a sample covariance matrix $X^\top X$ with $\Sigma = \mathrm{Id}$, such a local law was proved in **?**. It was later extended to sample covariance matrices with non-identity Σ **?**, and separable covariance matrices **?**. On the other hand, one may derive the asymptotic result in Theorem **??**

with error o(1) using the free addition of two independent random matrices in free probability theory **?**. 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.

5 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 **??**) involves an error term that scales down with ρ_1. Tightening this error bound might cover the unexplained observations in Figure **??**. 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.

Broader Impacts

In this work, we provide a theoretical framework to help understand when multi-task learning performs well. We approach this question by studying the bias-variance tradeoff of multi-task learning. We provide new technical tools to analyze the bias and variance of multi-task learning.

Our theoretical framework has the potential to impact many other neighboring areas in the ML community. (i) Our concentration bounds can apply to different settings such as soft parameter sharing **?** , kernel methods **?**, and convex formulation of multi-task learning **?**. (ii) Our analysis of the bias-variance tradeoff can extend to transfer learning and domain adaptation **?**. (iii) Our insights on positive and negative transfer can be useful in multimodal learning, where the data sources are usually heterogeneous. (iv) Our fine-gained study on sample sizes have the potential to provide new insight in meta learning, where limited labeled samples presents a significant challenge. Finally, since multi-task learning connects to a wide range of areas **?** such as semi-supervised learning, representation learning, and reinforcement learning **?**.We believe that the tools we have developed and the framework we have provided can inspire followup works in these areas.

Our proposed algorithmic consequences also have the potential to help both researchers and practitioners to better develop their use cases of multi-task learning. For one example, many medical applications use multi-task learning to train large-scale image classification models by combining multiple datasets **??**. Unlike the applications of multi-task learning in text classification where we can collect large amounts of labeled data **?**, in medical applications it is typically difficult and expensive to acquire labeled data. For such settings, our proposed training scheduler might improve the model training efficiency by using less labeled data. For another example, practitioners in industry often need to improve prediction performance by training many related tasks. But the results are not always positive and practitioners have a hard time figuring out why. Our proposed metric can provide guidance for selecting which tasks should be trained together in a multi-task model.

References

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

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

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.

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.

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

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

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, 2013.

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

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

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

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

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

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

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

Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.

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.

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

Sen Wu, Hongyang 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.

A Proofs of the Bias and Variance Bounds

In random matrix theory, it is 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.

Basic setting. We denote the two sample covariance matrices by $\mathcal{Q}_1 := X_1^\top X_1$ and $\mathcal{Q}_2 := X_2^\top X_2$. 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

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

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 $d_1 := p/n_1$ and $d_2 := p/n_2$ satisfy that

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

Here the lower bound $1 + \tau \leqslant d_2$ is to ensure that the sample covariance matrix $\mathcal{Q}_2$ is non-singular with high probability; see Lemma **??** below.

We assume that Σ_1 and Σ_2 have eigendecompositions

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

where the eigenvalues satisfy that

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

We assume that $M = \Sigma_1^{1/2}\Sigma_2^{-1/2}$ has singular value decomposition

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

where the singular values satisfy that

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

We summarize our basic assumptions here for future reference. Note that this assumption is in accordance with Assumption **??**, except that we rescale the entries of Z_1 and Z_2 here.

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

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

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

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

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

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

for large enough $n \geqslant n_0(\varepsilon, D)$, and we shall use the notation $\xi \prec \zeta$. If for some complex family ξ we have $|\xi| \prec \zeta$, then we will also write $\xi \prec \zeta$ or $\xi = \mathrm{O}_\prec(\zeta)$.

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

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 (**??**) 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 (**??**), then using Markov's inequality one can show that Z has bounded support $n^{-1/2}$.

Then we state the following lemma on the eigenvalues of $Z_1^\top Z_1$ and $Z_2^\top Z_2$, which are denoted as $\lambda_1(Z_1^\top Z_1) \geqslant \cdots \geqslant \lambda_p(Z_1^\top Z_1)$ and $\lambda_1(Z_2^\top Z_2) \geqslant \cdots \geqslant \lambda_p(Z_2^\top Z_2)$. We have used it in our previous proofs; see (**??**).

