Sharp Bias-variance Tradeoffs of Hard Parameter Sharing in High-dimensional Linear Regression

Anonymous Author
Anonymous Institution

Abstract

Hard parameter sharing for multi-task learning is widely used in empirical research despite the fact that its generalization properties have not been well established in many cases. This paper studies its generalization properties in a fundamental setting: How does hard parameter sharing work given multiple linear regression tasks? We develop new techniques and establish a number of new results in the high-dimensional setting, where the sample size and feature dimension increase at a fixed ratio. First, we show a sharp bias-variance decomposition of hard parameter sharing, given multiple tasks with the same features. Second, we characterize the asymptotic bias-variance limit for two tasks, even when they have arbitrarily different sample size ratios and covariate shifts. We also demonstrate that these limiting estimates for the empirical loss are incredibly accurate in moderate dimensions. Finally, we explain an intriguing phenomenon where increasing one task's sample size helps another task initially by reducing variance but hurts eventually due to increasing bias. This suggests progressively adding data for optimizing hard parameter sharing, and we validate its efficiency in text classification tasks.

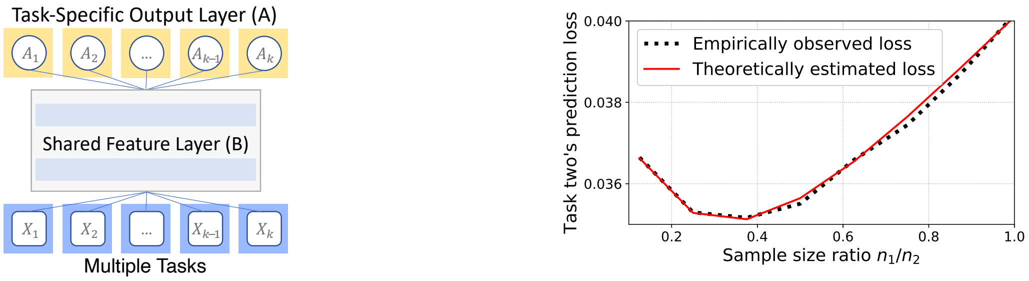
1 Introduction

Hard parameter sharing (HPS) for multi-task learning is widely used in empirical research and goes back to the seminal work of Caruana (1997). Recent work has revived interests in this approach because it improves performance and reduces the cost of collecting labeled data (Ruder, 2017). It is generally applied by sharing the feature layers between all tasks while keeping an output layer for every task. Often, hard parameter sharing offers two critical advantages if successfully applied. First, it reduces model parameters since all tasks use the same feature space. Second, it reduces the amount of labeled data needed from each task by augmenting the entire training dataset.

Hard parameter sharing works as an inductive transfer mechanism and a regularizer that reduces overfitting, both of which have great intuitive appeal (Ruder, 2017). For example, by restricting the shared space's size, HPS encourages information sharing among multiple tasks (Kumar and Daumé III, 2012). Another source of inductive bias comes from the tasks and depends on datasets' properties such as sample sizes and task covariances (Wu et al., 2020). However, how these dataset properties impact HPS has not been well established. Part of the challenge may be that HPS' generalization performance depends intricately on the sample size ratios and covariate shifts between tasks, and is not amenable to standard concentration results. Previous results based on Rademacher complexity or VC dimensions have considered cases where all tasks' sample sizes are equal to logarithmic factors of the feature dimension (Baxter, 2000; Maurer et al., 2016), and when all tasks' sample sizes increase simultaneously (Ando and Zhang, 2005; Maurer, 2006).

This paper presents new techniques to study hard parameter sharing and establishes a number of new results. We consider regression analysis, which is arguably one of the most fundamental problems in statistics and machine learning. We are interested in the *high-dimensional* setting, where each dataset's sample size and feature dimension grow linearly instead of logarithmically. This setting captures the fact that a single task's sample size is usually insufficient for accurate learning in many applications. For example, if a dataset's sample size is only a constant factor of dimension in linear regression, the variance is also constant (cf. Fact 2.3). The high-dimensional setting is challenging but is crucial for understanding how datasets' sample sizes impact generalization performance.

Preliminary work. Under review by AISTATS 2021. Do not distribute.



(a) A hard parameter sharing architecture

(b) Varying sample size ratio

Figure 1: Left: an illustrative picture of HPS. Right: an illustrative example of using HPS for two tasks X_1, Y_1 and X_2, Y_2 with sample size n_1, n_2, respectively. Increasing n_1/n_2 decreases task two's prediction loss initially but increase afterward. This phenomenon occurs due to different bias-variance tradeoffs as n_1/n_2 increases. Our result provides an estimated loss (solid line) that accurately matches the empirical loss (dotted line). See Section 4 for the precise setting.

1.1 Setup and Main Results

Suppose we have t datasets. For each dataset i from 1 to t, let n_i denote its sample size. Let $X^{(i)} \in \mathbb{R}^{n_i \times p}$ denote dataset i's feature covariates. We assume that the label vector $Y^{(i)} \in \mathbb{R}^{n_i}$ for $X^{(i)}$ follows a linear model plus random noise. We study the standard hard parameter sharing architecture: a shared feature representation layer $B \in \mathbb{R}^{p \times r}$ for all datasets and a separate output layer $A_i \in \mathbb{R}^r$ for every dataset i. See Figure 1a for an illustration. We study the following minimization problem:

$$f(A, B) = \sum_{i=1}^{t} \|X^{(i)} B A_i - Y^{(i)}\|^2, \qquad (1.1)$$

where $A = [A_1, A_2, \dots, A_t] \in \mathbb{R}^{r \times t}$. Given a solution from minimizing $f(A, B)$, denoted by $(\hat{A}, \hat{B})$ (which we will specify below), let $\hat{\beta}_i^{\text{HPS}} = \hat{B}\hat{A}_i$ denote the HPS estimator for task i. The critical questions are: (i) How well does the estimator work? In particular, how does the performance of the estimator scale with sample size? (ii) For datasets with different sample sizes and covariate shifts, how do they affect the estimator?

Main results. Our first result (Theorem 2.1) applies to multi-label prediction settings where all datasets have the same features (and sample size), and we want to make several predictions for every input (cf. examples in Hsu et al. (2009)). We analyze the global minimizer of $f(A, B)$, and provide a sharp bias-variance decomposition of its (out-of-sample) prediction loss for any task. This setting is tractable even though in general, $f(A, B)$ is non-convex in A and B (e.g. matrix completion is a special case for suitably designed $X^{(i)}, Y^{(i)}$). Our result implies that when all tasks have the same features but different labels, for any task, HPS helps reduce the task's variance compared to single-task learning but increases bias.

Our second result (Theorem 3.1) applies to two tasks with arbitrarily different sample size ratios and covariate shifts. While we can no longer characterize $f(A, B)$'s global minimum because of non-convexity, we can still provide a sharp bias-variance tradeoff of any local minimizer's prediction loss for both tasks. Despite being a simple setting, we observe several nontrivial phenomena by varying sample size ratios and covariate shifts between the two tasks. See Figure 1b for an illustration of the former. Consequently, using our precise loss estimates, we observe several qualitative properties of HPS for varying dataset properties.

Sample efficiency (Example 2.4): One advantage of combining multiple datasets is that the requirement for labeled data reduces compared to single-task learning, a phenomenon that Zamir et al. (2018) has observed empirically. Our results further imply that HPS's sample efficiency depends on model-specific variances across tasks vs. the noise variance and is generally high when the latter is large.

Sample size ratio (Example 3.2): Increasing one task's sample size does not always reduce another task's loss. In a simplified setting, we find that the task loss either decreases first before increasing afterward or decreases monotonically depending on how fast the bias grows. These two trends result from different bias-variance tradeoffs. This result is surprising because previous generalization bounds in multi-task learning typically scale down as all tasks' sample sizes increase, thus do not apply for different sample size ratios.

Covariate shift (Example 3.4): In addition to sample sizes, variance also scales with two datasets' covariate shifts. For a large sample size ratio, HPS's variance is smallest when there is no covariate shift. Counterintuitively, for a small sample size ratio, having covariate shifts reduces variance through a complementary spectrum. We achieve this result through a novel characterization on the inverse of the sum of two sample covariance matrices with arbitrary covariate shifts. See our discussion of proof techniques below for details.

Finally, we discuss the practical implications of our work. Our sample size ratio study implies a concrete progressive training procedure that gradually adds more data until performance drops. For example, in the setting of Figure 1b, this procedure will stop right at the minimum of the local basin. We conduct further studies of this procedure on six text classification datasets and observe that it reduces the computational cost by 65% compared to a standard round-robin training procedure while keeping the average accuracy of all tasks simultaneously.

Proof techniques. There are two main ideas in our analysis. The proof of our first result uses a geometric intuition that hard parameter sharing finds a "rank-r" approximation of the datasets. We carefully keep track of the concentration error between the global minimizer of $f(A, B)$ and its population version (cf. equation (2.1)). The proof of our second result is significantly more involved because of different sample sizes and covariate shifts. We show that the inverse of the sum of two sample covariance matrices with arbitrary covariate shifts converges to a deterministic diagonal matrix asymptotically (cf. Theorem 3.1). We use recently developed techniques from random matrix theory to show a sharp convergence rate. One limitation of our analysis is that in Example 3.2, there is an error term that can result in vacuous bounds for very small n_1 (cf. equation (3.7)). We believe our result has provided significant initial insights, and it is an interesting question to tighten our result.

1.2 Related Work

There is a large body of classical and recent works on multi-task learning. We focus our discussion on theoretical results and refer interested readers to several excellent surveys for general references (Pan and Yang, 2009; Zhang and Yang, 2017; Vandenhende et al., 2020). The early work of Baxter (2000); Ben-David and Schuller (2003); Maurer (2006) studied multi-task learning from a theoretical perspective, often using uniform convergence or Rademacher complexity based techniques. An influential paper by Ben-David et al. (2010) provides uniform convergence bounds that combine multiple datasets in certain settings. One limitation of uniform convergence based techniques is that the results often assume that all tasks have the same sample size, see e.g. Baxter (2000); Maurer et al. (2016). Besides, these techniques do not apply to the high-dimensional setting where the sample size is only a small constant times the dimension.