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

$$\lambda_1(Z_1^\top Z_1) \leqslant (1 + \sqrt{d_1})^2 + n^\varepsilon q, \tag{A.9}$$

and

$$(1 - \sqrt{d_2})^2 - n^\varepsilon q \leqslant \lambda_p(Z_2^\top Z_2) \leqslant \lambda_1(Z_2^\top Z_2) \leqslant (1 + \sqrt{d_2})^2 + n^\varepsilon q. \tag{A.10}$$

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

In Section **??**, we introduce the concept of resolvent, and give an almost optimal convergent estimate on it—Theorem **??**. This estimate is conventionally called *local law* in random matrix theory literature. Based on Theorem **??**, we then complete the proof of Lemma **??** and Lemma **??**. The proof of Theorem **??** is presented in Section **??**.

A.1 Resolvent and Local Law

Our main goal is to study the matrix inverse $(\mathcal{Q}_1 + \mathcal{Q}_2)^{-1}$. Using (**??**), we can rewrite it as

$$(\mathcal{Q}_1 + \mathcal{Q}_2)^{-1} = \Sigma_2^{-1/2} V \left(\Lambda U^\top Z_1^\top Z_1 U \Lambda + V^\top Z_2^\top Z_2 V\right)^{-1} V^\top \Sigma_2^{-1/2}. \tag{A.11}$$

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 **???**. 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 (Green's function) as

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

For simplicity of notations, we define the index sets

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

We will consistently use the latin letters $i, j \in \mathcal{I}_1$, greek letters $\mu, \nu \in \mathcal{I}_2 \cup \mathcal{I}_3$, and $\mathfrak{a}, \mathfrak{b} \in \mathcal{I}$. We label the indices of the matrices according to

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

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

$$\mathcal{G}_L = \left(WW^\top - z\right)^{-1} = \mathcal{G}, \quad \mathcal{G}_{LR} = \mathcal{G}_{RL}^\top = \mathcal{G}W, \quad \mathcal{G}_R = z\left(W^\top W - z\right)^{-1}. \tag{A.15}$$

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

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

Our proof will use the spectral decomposition of G. Let $W = \sum_{k=1}^p \sqrt{\lambda_k}\xi_k\zeta_k^\top$ be a singular value decomposition of W, where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0 = \lambda_{p+1} = \ldots = \lambda_n$ are the eigenvalues, $\{\xi_k\}_{k=1}^p$ are the left-singular vectors, and $\{\zeta_k\}_{k=1}^n$ are the right-singular vectors. Then using (**??**), we get that for $i, j \in \mathcal{I}_1$ and $\mu, \nu \in \mathcal{I}_2 \cup \mathcal{I}_3$,

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

We now describe the asymptotic limit of $\mathcal{G}(z)$. First 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

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

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

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

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

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

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 (**??**) in are reduced to

$$r_1 b_2 + r_2 b_3 = 1 - \gamma_n, \quad b_2 + \frac{1}{n}\sum_{i=1}^p \frac{\sigma_i^2 b_2}{\sigma_i^2 r_1 b_2 + (1 - \gamma_n - r_1 b_2)} = 1. \tag{A.21}$$

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

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

is a strictly increasing function on $[0, r_1^{-1}(1 - \gamma_n)]$. Moreover, we have $f(0) = 0 < 1$ and $f(r_1^{-1}(1 - \gamma_n)) = 1 + \gamma_n > 1$. Hence by mean value theorem, there exists a unique solution $b_2 \in (0, r_1^{-1}(1 - \gamma_n))$. 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

$$r_1\tau \leqslant r_1 b_2 < \min\{(1-\gamma_n) - r_1\tau, r_1(1-\tau)\}, \quad \tau < r_2 b_3 \leqslant 1 - \gamma_n - r_1\tau. \tag{A.22}$$

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. Moreover, we will also include some basic estimates on it.