Our proof techniques use the so-called local law of random matrices (Erdos and Yau, 2017), a recent development in the random matrix theory literature. In the single-task case, Bloemendal et al. (2014) first proved such a local law for sample covariance matrices with isotropic covariance. Knowles and Yin (2016) later extended this result to arbitrary covariance matrices. These techniques provide almost sharp convergence rates to the asymptotic limit compared to other methods such as free probability (Nica and Speicher, 2006). To the best of our knowledge, we are not aware of any previous results in the multi-task case, even for two tasks (with arbitrary covariate shifts).

The problem we study here is also related to high-dimensional prediction in transfer learning (Li et al., 2020; Bastani, 2020) and distributed learning (Dobriban et al., 2018). For example, Li et al. (2020) provide minimax-optimal rates to predict a target regression task given multiple sparse regression tasks. One closely related work is Wu et al. (2020), which studied hard parameter sharing for two linear regression tasks. However, their results only apply to sample size regimes at least logarithmic factors of dimension.

Organizations. The rest of this paper is organized as follows. In Section 2, we present the bias-variance decomposition for hard parameter sharing. In Section 3, we describe how varying sample sizes and covariate shifts impact hard parameter sharing using random matrix theory. In Sections 4, we validate our results in simulations. In Section 5, we summarize our work and discuss future work. Section A describes our study on text classification tasks. Section B, C, and D present proofs of our results.

Notations. For an $n \times p$ matrix X, let $\lambda_{\min}(X)$ denote its smallest singular value and $\|X\|$ denote its largest singular value. Let $\lambda_1(X), \lambda_2(X), \cdots, \lambda_{p\wedge n}(X)$ denote the singular values of X in decreasing order. Let X^+ denote the Moore-Penrose psuedoinverse of X. We refer to random matrices of the form $\frac{X^\top X}{n}$ as sample covariance matrices. We say that an event Ξ holds with high probability if the probability that Ξ happens goes to 1 as p goes to infinity.

2 A Bias-variance Decomposition for Multiple Tasks

In this section, we show that the prediction loss of hard parameter sharing admits a clean bias-variance decomposition, when all tasks have the same features.

Setting. Suppose we have t datasets whose sample sizes are all equal to n and whose features are all denoted by $X \in \mathbb{R}^{n\times p}$. The label vector of the i-th task follows a linear model $Y^{(i)} = X\beta^{(i)} + \varepsilon^{(i)}$. We assume: (i) $X = Z\Sigma^{1/2} \in \mathbb{R}^{n\times p}$ for a positive semidefinite matrix $\Sigma \in \mathbb{R}^{p\times p}$, and every entry of $Z \in \mathbb{R}^{n\times p}$ is drawn independently from a one dimensional distribution with zero mean, unit variance, and constant φ-th moment for a fixed $\varphi > 4$.

(ii) every entry of $\varepsilon^{(i)} \in \mathbb{R}^{n\times t}$ is drawn indepdently from a one dimensional distribution with zero mean, variance σ^2, and bounded moment up to any order.[1]

For an estimator $\hat{\beta}_i$ of task i, we are interested in its (out-of-sample) prediction loss

$$L(\hat{\beta}_i) = \left\|\Sigma^{1/2}(\hat{\beta}_i - \beta^{(i)})\right\|^2.$$

Recall that r is the width of B. We focus on cases where $r < t$, because otherwise the global minimum of $f(A,B)$ reduces to single-task learning (cf. Proposition 1 of Wu et al. (2020)).

Our first main result shows that hard parameter sharing essentially approximates all tasks through a rank-r subspace. To formalize this geometric intuition, we introduce the matrix $B^\star := [\beta^{(1)}, \beta^{(2)}, \dots, \beta^{(t)}] \in \mathbb{R}^{p\times t}$ which contains all the linear model parameters. Let $A^\star A^{\star\top}$ denote the best rank-r subspace approximation of $B^{\star\top}\Sigma B^\star$ (which is task labels' "covariance"):[2]

$$A^\star := \mathop{\arg\min}_{U\in\mathbb{R}^{t\times r}:U^\top U = \mathrm{Id}_{r\times r}} \langle UU^\top, B^{\star\top}\Sigma B^\star\rangle. \quad (2.1)$$

Let $a_i^\star \in \mathbb{R}^r$ denote the i-th column of $A^\star A^{\star\top}$. We show that the prediction loss of HPS decomposes into a bias term $L(B^\star a_i^\star)$ that measures the prediction loss of $B^\star a_i^\star$, plus a variance term that scales with $\|a_i^\star\|^2$. Let $(\hat{A}, \hat{B})$ be the global minimizer of $f(A,B)$. Recall that the HPS estimator is defined as $\hat{\beta}_i^{\mathrm{HPS}} = \hat{B}\hat{A}_i$. Our result is stated as follows.

Theorem 2.1. *Assume that $n > \rho\cdot p$ for a fixed constant $\rho > 1$. Let c_φ be any fixed value within $(0, \frac{\varphi-4}{2\varphi})$. For any task $i = 1, 2, \dots, t$, with high probability over the randomness of the input, the prediction loss of $\hat{\beta}_i^{\mathrm{HPS}}$ satisfies that*

$$\left|L(\hat{\beta}_i^{\mathrm{HPS}}) - L(B^\star a_i^\star) - \sigma^2\|a_i^\star\|^2 \operatorname{Tr}\left[\Sigma(X^\top X)^{-1}\right]\right|$$
$$\leqslant n^{-\frac{c_\varphi}{2}} \cdot \frac{t\left(\|\Sigma^{1/2}B^\star\|^2 + \sigma^2\right)^2}{\lambda_r(B^{\star\top}\Sigma B^\star) - \lambda_{r+1}(B^{\star\top}\Sigma B^\star)}.$$

Comparison to single-task learning (STL). Theorem 2.1 provides a sharp generalization error bound that is asymptotically tight when n goes to infinity. One direct implication of our result is that compared to STL, the variance always decreases, since STL's variance is equal to $\sigma^2 \operatorname{Tr}[\Sigma(X^\top X)^{-1}]$. On the other hand, the bias always increases.

How does hard parameter sharing scale with sample size n? Obviously, the concentration error decreases with n. First, we consider the variance of $\hat{\beta}_i^{\mathrm{HPS}}$, which is $\sigma^2\|a_i^\star\| \operatorname{Tr}\left[\Sigma(X^\top X)^{-1}\right]$? It turns out that this quantity converges to a fixed limit in the high-dimensional setting, which is formally stated in the following assumption.

Assumption 2.2. Let $\tau > 0$ be a small enough constant. In the high-dimensional setting, the sample size n grows to infinity proportionally with the feature dimension p, i.e. $n/p \to \rho \in (\tau, 1/\tau)$ as p goes to infinity.

Under the above assumption, we can use the following result to simplify the variance of $\hat{\beta}_i^{\mathrm{HPS}}$.

Fact 2.3 (cf. Theorem 2.4 in Bloemendal et al. (2014)). With high probability over the randomness of X, we have that

$$\operatorname{Tr}\left[\Sigma(X^\top X)^{-1}\right] = \frac{p}{n-p} \pm \mathrm{O}(n^{-c_\varphi}).$$

Remark. The above result has a long history in random matrix theory. For a multivariate Gaussian random matrix, this result follows from the classical result for the mean of inverse Wishart distribution (Anderson, 2003). For non-Gaussian random matrices, this result can be derived via the well-known Stieltjes transform method (cf. Lemma 3.11 of Bai and Silverstein (2010)). Applying Fact 2.3 to Theorem 2.1, we obtain that hard parameter sharing's variance is

$$\sigma^2\|a_i^\star\|^2 \operatorname{Tr}\left[\Sigma(X^\top X)^{-1}\right] = \sigma^2\|a_i^\star\|^2 \frac{p}{n-p} \pm \mathrm{O}(p^{-c_\varphi}).$$

Next, we consider the bias of $\hat{\beta}_i^{\mathrm{HPS}}$, which is $L(B^\star a_i^\star)$. We illustrate the bias through a random-effect model, which has been studied for a single-task case (Dobriban and Sheng, 2020). Suppose every $\beta^{(i)}$ consists of two random components, one that is shared among all tasks and one that is task-specific. Thus, each task contributes a certain amount to the shared component and injects a task-specific bias. Let β_0 denote the shared component whose entries are sampled i.i.d. from an isotropic Gaussian distribution of mean zero and variance $p^{-1}\kappa^2$. Let $\beta^{(i)}$ be equal to β_0 plus a task-specific component that is a random Gaussian vector with i.i.d. entries of mean zero and variance $p^{-1}d^2$. Thus, for any two different $\beta^{(i)}$ and $\beta^{(j)}$, their distance is roughly $2d^2$. Concretely, we can think of $\kappa = 1$ and $d^2/\sigma^2 = \mathrm{O}(1)$.

Example 2.4 (Sample efficiency)**.** *In the random-effect model described above, we further assume that Σ is isotropic as an example. We show that when the rank r is one, the average prediction loss of hard parameter sharing is as follows*

$$\frac{1}{t}\sum_{i=1}^{t} L(\hat{\beta}_i^{\mathrm{HPS}}) = \left(1 - \frac{1}{t}\right)d^2 + \frac{1}{t}\cdot\frac{\sigma^2 p}{n-p} \pm \mathrm{O}(n^{-\frac{c_\varphi}{2}}).$$

[1] There exists a fixed function $C : \mathbb{N} \to \mathbb{R}^+$ such that for any $k \in \mathbb{N}$, the k-th moment is bounded by $C(k)$.