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

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

The proof is a standard application of the contraction principle. For reader's convenience, we will include its proof in Appendix **??**. 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 **??**.

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

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

Suppose they satisfy the system of equations

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

for some (random) errors satisfying $\max_{\alpha=2}^{3} |\mathcal{E}_\alpha| \leqslant \delta(z)$, where $\delta(z)$ is a deterministic z-dependent function with $\delta(z) \leqslant (\log n)^{-1}$. Then we have

$$\max_{\alpha=2}^{3} |m_\alpha(z) - m_{\alpha c}(z)| \leqslant C_0\delta(z). \tag{A.26}$$

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

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

The proof of this theorem will be given in Section **??**. We now use it to complete the proof of Lemma **??** and Lemma **??**.

Proof of Lemma **??**. In the setting of Lemma **??**, we write

$$\mathcal{R} := (X_1^\top X_1 + X_2^\top X_2)^{-1} = n^{-1}\left(\Sigma_1^{1/2} Z_1^\top Z_1 \Sigma_1^{1/2} + \Sigma_2^{1/2} Z_2^\top Z_2 \Sigma_2^{1/2}\right)^{-1},$$

where the extra n^{-1} is due to the choice of the scaling—in the setting of Lemma **??** the variances of the entries of $Z_{1,2}$ are equal to 1, while here they are taken to be n^{-1}. Then as in (**??**), we can write

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

If the entries of $\sqrt{n}Z_1$ and $\sqrt{n}Z_2$ have arbitrarily high moments as in (**??**), then Z_1 and Z_2 have bounded support $q = n^{-1/2}$. Using Theorem **??**, 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 (**??**), we have

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

with (b_2, b_3) satisfying (**??**). Thus from (**??**) we get that

$$\operatorname{Tr}(A\mathcal{R}) = n^{-1}\operatorname{Tr}(r_1 b_2 M^\top M + r_2 b_3)^{-1} + \mathrm{O}_\prec(n^{-1/2}\|A\|).$$

This concludes (**??**) if we rename $r_1 b_2 \to a_1$ and $r_2 b_3 \to a_2$.

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 (**??**). This gives (**??**) using (**??**). □

Proof of Lemma **??**. In the setting of Lemma **??**, we can write

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

Here again the n^2 factor disappears due to the choice of scaling. With (**??**), we can write the above expression as $\Delta := \mathbf{v}^\top(\mathcal{G}^2)(0)\,\mathbf{v}$ where $\mathbf{v} := V^\top \Sigma_2^{-1/2}\beta$. Note that $\mathcal{G}^2(0) = \partial_z \mathcal{G}|_{z=0}$. Now using Cauchy's integral formula and Theorem **??**, 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 $\{z \in \mathbb{C} : |z| \leqslant (\log n)^{-1}\}$. 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 $m_{3c}'(0)$.

By the implicit differentiation of (**??**), we obtain that

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

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

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

Then by (**??**) and (**??**), we get

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

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

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

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

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

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

Proof of Corollary **??**. 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

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

By the moment conditions (**??**) 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 (**??**) and integration by parts, it is easy to verify that

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

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

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

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

With this corollary, we can easily extend Lemma **??** and Lemma **??** to the case with weaker moment assumptions. Due to length constraint, we will not go into further details here.

A.2 Proof of the Anisotropic Local Law

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

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

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

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

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

and prove the following result.

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

This section is organized as follows. In Section **??**, we collect some basic estimates and resolvent identities that will be used in the proof of Theorem **??** and Proposition **??**. Then in Section **??** we give the proof of Proposition **??**, which concludes Theorem **??** for i.i.d. Gaussian Z_1 and Z_2. In Section **??**, we describe how to extend the result in Theorem **??** from the Gaussian case to the case with generally distributed entries of Z_1 and Z_2. Finally, in Section **??**, we give the proof of Lemma **??** and Lemma **??**. 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 **??**, 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* **??** *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 (**??**), we let $\{\lambda_k\}_{1\leqslant k\leqslant p}$ be the eigenvalues of $WW^\top$. By Lemma **??** and the assumption (**??**), 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 (**??**), we obtain (**??**) and (**??**). □

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 **?**)**.** *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 (??))*

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

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 (??). For convenience, we will adopt the convention that for any minor $\mathcal{A}^{(T)}$ defined as above, $\mathcal{A}^{(T)}_{ab}=0$ 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 $\mathfrak{b},\mathfrak{c}\in\mathcal{I}\setminus\{\mathfrak{a}\}$,

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

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

Lemma A.14 (Lemma 3.8 of ?)**.** *Let (x_i), (y_j) be independent families of centered and independent random variables, and (A_i), (B_{ij}) be families of deterministic complex numbers. Suppose the entries x_i, y_j have variances at most n^{-1} and satisfy the bounded support condition (??) with $q\leqslant n^{-\phi}$ for some constant $\phi>0$. Then we have the following bounds:*

$$\Big|\sum_i A_ix_i\Big|\prec q\max_i|A_i|+\frac{1}{\sqrt{n}}\Big(\sum_i|A_i|^2\Big)^{1/2},\quad \Big|\sum_{i,j}x_iB_{ij}y_j\Big|\prec q^2B_d+qB_o+\frac{1}{n}\Big(\sum_{i\neq j}|B_{ij}|^2\Big)^{1/2},$$
$$\Big|\sum_i\bar{x}_iB_{ii}x_i-\sum_i(\mathbb{E}|x_i|^2)B_{ii}\Big|\prec qB_d,\quad \Big|\sum_{i\neq j}\bar{x}_iB_{ij}x_j\Big|\prec qB_o+\frac{1}{n}\Big(\sum_{i\neq j}|B_{ij}|^2\Big)^{1/2},$$

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

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

A.2.2 Entrywise Local Law

The main goal of this subsection is to prove the following entrywise local law. The anisotropic local law (**??**) then follows from the entrywise local law combined with a polynomialization method as we will explain later. Recall that in the setting of Proposition **??**, 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* **??** *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 **??** 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 (**??**).

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 (**?**, Section 3) and (**?**, Section 5), we introduce the Z variables

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

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

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

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

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* **??** *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 **??**. Hence (**??**) and (**??**) also hold for $G^{(\mathfrak{a})}$. Now applying Lemma **??** to (**??**) and (**??**), and using the a priori bound (**??**), we get that for any $i \in \mathcal{I}_1$,

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

where in the last step we used (**??**) to get that for any μ,

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

Similarly, applying Lemma **??** to Z_μ and Z_ν in (**??**) and using (**??**), 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 (??), Lemma ?? and (??), we obtain that

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

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 (??). □

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

Step 2: Self-consistent equations. This is the key step of the proof for Proposition ??, 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 (??) for some small error $|\mathcal{E}_{2,3}| \prec n^{-1/2}$. Then in Step 3 we will apply Lemma ?? 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

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

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

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

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

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

Plugging (??) into (??), 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 ?? hold. Then the following estimates hold uniformly in $z \in \mathbf{D}$:*

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

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

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

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

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

where we recall Definition **??**, 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 (**??**) 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 (**??**) 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 (**??**), we obtain that for all i and μ,

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

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

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

Plugging it into the definitions of m and m_1 in (**??**), we get

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

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

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 (**??**), 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 (**??**), (**??**), (**??**), (**??**) and (**??**), 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 (**??**) and (**??**) into (**??**), we conclude (**??**). □

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 **??** to (**??**) 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 (??), then by (??) and (??), 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 (??), 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 (??) shows that Ξ holds with high probability, and the condition (??) verifies the condition (??) of Lemma ??. Then applying Lemma ?? to (??), 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 (??) into (??)-(??), 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 (??), we conclude (??). □

Now we give the proof of (??).

Proof of (??). By (??), 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 (??), (??) and (??), we obtain that (??) holds at $z=0$ even without the indicator function $\mathbf{1}(\Xi)$. Furthermore, it gives that

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

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

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

Denoting $\omega_2 = -m_{2c}(0)$ and $\omega_3 = -m_{2c}(0)$. By (??), 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 (**??**) that (ω_2, ω_3) satisfy approximately the same equations as (**??**):

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

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

With Lemma **??**, we can complete the proof of Proposition **??**.