[2] To ensure that $A^\star$ is unique, we assume that $\lambda_{r+1}(B^{\star\top}\Sigma B^\star)$ is strictly smaller than $\lambda_r(B^{\star\top}\Sigma B^\star)$.

We describe a proof sketch. First, we show that the bias equation $L(B^\star a_i^\star)$ simplifies to the following

$$\frac{1}{t}\sum_{i=1}^{t} L(B^\star a_i^\star) = \frac{1}{t}\|B^\star A^\star {A^\star}^\top - B^\star\|_F^2 \approx \left(1-\frac{1}{t}\right) d^2.$$

To see this, recall that r is one and $A^\star {A^\star}^\top$ is the best rank-1 approximation of ${B^\star}^\top \Sigma B^\star = {B^\star}^\top B^\star$. Hence, the above expression is equal to the sum of ${B^\star}^\top B^\star$'s bottom $t-1$ singular values. Based on the definition of the random-effect model, the (i,j)-th entry of ${B^\star}^\top B^\star$ is equal to (ignoring lower order terms)

$$\beta_i^\top \beta_j = \|\beta_0\|^2 + \begin{cases} 0, & \text{if } i \neq j \\ d^2, & \text{if } i = j \end{cases}$$

Note that $\|\beta_0\|^2$ is approximately κ^2. Then, one can verify that the top eigenvalue of ${B^\star}^\top B^\star$ is $t\kappa^2 + d^2$ and the rest of its eigenvalues are all d^2. Therefore, by taking a rank-1 approximation of ${B^\star}^\top B^\star$, we get the average prediction loss of $B^\star a_i^\star$.

Second, using Fact 2.3, one can see that the average variance is

$$\begin{aligned}\frac{1}{t}\sum_{i=1}^{t} \sigma^2 \|a_i^\star\|^2 \operatorname{Tr}\left[\Sigma(X^\top X)^{-1}\right] &= \frac{\sigma^2}{t}\sum_{i=1}^{t}\|a_i^\star\|^2 \frac{p}{n-p} \\ &= \frac{1}{t}\frac{\sigma^2 p}{n-p},\end{aligned}$$

because $A^\star$ has rank-1 and $\sum_{i=1}^t \|a_i^\star\|^2 = 1$. Combined together, we have derived the average prediction loss in the random-effect model.

Comparison to single-task learning. Recall that the average prediction loss of STL scales as $\sigma^2 \cdot \operatorname{Tr}\left[\Sigma(X^\top X)^{-1}\right] = \frac{\sigma^2 p}{n-p}$ by Fact 2.3. Comparing HPS to STL, we have the following qualitative properties.

(i) The prediction loss of HPS is smaller than STL if and only if $d^2 < \frac{\sigma^2 p}{n-p}$, that is, the "task-specific variance" of $\beta^{(i)}$ is smaller than the "noise variance".

(iii) HPS requires at most $p + \frac{n-p}{t-(t-1)\frac{d^2(n-p)}{\sigma^2 p}}$ samples that is less than n samples to get comparable loss to STL. This follows by using this sample size in the average prediction loss equation in Example 2.4.

(ii) When $d^2 < \frac{\sigma^2 p}{n-p}$, increasing r does not help. To see this, one can verify what when r increases by one, bias reduces by $\frac{d^2}{t}$, but variance increases by $\frac{\sigma^2 p}{t(n-p)} > \frac{d^2}{t}$ (details omitted).

Proof overview. The key step for proving Theorem 2.1 is a characterization of $f(A,B)$'s global minimizer. In the setting of this theorem, the minimization problem (1.1) becomes

$$f(A,B) = \sum_{j=1}^{t}\left\|XBA_j - Y^{(j)}\right\|^2, \qquad (2.2)$$

where we recall that $B \in \mathbb{R}^{p\times r}$ and $A_1, A_2, \dots, A_t \in \mathbb{R}^r$. Using the local optimality condition over B, that is, $\frac{\partial f}{\partial B} = 0$, we obtain $\hat{B}$ as a function of A as follows

$$\begin{aligned}\hat{B}(A) &:= (X^\top X)^{-1} X^\top \left(\sum_{j=1}^{t} Y^{(j)} A_j^\top\right)(AA^\top)^+ \\ &= (X^\top X)^{-1} X^\top Y A^\top (AA^\top)^+, \qquad (2.3)\end{aligned}$$

where $Y = [Y^{(1)}, Y^{(2)}, \dots, Y^{(t)}]$. Here we have used that $X^\top X$ is invertible since $n > \rho \cdot p$ and $\rho > 1$ (cf. Fact E.1). Plugging $\hat{B}(A)$ into equation (2.2), we obtain the following objective that only depends on A (in matrix notation):

$$g(A) = \left\|X(X^\top X)^{-1}X^\top Y A^\top (AA^\top)^+ A - Y\right\|_F^2. \qquad (2.4)$$

Let $\hat{A}$ be the global minimizer of $g(A)$. Then $(\hat{A}, \hat{B}(\hat{A}))$ is the global minimizer of $f(A,B)$. Our main idea is to show that the subspaces spanned by the rows of $\hat{A}$ and $A^\star$ are close to each other. We carefully keep track of the concentration error between $\hat{A}$ and $A^\star$. The proof can be found in Section B.

3 Bias-variance Limits: Different Sample Sizes and Covariate Shifts

The previous section assumes that all tasks have the same sample size and feature vectors. This section discusses how having different sample sizes and different covariance matrices impact hard parameter sharing. The setting where covariates differ across tasks is often known as "covariate shift".

Unlike the previous section, we can no longer characterize the global minimum of $f(A,B)$. This is because $f(A,B)$ is in general non-convex. Instead, our result implies sharp bias-variance tradeoffs for any *local minimizer* of $f(A,B)$. We focus on the two-task case to better understand the impact of having different sample sizes and covariates. Let n_1, n_2 denote task one and two's sample size, respectively. Suppose

$$\begin{aligned}X^{(1)} &= Z^{(1)}(\Sigma^{(1)})^{1/2} \in \mathbb{R}^{n_1\times p}, \text{ and} \\ X^{(2)} &= Z^{(2)}(\Sigma^{(2)})^{1/2} \in \mathbb{R}^{n_2\times p},\end{aligned}$$

where the entries of $Z^{(1)}$ and $Z^{(2)}$ are drawn independently from a one dimensional distribution with zero mean, unit variance, and constant φ-th moment for a fixed $\varphi > 4$. $\Sigma^{(1)} \in \mathbb{R}^{p\times p}$ and $\Sigma^{(2)} \in \mathbb{R}^{p\times p}$ denote the population covariance matrices of task 1 and task 2, respectively.

Bias-variance equations. Our key result characterizes the asymptotic limit of the inverse of the sum of two arbitrarily different sample covariance matrices. Without loss of generality, we consider task two's prediction loss and the same result applies to task one. We consider the case of $r = 1 < t = 2$, since when $r > 1$, the global minimum of $f(A, B)$ reduces to single-task learning (cf. Proposition 1 of Wu et al. (2020)). When $r = 1$, B is a vector and A_1, A_2 are both scalars. To motivate our study, we consider a special case where $A_1 = A_2 = 1$. Hence the HPS estimator is equal to B. By solving B in equation (1.1), we obtain the estimator for task two as follows:

$$\hat{\beta}_2^{\text{HPS}} = \hat{\Sigma}^{-1}({X^{(1)}}^\top Y^{(1)} + {X^{(2)}}^\top Y^{(2)}), \quad \text{where}$$
$$\hat{\Sigma} = {X^{(1)}}^\top X^{(1)} + {X^{(2)}}^\top X^{(2)}. \tag{3.1}$$

The matrix $\hat{\Sigma}$ adds up both tasks' sample covariance matrices, and the expectation of $\hat{\Sigma}$ is equal to a mixture of their population covariance matrices, with mixing proportions determined by their sample sizes.

To derive the bias and variance equation, we consider the expected loss conditional on the covariates as follows (the empirical loss is close to this expectation as will be shown in equation (D.13)):

$$\mathop{\mathbb{E}}_{\varepsilon}\left[L(\hat{\beta}_2^{\text{HPS}}) \mid X^{(1)}, X^{(2)}\right]$$
$$= \left\|{\Sigma^{(2)}}^{1/2}\hat{\Sigma}^{-1}{X^{(1)}}^\top X^{(1)}(\beta^{(1)} - \beta^{(2)})\right\|^2 \tag{3.2}$$
$$+ \sigma^2 \operatorname{Tr}\left[\Sigma^{(2)}\hat{\Sigma}^{-1}\right]. \tag{3.3}$$

Equations (3.2) and (3.3) correspond to the bias and variance of HPS for two tasks, respectively. Our main result in this section characterizes the asymptotic bias-variance limits in the high-dimensional setting. Intuitively, the spectrum of $\hat{\Sigma}^{-1}$ (and hence its trace) not only depends on both tasks' sample sizes, but also depends on the "alignment" between $\Sigma^{(1)}$ and $\Sigma^{(2)}$. However, capturing this intuition quantitatively turns out to be technically challenging. We introduce a key quantity $M := (\Sigma^{(1)})^{1/2}(\Sigma^{(2)})^{-1/2}$, and as we show below, the trace of $\hat{\Sigma}^{-1}$ has an intricate dependence on the spectrum of M.

Let $\lambda_1, \lambda_2, \dots, \lambda_p$ denote M's singular values in descending order. Our main result is stated as follows.

Theorem 3.1. *Let c_φ be any fixed value within $(0, \frac{\varphi-4}{2\varphi})$. Assume that: a) the sample sizes n_1 and n_2 both satisfy Assumption 2.2; b) M's singular values are all greater than τ and less than $1/\tau$; c) task one's sample size is greater than τp and task two's sample size is greater than $(1+\tau)p$. With high probability over the randomness of $X^{(1)}$ and $X^{(2)}$, we have that the variance equation* (3.3) $\operatorname{Tr}[\Sigma^{(2)}\hat{\Sigma}^{-1}]$ *(leaving out σ^2) satisfies the following estimate:*

$$\left|\operatorname{Tr}\left[\Sigma^{(2)}\left(\hat{\Sigma}^{-1} - \tfrac{(a_1\Sigma^{(1)} + a_2\Sigma^{(2)})^{-1}}{n_1+n_2}\right)\right]\right| \leqslant p^{-c_\varphi}, \tag{3.4}$$

where a_1 and a_2 are the solutions of the following self-consistent equations

$$a_1 + a_2 = 1 - \frac{p}{n_1+n_2}, \tag{3.5}$$
$$a_1 + \frac{1}{n_1+n_2}\cdot\left(\sum_{i=1}^p \frac{\lambda_i^2 a_1}{\lambda_i^2 a_1 + a_2}\right) = \frac{n_1}{n_1+n_2}. \tag{3.6}$$

Due to space limit, we defer the bias limit result to Appendix (C). Our result extends Fact 2.3 to the inverse of the sum of two sample covariance matrices. To see this, when n_1 is zero, we solve equations (3.5) and (3.6) to obtain that $a_1 = 0$ and $a_2 = (n_2 - p)/n_2$, and apply them to equation (3.4). For general A_1, A_2 that are not equal to one, we can still apply our result by rescaling $X^{(1)}$ and M with A_1/A_2. We defer a proof sketch of Theorem 3.1 until the end of the section.

How does hard parameter sharing scale with sample sizes and covariate shift M**?** One can see that the variance limit depends intricately on both tasks' samples sizes and covariate shift. Next, we illustrate how varying them impact the prediction loss.

Example 3.2 (Sample size ratio)**.** *We first consider the impact of varying sample sizes. Consider the random-effect model from Section 2, with both tasks having an isotropic population covariance matrix.*

Applying Theorem 3.1 to the above setting, we get that

$$\frac{1}{n_1+n_2}\operatorname{Tr}[\Sigma^{(2)}(a_1\Sigma^{(1)} + a_2\Sigma^{(2)})^{-1}]$$
$$= \frac{1}{n_1+n_2}\operatorname{Tr}\left[((a_1+a_2)\operatorname{Id}_p)^{-1}\right] = \frac{p}{n_1+n_2-p},$$

because $a_1 + a_2 = 1 - \frac{p}{n_1+n_2}$ by equation (3.5). Similarly, we can calculate the bias limit (details omitted). Combined together, we obtain the following corollary of Theorem 3.1.

Corollary 3.3. *In the setting of Example 3.2, assume that (i) both tasks sample sizes are at least $3p$; (ii) noise variance is smaller than the shared signal variance: $\sigma^2 \lesssim \kappa^2$; (iii) task-specific variance is much smaller than the shared signal variance: $d^2 \leqslant p^{-c}\kappa^2$ for a small constant $c > 0$. Let $\varepsilon = (1 + \sqrt{p/n_1})^4 - 1$, which decreases as n_1 increases. Let $\hat{A}, \hat{B}$ be the global minimizer of $f(A, B)$. With high probability over the randomness of the input, the prediction loss of $\hat{\beta}_2^{\text{HPS}} = \hat{B}\hat{A}_2$ for task two satisfies that*

$$\left|L(\hat{\beta}_2^{\text{HPS}}) - \frac{2d^2 n_1^2(n_1+n_2)}{(n_1+n_2-p)^3} - \frac{\sigma^2 p}{n_1+n_2-p}\right|$$
$$\leqslant \varepsilon \cdot \frac{2d^2 n_1^2(n_1+n_2)}{(n_1+n_2-p)^3} + \mathrm{O}(p^{-c/2}). \tag{3.7}$$

In the above inequality, the d^2 scaling term is the bias limit, and the σ^2 scaling term is the variance limit. This result allows for a more concrete interpretation since the dependence on datasets' properties is explicit. The proof of Corollary 3.3 can be found in Appendix D. As a remark, by combining the bias and variance limits, we can also obtain a bias-variance tradeoff for any local minimizer of $f(A, B)$. The proof is similar to Corollary 3.3, so we omit the details.

Next, we use the bias-variance limits to study how varying sample sizes impacts HPS. For example, imagine if we want to decide whether to collect more of task one's data or not, how does increasing n_1 affect the prediction loss? We assume that n_2 is fixed for simplicity. The variance limit in equation (3.7) obvious decreases with n_1. It turns out that the bias term always increases with n_1, which can be verified by showing that the bias limit's derivative is always nonnegative. By comparing the derivative of the bias and variance limits with respect to n_1 (details omitted), we obtain the following dichotomy.

(i) When $\frac{d^2}{\sigma^2} < \frac{p}{4n_2 - 6p}$, the prediction loss decreases monotonically as n_1 increases. Intuitively, this regime of d^2 always helps task two.

(ii) When $\frac{d^2}{\sigma^2} > \frac{p}{4n_2 - 6p}$, the prediction loss always decreases first from $\frac{\sigma^2 p}{n_2 - p}$ (when $n_1 = 0$), and then increases to d^2 (when $n_1 \to \infty$). To see this, near the point where n_1 is zero, one can verify (from the derivatives) that bias increases less while variance decreases more, and there is *exactly* one critical point where the derivative is zero, which corresponds to the *optimal sample size ratio.*

Example 3.4 (Covariate shift). *Our second example focuses on how varying covariate shifts impacts the variance limit in equation* (3.4)*. For large enough p,*

$$\operatorname{Tr}\left[\Sigma^{(2)}\hat{\Sigma}^{-1}\right] \to \frac{1}{n_1+n_2}\operatorname{Tr}\left[\Sigma^{(2)}(a_1\Sigma^{(1)} + a_2\Sigma^{(2)})^{-1}\right]$$
$$= \frac{1}{n_1+n_2}\operatorname{Tr}\left[(a_1 M^\top M + a_2 \operatorname{Id})^{-1}\right].$$

Hence the variance limit is determined by the spectrum of M. To illustrate the above result, suppose that half of M's singular values are equal to $\lambda > 1$ and the other half are equal to λ^{-1}. In particular, when $\lambda = 1$, there is no covariate shift. As λ increases, the severity of covariate shift increases. We observe the following dichotomy.

(i) If $n_1 \geqslant n_2$, then the variance limit is smallest when there is no covariate shift.

(ii) If $n_1 < n_2$, then the variance limit is largest when there is no covariate shift.

We explain why the dichotomy happens. The variance limit in this example is equal to $\frac{p}{2(n_1+n_2)} f(\lambda)$, where

$$f(\lambda) = (\lambda^{-2} a_1 + a_2)^{-1} + (\lambda^2 a_1 + a_2)^{-1}.$$

Using the fact that $a_1 + a_2 = 1 - \frac{p}{n_1+n_2}$, we can verify

$$f(\lambda) - f(1) = \left(2a_1 - \frac{n_1+n_2-p}{n_1+n_2}\right) g(\lambda, a_1),$$

where $g(\lambda, a_1) \geqslant 0$. We claim that $a_1 \geqslant \frac{n_1+n_2-p}{2(n_1+n_2)}$ if and only if $n_1 \geqslant n_2$, which explains the dichonomy. In fact, if $a_1 > a_2$, then equation (3.5) gives that $a_1 > \frac{n_1+n_2-p}{2(n_1+n_2)}$, and equation (3.6) gives that

$$\frac{n_1}{n_1+n_2} > a_1 + \frac{p}{2(n_1+n_2)}\left(\frac{\lambda^2}{\lambda^2+1} + \frac{\lambda^{-2}}{\lambda^{-2}+1}\right) > \frac{1}{2}.$$

This implies $n_1 > n_2$. The other direction follows from similar arguments.

Proof overview of Theorem 3.1. For the rest of this section, we present an overview of the proof of Theorem 3.1. The central quantity of interest is the inverse of the sum of two sample covariance matrices. We note that the variance equation $\operatorname{Tr}[\Sigma^{(2)}\hat{\Sigma}^{-1}]$ is equal to $(n_1+n_2)^{-1}\operatorname{Tr}\left[W^{-1}\right]$, where W is

$$\frac{\Lambda U^\top (Z^{(1)})^\top Z^{(1)} U \Lambda + V^\top (Z^{(2)})^\top Z^{(2)} V}{n_1+n_2}. \qquad (3.8)$$

Here $U\Lambda V^\top$ is defined as the SVD of M. This formulation is helpful because we know that $(Z^{(1)})^\top Z^{(1)}$ and $(Z^{(2)})^\top Z^{(2)}$ are both sample covariance matrices with isotropic population covariance, and U, V are both orthonormal matrices. For example, if $Z^{(1)}, Z^{(2)}$ are both Gaussian random matrices, by rotational invariance, $Z^{(1)}U, Z^{(2)}V$ are still Gaussian random matrices.

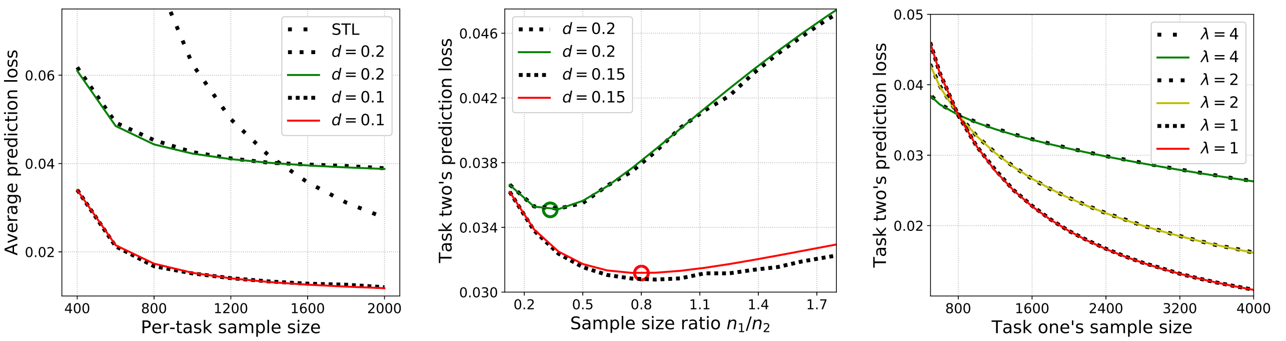
Our proof uses the Stieltjes transform or the resolvent method in random matrix theory. We briefly describe the key ideas and refer the interested readers to classical texts such as Bai and Silverstein (2010); Tao (2012); Erdos and Yau (2017). For any probability measure μ supported on $[0, \infty)$, the Stieltjes transform of μ is a complex function defined as

$$m_\mu(z) := \int_0^\infty \frac{\mathrm{d}\mu(x)}{x - z}, \text{ for any complex } z \in \mathbb{C} \setminus \{0\}.$$

Thus, the Stieltjes transform method reduces the study of a probability measure μ to the study of a complex function $m_\mu(z)$.