Proof of Proposition **??**. With (**??**), one can use the polynomialization method in (**?**, Section 5) to get the anisotropic local law (**??**) 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 **??** for a general X satisfying the bounded support condition (**??**) with $q \leqslant n^{-\phi}$ for some constant $\phi > 0$. Proposition **??** implies that (**??**) 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 **??**,

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

for any deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}}$ and $z \in \mathbf{D}$. Here we abbreviated $Z := \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ and $Z^{Gauss} := \begin{pmatrix} Z_1^{Gauss} \\ Z_2^{Gauss} \end{pmatrix}$. We prove the above statement using a continuous comparison argument introduced in **?**. The proof is similar to the ones in Sections 7-8 of **?**, so we only give 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 (**??**) through interpolation matrices Z^θ between Z^0 and Z^1. We have see that (**??**) holds for Z^0 by Proposition **??**. Using (**??**) 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 (**??**) to $F(Z) := F_{\mathbf{u}\,\mathbf{v}}^s(Z,z)$ 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 (**??**) 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 (**??**) holds, then combining (**??**) with a Grönwall's argument we obtain that for any fixed $s\in 2\mathbb{N}$ and constant $\varepsilon>0$:

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

Together with Markov's inequality, we conclude (**??**). Underlying the proof of (**??**) is an expansion approach, which is very similar to the ones for Lemma 7.10 of **?** and Lemma 6.11 of **?**. So we omit the details.

A.2.4 Proofs of the Limiting Equations

Finally, we give the proof of Lemma **??** and Lemma **??** using the contraction principle.

Proof of Lemma **??**. One can check that the equations in (**??**) are equivalent to the following ones:

$$r_1m_{2c}=-(1-\gamma_n)-r_2m_{3c}-z\left(m_{3c}^{-1}+1\right),\quad g_z(m_{3c}(z))=1, \tag{A.76}$$

where

$$g_z(m_{3c}):=-m_{3c}+\frac{1}{n}\sum_{i=1}^{p}\frac{m_{3c}}{z-\sigma_i^2(1-\gamma_n)+(1-\sigma_i^2)r_2m_{3c}-\sigma_i^2z\left(m_{3c}^{-1}+1\right)}.$$

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 (**??**), and the solution satisfies (**??**). Now we abbreviate $\varepsilon(z):=m_{3c}(z)-m_{3c}(0)$, and from (**??**) 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 (**??**) defines a mapping $h:\mathbb{C}\to\mathbb{C}$, which maps $\varepsilon^{(k)}$ to $\varepsilon^{(k+1)}=h(\varepsilon^{(k)})$.

With direct calculation, one can get the derivative

$$g_z'(m_{3c}(0))=-1-\frac{1}{n}\sum_{i=1}^{p}\frac{\sigma_i^2(1-\gamma_n)-z\left[1-\sigma_i^2\left(2m_{3c}^{-1}(0)+1\right)\right]}{\left[z-\sigma_i^2(1-\gamma_n)+(1-\sigma_i^2)r_2m_{3c}(0)-\sigma_i^2z\left(m_{3c}^{-1}(0)+1\right)\right]^2}.$$

Then it is easy to check that there exist constants $\widetilde{c},\widetilde{C}>0$ depending only on τ in (**??**) and (**??**) 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 (??) and (??), 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 (??) subject to the condition $\|\varepsilon\|_\infty \leqslant r$.

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

Note that with (??) and $\varepsilon^{(0)} = 0$, we get from (??) 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 (??) for $m_{3c}(z)$. Using the first equation in (??), we immediately obtain the bound $r_1|m_{2c}(z) - m_{2c}(0)| \leqslant C|z|$. This gives (??) 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 (??) and treat $m_{2c}(z)$ as the solution to the following equation:

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

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

Proof of Lemma ??. Under (??), we can obtain equation (??) 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 (??) from (??), 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 ??, so we omit the details. □