Let $\mu = p^{-1}\sum_i \delta_{\sigma_i}$ denote the empirical spectral distribution of W, where the σ_i's are the eigenvalues of W and δ_{σ_i} is the point mass measure at σ_i. Then it is easy to see that the Stieltjes transform of μ is equal to

$$m_\mu(z) := \frac{1}{p}\sum_{i=1}^{p} \frac{1}{\sigma_i - z} = p^{-1}\operatorname{Tr}\left[(W - z\operatorname{Id})^{-1}\right].$$



(a) Example 2.4

(b) Example 3.2

(c) Example 3.4

Figure 2: Our estimated losses (solid line) match the empirical losses (dotted line) accurately under various settings in dimension $p = 200$. **Left.** Validating Example 2.4 for ten tasks: the noise variance σ^2 is $1/4$. **Middle.** Validating Example 3.2 for two tasks: we discover an interesting phenomena by fixing task two's sample size and increasing task one's sample size. Moreover, our result accurately predicts the critical point (marked in circle) of the loss curve. **Right.** We show how different levels of covariate shift affect hard parameter sharing when there is no bias. Having covariate shift increases task two's prediction loss when task two's sample size is smaller than task one. Otherwise, having covariate shift (surprisingly) decreases task two's prediction loss.

The above matrix $(W - z\,\mathrm{Id})^{-1}$ is known as W's resolvent or Green's function. We prove the convergence of W's resolvent using the so-called "local law" with a sharp convergence rate (Bloemendal et al., 2014; Erdos and Yau, 2017; Knowles and Yin, 2016). The complete proof is provided in Section C.

4 Simulation Studies

We demonstrate the accuracy of our results in simulations. While our theory is asymptotic (with error terms that are negligible when p is sufficiently large), we observe that they are incredibly accurate in a moderate dimension of $p = 200$.

Sample efficiency. First, we validate the result of Example 2.4. Figure 2a shows the average prediction loss over ten tasks as we increase the number of samples per-task from 400 to 2000. In all the parameter settings, our results estimate the empirical losses accurately. We also observe a trend that the average prediction loss increases as we increase distance d from 0.1 to 0.2. Our work explains the differences between these two settings since $d^2 = 0.1^2$ is always smaller than $\frac{\sigma^2 p}{n-p}$, but $d^2 = 0.2^2$ is not. Indeed, we observe a crossover point between hard parameter sharing and STL. Finally, for $d = 0.2$, looking horizontally, we find that HPS requires fewer samples per-task than STL to achieve the same loss level.

Sample size ratio. Second, we validate the result of Example 3.2. Figure 2b shows task two's prediction loss as we increase the sample ratio n_1/n_2 from $1/10$ to $7/10$. We consider a regime where task two consists of $80,000$ samples, and task one's sample size varies from $8,000$ to $56,000$. The task-specific variance (which scales with model distance) is $d = 0.2$, the noise variance is $\sigma^2 = 4^2$, and the shared signal variance is 1. We observe that as we increase the sample ratio, task two's prediction loss decreases initially but later will increase when the sample ratio is above a certain level. On the other hand, when $d = 0.15$, task two's prediction loss decreases faster. Intuitively, this is because bias increases less for smaller d^2.

Covariate shift. Finally, we validate the result of Example 3.4. Figure 2c shows task two's prediction loss as we increase task one's sample size. Recall that λ measures the severity of covariate shifts—a larger λ means a larger covariate shift. We indeed observe the dichotomy in Example 3.4 at $n_1 = 800$. The sample size n_2 is 800 and the noise variance σ^2 is $1/4$.

5 Conclusions and Discussions

This work studied generalization properties of a widely used hard parameter sharing approach for multi-task learning. We provided sharp bias-variance tradeoffs of HPS in high-dimensional linear regression. Using these results, we analyzed how varying sample sizes and covariate shifts impact HPS, and rigorously explained several empirical phenomena such as negative transfer and covariate shift related to these dataset properties. We validated our theory and conducted further studies on text classification tasks. We describe open questions for future work. First, it would be interesting to tighten our estimate in Corollary 3.3, which would extend the observation in Figure 2b to small n_1. Second, it would be interesting to extend our result to classification problems such as logistic regression.

References

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.

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

Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 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, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

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.

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

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.

Edgar Dobriban and Yue Sheng. Wonder: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21 (66):1–52, 2020.

Edgar Dobriban, Stefan Wager, et al. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 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. Delocalization and diffusion profile for random band matrices. *Commun. Math. Phys.*, 323:367–416, 2013b.

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, 2013c.

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

Daniel J Hsu, Sham M Kakade, John Langford, and Tong Zhang. Multi-label prediction via compressed sensing. In *Advances in neural information processing systems*, pages 772–780, 2009.

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.

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, 2018.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *arXiv preprint arXiv:2006.10593*, 2020.

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.

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.

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.

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

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.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

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.

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 R. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*, 2020.

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

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.

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

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 Further Studies on Text Classification Tasks

Our results and simulations are all in the high-dimensional linear regression setting. How well do they extend to other scenarios? In this section, we conduct further studies on six text classification datasets. Our datasets include a movie review sentiment dataset (MR) (Pang and Lee, 2005), a sentence subjectivity dataset (SUBJ) (Pang and Lee, 2004), a customer reviews dataset (CR) (Hu and Liu, 2004), a question type dataset (TREC) (Li and Roth, 2002), an opinion polarity dataset (MPQA) (Wiebe et al., 2005), and the Stanford sentiment treebank (SST) dataset (Socher et al., 2013). Our model consists of a word embedding layer with GloVe embeddings (Pennington et al., 2014) followed by a long-short term memory (LSTM) or a multi-layer perception (MLP) layer (Lei et al., 2018).[3]

Sample size ratio. First, we show that our observation in Figure 2b also occurs in the text classification tasks. In Figure 3a, we observe that for multiple example task pairs, increasing task one's sample size improves task two's prediction accuracy initially, but hurts eventually. On the y-axis, we plot task two's test accuracy using HPS, subtracted by its STL test accuracy. We fix task two's sample size at 1000 and increase task one's sample size from 100 to 3000.

These examples and the one in Figure 2b suggest a natural progressive training schedule, where we add samples progressively until performance drops. Concretely, here is one implementation of this idea.

- We divide the training data into S batches. We divide the training procedure into S stages. During every stage, we progressively add one more data batch.
- During every stage, we train for T epochs using only the S batches. If the validation accuracy drops compared to the previous round's result or reach a desired threshold τ, we terminate.

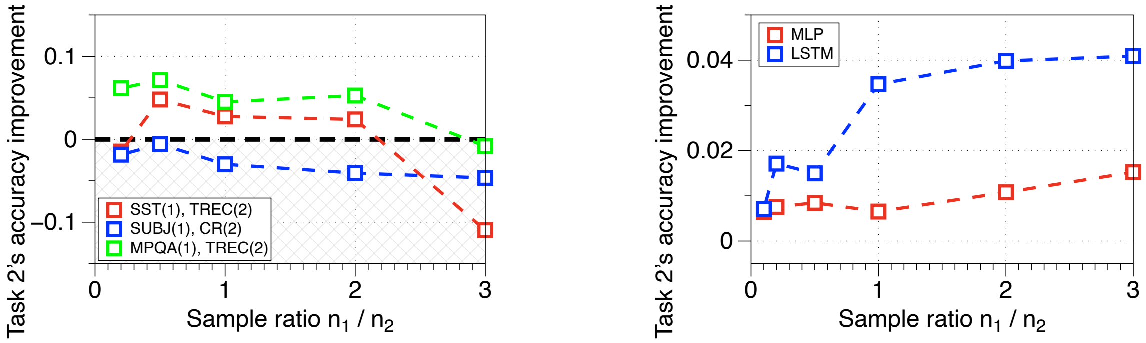
For example, if we apply this procedure to the settings of Figure 3a and 2b, it will terminate once reaching the optimal sample ratio. The advantage of this procedure is that it reduces the computational cost compared to standard round-robin training schedules. For example, if the procedure terminates at 30% of all batches, then SGD only passes over 30% of its data, whereas standard round-robin training passes over 100% of task one's data.

We evaluate the progressive training procedure on the six text classification datasets. First, we conduct multi-task training over all the 15 two-task pairs from the six datasets. We focus on task two's test accuracy and set τ as task two's test accuracy obtained via the standard round-robin training schedule. We include all of task two's data and progressively add task one's data using the procedure described above. Since the prediction accuracy has been controlled the same, we compare the computational cost. We find that when averaged over all the 15 two-task pairs, this procedure requires only 45% of the computational cost to reach the desired accuracy τ for task two. Second, we conduct multi-task training on all six datasets jointly. We extend our procedure to all six datasets. We include the data from all tasks except SST. For SST, we progressively add data similar to the above procedure. We set τ to be the average test accuracy of all the six tasks obtained using standard round-robin training. We find that adding samples progressively from SST requires less than 35% of the computational cost to reach the same average test accuracy τ.

Covariate shift. Recall from Example 3.4 that having covariate shifts worsens the variance (hence the loss) of hard parameter sharing when the sample ratio increases. This highlights the need for correcting covariate shifts when the sample size ratio rises. To this end, we study a covariance alignment procedure proposed in Wu et al. (2020), designed to correct covariate shifts. The idea is to add an alignment module between the input and the shared module B. This module is then trained together with B and the output layers. We refer to Wu et al. (2020) for more details about the procedure and the implementation.

We conduct multi-task training on all 15 task pairs from the six datasets. In Figure 3b, we measure the performance gains from performing covariance alignment vs. HPS. To get a robust comparison, we average the improvements over the 15 task pairs. The result shows that as the sample size ratio increases, performing covariance alignment provides more significant gains over HPS. We fix task two's sample size at $1,000$, and increase task one's sample size from $1,000$ to $3,000$.

[3] For MLP, we apply an average pooling layer over word embeddings. For LSTM, we add a shared feature representation layer on top of word embeddings.



(a) HPS vs. STL

(b) HPS vs. covariance alignment

Figure 3: Comparing hard parameter sharing (HPS) to single-task learning (STL) and a covariance alignment approach proposed by Wu et al. (2020): In Figure 3a, we observe that for multiple task pairs, increasing task one's sample size improves task two's prediction accuracy initially, but hurts eventually – a phenomenon similar to Figure 2b. In Figure 3b, we observe that as task one's sample size increases, covariance alignment improves more over HPS.

B Missing Proof of Theorem 2.1

We fill in missing details in the proof. Our first claim shows that the subspace spanned by the rows of $\hat{A}$ is close to that of $A^\star$.

Claim B.1. *Let $U_{\hat{A}} U_{\hat{A}}^\top \in \mathbb{R}^{t\times t}$ denote the subspace projection $\hat{A}^\top(\hat{A}\hat{A}^\top)^+\hat{A}$. In the setting of Theorem 2.1, we have that*

$$\left\|U_{\hat{A}}U_{\hat{A}}^\top - A^\star {A^\star}^\top\right\|_F^2 \leqslant n^{-c_\varphi}\cdot \frac{t(\left\|\Sigma^{1/2}B^\star\right\|^2+\sigma^2)}{\lambda_r({B^\star}^\top\Sigma B^\star)-\lambda_{r+1}({B^\star}^\top\Sigma B^\star)}.$$

The proof of the above claim is based on the following characterization.

Claim B.2. *In the setting of Theorem 2.1, we have that*

$$\mathop{\mathbb{E}}_{\{\varepsilon^{(j)}\}_{j=1}^t, X}[g(A)] = n\left\|\Sigma^{1/2}B^\star\left(A^\top(AA^\top)^+A - \mathrm{Id}_{t\times t}\right)\right\|_F^2 + \sigma^2(n\cdot t - p\cdot r). \tag{B.1}$$

As a result, the minimum of $\mathbb{E}\left[g(A)\right]$, denoted by $A^\star {A^\star}^\top$, is the best rank-r approximation of ${B^\star}^\top\Sigma B^\star$.

One can see that the expected optimization objective also admits a nice bias-variance decomposition. Furthermore, its minimum only depends on the bias term since the variance term is fixed, and the minimizer of the bias term is precisely $A^\star {A^\star}^\top$.

The next piece of our proof deals with the prediction loss of hard parameter sharing.

Claim B.3. *In the setting of Theorem 2.1, let $\hat{a}_i = \hat{A}^\top(\hat{A}\hat{A}^\top)^+\hat{A}_i$. We have that the prediction loss of $\hat{\beta}_i^{\mathrm{HPS}} := \hat{B}\hat{A}_i$ satisfies that*

$$\left|L(\hat{\beta}_i^{\mathrm{HPS}}) - L(B^\star\hat{a}_i) - \sigma^2\|\hat{a}_i\|^2\cdot\operatorname{Tr}\left[\Sigma(X^\top X)^{-1}\right]\right| \leqslant n^{-1/4}\left(L(B^\star\hat{a}_i) + \sigma^2\cdot\|\hat{a}_i\|^2\right).$$

Provided with these results, we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. Using Claim B.3, we get that the prediction loss of $\hat{\beta}_i^{\mathrm{HPS}}$ is equal to $L(B^\star\hat{a}_i) + \sigma^2\|\hat{a}_i\|^2\cdot\operatorname{Tr}\left[\Sigma(X^\top X)^{-1}\right]$ up to a multiplicative error of order $n^{-1/4}$. For the latter, we use Claim B.1 to upper bound the difference between $\|\hat{a}_i\|^2$ and $\|a_i^\star\|^2$. For $L(B^\star\hat{a}_i)$, we again use Claim B.1 to upper bound the distance between $\hat{a}_i$ and $a_i^\star$. Combined together, we obtain the difference if we replace $\hat{a}_i$ with $a_i^\star$ in Claim B.3, and the proof is complete. □

Next we present the proof of Claim B.1, Claim B.2, and Claim B.3.

Proof of Claim B.2. To facilitate the analysis, we consider the following matrix notations. Denote

$$\mathcal{E} := [\varepsilon^{(1)}, \varepsilon^{(2)}, \cdots, \varepsilon^{(t)}], \quad \text{and} \quad \mathcal{W} := X(X^\top X)^{-1} X^\top \mathcal{E} A^\top (AA^\top)^+.$$

For any $j = 1, 2, \dots, t$, let

$$H_j := B^\star A^\top (AA^\top)^+ A_j - \beta^{(j)}, \quad \text{and} \quad E_j := \mathcal{W} A_j - \varepsilon^{(j)}.$$

Then we can write the function $g(A)$ conveniently as

$$g(A) = \sum_{j=1}^{t} \|XH_j + E_j\|^2 .$$

We will divide $g(A)$ into three parts. For simplicity, we will use matrix notations in the proof, that is, stacking $[H_j]_j$ gives matrix $B^\star A^\top (AA^\top) A - B^\star$, and stacking $[E_j]_j$ gives $\mathcal{W}A - \mathcal{E}$.

Part 1: The first part is the square of XH_j,

$$\sum_{j=1}^{t} \|XH_j\|^2 = \left\| X(B^\star A^\top (AA^\top) A - B^\star) \right\|_F^2 = \left\| X(B^\star U_A U_A^\top - B^\star) \right\|_F^2 , \tag{B.2}$$

where $U_A U_A^\top \in \mathbb{R}^{t \times t}$ denotes the subspace projection $A^\top (AA^\top)^+ A$. Taking expectation of equation (B.2) over X, we get

$$\sum_{j=1}^{t} \|XH_j\|^2 = n \left\| \Sigma^{1/2} (B^\star U_A U_A^\top - B^\star) \right\|^2 .$$

Part 2: The second part is the cross term, which is equal to the following using the matrix notations:

$$\sum_{j=1}^{t} \langle XH_j, E_j \rangle = \langle X(B^\star U_A U_A^\top - B^\star), \mathcal{W}A - \mathcal{E} \rangle = -\langle X(B^\star U_A U_A^\top - B^\star), \mathcal{E} \rangle, \tag{B.3}$$

which is zero in expectation over $\mathcal{E}$.

Part 3: The last part is the square of E_j:

$$\sum_{j=1}^{t} \|E_j\|^2 = \|\mathcal{W}A - \mathcal{E}\|_F^2 = \|\mathcal{E}\|_F^2 - \langle \mathcal{W}A, \mathcal{E} \rangle, \tag{B.4}$$

where in the second step we use $\|\mathcal{W}A\|^2 = \langle \mathcal{W}A, \mathcal{E} \rangle$ by algebraic calculation. Hence, it suffices to show that the expectation of equation (B.4) is equal to $\sigma^2 (n \cdot t - p \cdot r)$. First, we have that $\mathbb{E}\left[\|\mathcal{E}\|_F^2\right] = \sigma^2 \cdot n \cdot t$. Second, we show that

$$\mathop{\mathbb{E}}_{\mathcal{E}} \left[\langle \mathcal{W}A, \mathcal{E} \rangle \right] = \mathop{\mathbb{E}}_{\mathcal{E}} \left[\mathrm{Tr}\left[\mathcal{E}^\top U_X U_X^\top \mathcal{E} U_A U_A^\top \right] \right] = p\sigma^2 \cdot \mathrm{Tr}\left[U_A U_A^\top \right] = p\sigma^2 \cdot r,$$

where $U_X U_X^\top = X(X^\top X)^{-1} X^\top$. The first step follows by applying the definition of $\mathcal{W}$. The last step is because $U_A U_A^\top$ has rank r. Hence, it suffices to show the second step is correct. For any $1 \leqslant i, j \leqslant t$, let $\delta_{i,j} = 1$ if $i = j$, and 0 otherwise. Because $\varepsilon^{(i)}$ and $\varepsilon^{(j)}$ are pairwise independent, we have that

$$\mathop{\mathbb{E}}_{\mathcal{E}} \left[(\mathcal{E}^\top U_X U_X^\top \mathcal{E})_{ij} \right] = \mathop{\mathbb{E}}_{\mathcal{E}} \left[{\varepsilon^{(i)}}^\top U_X U_X^\top \varepsilon^{(j)} \right] = \sigma^2 \cdot \mathrm{Tr}\left[U_X U_X^\top \right] \cdot \delta_{ij} = p\sigma^2 \cdot \delta_{ij}.$$

The last step uses the fact that $\mathrm{Tr}[U_X U_X^\top] = p$. Hence, the second step is correct.

Combining the three parts, the proof is complete. □

Proof of Claim B.1. Corresponding to the right-hand side of (B.1), we define the function

$$h(A) := n \left\| \Sigma^{1/2} B^\star \left(A^\top (AA^\top)^+ A - \mathrm{Id}_{t\times t} \right) \right\|_F^2 + \sigma^2 (n \cdot t - p \cdot r). \tag{B.5}$$

Let c be a fixed constant that is sufficiently small. Let c_∞ be any fixed value within $(0, 1/2 - c)$. To show that $U_{\hat{A}} U_{\hat{A}}^\top$ is close to $A^\star {A^\star}^\top$, we first show that $g(A)$ is close to $h(A)$ as follows:

$$|g(A) - h(A)| \lesssim n^{-c_\varphi} \cdot n \left\| \Sigma^{1/2} B^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}) \right\|_F^2 + n^{-c_\infty} \cdot \sigma^2 \cdot n \cdot t. \tag{B.6}$$

We consider the concentration error of each part of $g(A)$.

For equation (B.2), applying Corollary E.2 to $XH_j = Z\Sigma^{1/2} H_j$, we obtain that $\left\| Z\Sigma^{1/2} H_j \right\|^2 = n \|\Sigma^{1/2} H_j\|^2 \cdot (1 + \mathrm{O}(n^{-c_\varphi}))$ with high probability. This implies that

$$\left| \sum_{j=1}^t \|XH_j\|^2 - \sum_{j=1}^t n \|\Sigma^{1/2} H_j\|^2 \right| \lesssim n^{-c_\varphi} \cdot n \left\| \Sigma^{1/2} B^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}) \right\|^2. \tag{B.7}$$

For equation (B.3), using Corollary E.3, we obtain the following with high probability:

$$\begin{aligned} |\langle X B^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}), \mathcal{E} \rangle| &\leqslant n^c \cdot \sigma \cdot \left\| X B^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}) \right\|_F \\ &\leqslant n^c \cdot \sigma \cdot \|Z\| \cdot \left\| \Sigma^{1/2} B^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}) \right\|_F \\ &\lesssim n^{c+1/2} \cdot \sigma \cdot \left\| \Sigma^{1/2} B^\star (U_A U_A^\top - \mathrm{Id}_{t\times t}) \right\|_F \end{aligned} \tag{B.8}$$

In the second step, we use the fact that $X = Z\Sigma^{1/2}$. In the third step, we use Fact E.1(ii) to bound the operator norm of Z by $\mathrm{O}(\sqrt{n})$. By the AM-GM inequality, equation (B.8) is bounded by the right-hand side of (B.6).

For equation (B.4), using Corollary E.3, we obtain that with high probability,

$$\left| \|\mathcal{E}\|_F^2 - \sigma^2 \cdot n \cdot t \right| = \left| \mathrm{Tr}\left[\mathcal{E}^\top \, \mathrm{Id}_{n\times n} \, \mathcal{E} \right] - \sigma^2 \cdot n \cdot t \right| \leqslant n^c \cdot \sigma^2 \| \mathrm{Id}_{n\times n} \|_F = n^{1/2+c} \cdot \sigma^2. \tag{B.9}$$

For the inner product between $\mathcal{W}A$ and $\mathcal{E}$, we have that with high probability,

$$\begin{aligned} \left| \langle \mathcal{W}A, \mathcal{E} \rangle - \sigma^2 \cdot p \cdot r \right| &= \left| \mathrm{Tr}\left[\left(\mathcal{E}^\top U_X U_X^\top \mathcal{E} - p\sigma^2 \cdot \mathrm{Id}_{t\times t} \right) U_A U_A^\top \right] \right| \\ &\leqslant \left\| U_A U_A^\top \right\|_F \cdot \left\| \mathcal{E}^\top U_X U_X^\top \mathcal{E} - p\sigma^2 \cdot \mathrm{Id}_{t\times t} \right\| \\ &\leqslant \sqrt{r} \cdot n^c \cdot \sigma^2 \cdot \|U_X U_X^\top\|_F \\ &\leqslant \sqrt{r} \cdot n^{1/2+c} \cdot \sigma^2. \end{aligned} \tag{B.10}$$

Here in the third step, we apply equation (E.6) to $\left\| \mathcal{E}^\top U_X U_X^\top \mathcal{E} - p\sigma^2 \cdot \mathrm{Id}_{t\times t} \right\|$ and use that $\left\| U_A U_A^\top \right\|_F = \sqrt{r}$ because U_A has rank r. In the fourth step, we use $\|U_X U_X^\top\|_F = \sqrt{p}$ because U_X has rank p.

Combining the concentration error estimate for all three parts, we obtain equation (B.6).

Next, we use equation (B.6) to prove the claim. Using triangle inequality, we upper bound the gap between $h(A^\star)$ and $h(\hat{A})$:

$$\begin{aligned} h(\hat{A}) - h(A^\star) &\leqslant |g(A^\star) - h(A^\star)| + (g(\hat{A}) - g(A^\star)) + \left| g(\hat{A}) - h(\hat{A}) \right| \\ &\leqslant |g(A^\star) - h(A^\star)| + \left| g(\hat{A}) - h(\hat{A}) \right| \\ &\lesssim n^{-c_\varphi} \cdot n \left\| \Sigma^{1/2} B^\star \right\|_F^2 + n^{-c_\infty} \cdot \sigma^2 \cdot n \cdot t. \end{aligned} \tag{B.11}$$

The second step used the fact that $\hat{A}$ is the global minimizer of $g(\cdot)$, so that $g(\hat{A}) \leqslant g(A^\star)$. The third step used equation (B.6) and the fact that the spectral norm of $U_A U_A^\top - \mathrm{Id}_{t\times t}$ is at most one. Using equation (B.5), we can verify that

$$h(\hat{A}) - h(A^\star) = n \, \mathrm{Tr}\left[{B^\star}^\top \Sigma B^\star (A^\star {A^\star}^\top - U_{\hat{A}} U_{\hat{A}}^\top) \right].$$

Let $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_t$ be the eigenvalues of $B^{\star\top}\Sigma B^{\star}$. Let v_i be the corresponding eigenvector of λ_i. Then, we have $A^\star A^{\star\top} = \sum_{i=1}^r v_i v_i^\top$, and

$$\begin{aligned}
h(\hat A) - h(A^\star) = n\sum_{i=1}^r \lambda_i - n\sum_{i=1}^t \lambda_i \|U_{\hat A}^\top v_i\|^2 &= n\sum_{i=1}^r \lambda_i \left(1 - \|U_{\hat A}^\top v_i\|^2\right) - n\sum_{i=r+1}^t \lambda_i \|U_{\hat A}^\top v_i\|^2 \\
&\geqslant n(\lambda_r - \lambda_{r+1}) \sum_{i=r+1}^t \|U_{\hat A}^\top v_i\|^2, \qquad \text{(B.12)}
\end{aligned}$$

where we use $\sum_{i=1}^r \Big(1 - \|U_{\hat A}^\top v_i\|^2\Big) = r - \sum_{i=1}^r \|U_{\hat A}^\top v_i\|^2 = \sum_{i=r+1}^t \|U_{\hat A}^\top v_i\|^2$ in the last step. On the other hand, we have

$$\begin{aligned}
\|A^\star A^{\star\top} - U_{\hat A} U_{\hat A}^\top\|_F^2 &= 2r - 2\langle A^\star A^{\star\top}, U_{\hat A} U_{\hat A}{}^\top\rangle \\
&= 2\sum_{i=r+1}^t \|U_{\hat A}^\top v_i\|^2.
\end{aligned}$$

Thus from equation (B.11) and (B.12), we obtain that

$$\|A^\star A^{\star\top} - U_{\hat A} U_{\hat A}^\top\|_F^2 = 2\sum_{i=r+1}^t \|U_{\hat A}^\top v_i\|^2 \lesssim \frac{n^{-c_\varphi} \cdot \left\|\Sigma^{1/2} B^\star\right\|_F^2 + n^{-c_\infty} \cdot \sigma^2 t}{\lambda_r - \lambda_{r+1}}.$$

Hence the proof is complete. □

Proof of Claim B.3. The proof is similar to that of equation (B.6). The prediction loss of hard parameter sharing for task i is equal to

$$\begin{aligned}
L(\hat\beta_i^{\mathrm{HPS}}) &= \left\|\Sigma^{1/2}(\hat B \hat A_i - \beta^{(i)})\right\|^2 \\
&= \left\|\Sigma^{1/2}((X^\top X)^{-1} X^\top Y \hat A^\top (\hat A \hat A^\top)^{+} \hat A_i - \beta^{(i)})\right\|^2 \\
&= \left\|\Sigma^{1/2}(B^\star \hat a_i - \beta^{(i)} + R_i)\right\|^2,
\end{aligned}$$

where we denote $R_i = (X^\top X)^{-1} X^\top \mathcal{E} \hat a_i$. We divide the prediction loss into three parts.

Part 1: The first part is the bias term: $\|\Sigma^{1/2}(B^\star \hat a_i - \beta^{(i)})\|^2 = L(B^\star \hat a_i)$.

Part 2: The second part is the cross term, whose expectation over $\mathcal{E}$ is zero. Let $b = B^\star \hat a_i - \beta^{(i)}$ for simplicity. Using Corollary E.3, the concentration error can be bounded as

$$\begin{aligned}
\left|\langle \Sigma^{1/2} b, \Sigma^{1/2} R_i\rangle\right| &= \left|\langle X(X^\top X)^{-1}\Sigma b \hat a_i^\top, \mathcal{E}\rangle\right| \\
&\leqslant \sum_{j=1}^t |\hat a_i(j)| \cdot \left|\langle X(X^\top X)^{-1}\Sigma b, \varepsilon^{(j)}\rangle\right| \\
&\leqslant \sum_{j=1}^t |\hat a_i(j)| \cdot n^c \sigma \left\|X(X^\top X)^{-1}\Sigma b\right\| \\
&\leqslant \sqrt{t}\|\hat a_i\| \cdot n^c \sigma \left\|X(X^\top X)^{-1}\Sigma b\right\|.
\end{aligned}$$

In the first step, we plug in the definition of R_i and re-arrange terms. In the second step, we use $\hat a_i(j)$ to denote the j-th coordinate of $\hat a_i$. In the third step, we use equation (E.5). In the last step, we use $\sum_j |\hat a_i(j)| \leqslant \sqrt{t}\|\hat a_i\|$ by Cauchy-Schwarz inequality. Finally, we have

$$\begin{aligned}
\left\|X(X^\top X)^{-1}\Sigma b\right\|_F &= \left[b^\top \Sigma (X^\top X)^{-1} X^\top X (X^\top X)^{-1} \Sigma b\right]^{1/2} \\
&\leqslant \|\Sigma^{1/2} b\| \cdot \left\|\Sigma^{1/2}(X^\top X)^{-1}\Sigma^{1/2}\right\|^{1/2} = \|\Sigma^{1/2} b\| \cdot \left\|(Z^\top Z)^{-1}\right\|^{1/2} \\
&\leqslant n^{-1/2} \cdot \|\Sigma^{1/2} b\|.
\end{aligned}$$

Above, we use $X = Z\Sigma^{1/2}$. In the last step, we use Fact E.1(ii) to bound the operator norm of $Z^\top Z$ by $\mathrm{O}(n^{-1})$. One can see that the concentration error from this part is upper bounded by the result in Claim B.3.

Part 3: The final part is the squared term of R_i. We rewrite it as

$$\begin{aligned}\|\Sigma^{1/2} R_i\|^2 &= \left\|\sum_{j=1}^{t} \hat{a}_i(j)\Sigma^{1/2}(X^\top X)^{-1}X^\top \varepsilon^{(j)}\right\|^2 \\ &= \sum_{1\leqslant j,k\leqslant t} \hat{a}_i(j)\hat{a}_i(k){\varepsilon^{(j)}}^\top X(X^\top X)^{-1}\Sigma(X^\top X)^{-1}X^\top\varepsilon^{(k)}. \end{aligned} \tag{B.13}$$

First, for any $1 \leqslant j,k \leqslant t$, the expectation is

$$\mathop{\mathbb{E}}_{\mathcal{E}}\left[{\varepsilon^{(j)}}^\top X(X^\top X)^{-1}\Sigma(X^\top X)^{-1}X^\top\varepsilon^{(k)}\right] = \delta_{jk}\cdot\sigma^2\operatorname{Tr}\left[\Sigma(X^\top X)^{-1}\right].$$

Second, using equation (E.6), the concentration error is at most

$$\begin{aligned}&\left|{\varepsilon^{(j)}}^\top X(X^\top X)^{-1}\Sigma(X^\top X)^{-1}X^\top\varepsilon^{(k)} - \delta_{jk}\cdot\sigma^2\operatorname{Tr}\left[\Sigma(X^\top X)^{-1}\right]\right| \\ \leqslant & n^c\cdot\sigma^2\left\|X(X^\top X)^{-1}\Sigma(X^\top X)^{-1}X^\top\right\|_F = n^c\cdot\sigma^2\left\|\Sigma^{1/2}(X^\top X)^{-1}\Sigma^{1/2}\right\|_F \\ \leqslant & \sigma^2\cdot p^{1/2}\cdot\left\|(Z^\top Z)^{-1}\right\|^{1/2} \lesssim \sigma^2\cdot n^{-1/2+c}. \end{aligned} \tag{B.14}$$

Above, we used Fact E.1(ii) in the last step to bound the operator norm of $(Z^\top Z)^{-1}$. Plugging equation (B.14) into equation (B.13), we obtain that

$$\left|\left\|\Sigma^{1/2}R_i\right\|^2 - \sigma^2\|\hat{a}_i\|^2\cdot\operatorname{Tr}\left[\Sigma(X^\top X)^{-1}\right]\right| \lesssim \sigma^2\cdot n^{-1/2+c}\sum_{1\leqslant j,k\leqslant t}|\hat{a}_i(j)||\hat{a}_i(k)| = n^{-1/2+c}\sigma^2\cdot\|\hat{a}_i\|^2.$$

Finally, combining the three parts together, we complete the proof. □

C Proof of Theorem 3.1

We first state the asymptotic limit for the bias equation (3.2).

Theorem C.1. *Let S be an arbitrary subset of the unit sphere in dimension p whose size is polynomial in p. In the setting of Theorem 3.1, the bias equation (3.2) satisfies the following limit with high probability for any unit vector $w \in S$:*

$$\left|w^\top\Sigma^{(1)}\left(\hat{\Sigma}^{-1}\Sigma^{(2)}\hat{\Sigma}^{-1} - \frac{1}{(n_1+n_2)^2}{\Sigma^{(2)}}^{-1/2}V\frac{a_3\Lambda^2+(a_4+1)\operatorname{Id}}{(a_1\Lambda^2+a_2\operatorname{Id})^2}V^\top{\Sigma^{(2)}}^{-1/2}\right)\Sigma^{(1)}w\right| \leqslant \frac{p^{-c_\varphi}}{(n_1+n_2)^2}, \tag{C.1}$$

where a_3 and a_4 are the solutions of the following self-consistent equations

$$a_3 + a_4 = \frac{1}{n_1+n_2}\sum_{i=1}^{p}\frac{1}{\lambda_i^2 a_1 + a_2},\quad a_3 + \frac{1}{n_1+n_2}\sum_{i=1}^{p}\frac{\lambda_i^2(a_2a_3 - a_1a_4)}{(\lambda_i^2a_1+a_2)^2} = \frac{1}{n_1+n_2}\sum_{i=1}^{p}\frac{\lambda_i^2a_1}{(\lambda_i^2a_1+a_2)^2}. \tag{C.2}$$

Proof Overview (cont'd). We continue the proof overview of Theorem 3.1 and C.1 from Section 3. Recall that $(W - z\operatorname{Id})^{-1}$ is the resolvent of matrix W. We say that $(W - z\operatorname{Id})^{-1}$ converges to a deterministic $p \times p$ matrix limit $R(z)$ if for any sequence of deterministic unit vectors $v \in \mathbb{R}^p$,

$$v^\top\left[(W - z\operatorname{Id})^{-1} - R(z)\right]v \to 0 \quad \text{when } p \text{ goes to infinity.}$$

To study W's resolvent, we observe that W is equal to $FF^\top$ for a p by n_1+n_2 matrix

$$F := (n_1+n_2)^{-1/2}[\Lambda U^\top(Z^{(1)})^\top, V^\top(Z^{(2)})^\top]. \tag{C.3}$$

Consider the following symmetric block matrix whose dimension is $p + n_1 + n_2$

$$H := \begin{pmatrix} 0 & F \\ F^\top & 0 \end{pmatrix}. \tag{C.4}$$

For this block matrix, we define its resolvent as

$$G(z) := \left[H - \begin{pmatrix} z \operatorname{Id}_{p\times p} & 0 \\ 0 & \operatorname{Id}_{(n_1+n_2)\times(n_1+n_2)} \end{pmatrix} \right]^{-1},$$

for any complex value $z \in \mathbb{C}$. Using Schur complement formula for the inverse of a block matrix, it is not hard to verify that

$$G(z) = \begin{pmatrix} (W - z\operatorname{Id})^{-1} & (W - z\operatorname{Id})^{-1}F \\ F^\top (W - z\operatorname{Id})^{-1} & z(F^\top F - z\operatorname{Id})^{-1} \end{pmatrix}. \tag{C.5}$$

Variance asymptotic limit. In Theorem C.7, we will show that for z in a small neighborhood around 0, when p goes to infinity, $G(z)$ converges to the following limit

$$\mathfrak{G}(z) := \begin{pmatrix} (a_1(z)\Lambda^2 + (a_2(z) - z)\operatorname{Id}_{p\times p})^{-1} & 0 & 0 \\ 0 & -\frac{n_1+n_2}{n_1} a_1(z)\operatorname{Id}_{n_1\times n_1} & 0 \\ 0 & 0 & -\frac{n_1+n_2}{n_2} a_2(z)\operatorname{Id}_{n_2\times n_2} \end{pmatrix}, \tag{C.6}$$

where $a_1(z)$ and $a_2(z)$ are the unique solutions to the following self-consistent equations

$$\begin{aligned} a_1(z) + a_2(z) &= 1 - \frac{1}{n_1+n_2}\left(\sum_{i=1}^p \frac{\lambda_i^2 a_1(z) + a_2(z)}{\lambda_i^2 a_1(z) + a_2(z) - z} \right), \\ a_1(z) + \frac{1}{n_1+n_2}\left(\sum_{i=1}^p \frac{\lambda_i^2 a_1(z)}{\lambda_i^2 a_1(z) + a_2(z) - z} \right) &= \frac{n_1}{n_1+n_2}. \end{aligned} \tag{C.7}$$

The existence and uniqueness of solutions to the above system are shown in Lemma C.10. Given this result, we now show that when $z = 0$, the matrix limit $\mathfrak{G}(0)$ implies the variance limit shown in equation (3.4). First, we have that $a_1 = a_1(0)$ and $a_2 = a_2(0)$ since the equations in (C.7) reduce to equations (3.5) and (3.6) when $z = 0$. Second, since W^{-1} is the upper-left block matrix of $G(0)$, we have that W^{-1} converges to $(a_1\Lambda^2 + a_2\operatorname{Id})^{-1}$. Using the fact that $\operatorname{Tr}[\Sigma^{(2)}\hat{\Sigma}^{-1}] = (n_1+n_2)^{-1}\operatorname{Tr}\left[W^{-1}\right]$, we get that when p goes to infinity,

$$\begin{aligned} \operatorname{Tr}\left[\Sigma^{(2)}\hat{\Sigma}\right] \to \frac{1}{n_1+n_2}\operatorname{Tr}\left[(a_1\Lambda^2 + a_2\operatorname{Id})^{-1}\right] &= \frac{1}{n_1+n_2}\operatorname{Tr}\left[(a_1 M^\top M + a_2\operatorname{Id})^{-1}\right] \\ &= \frac{1}{n_1+n_2}\operatorname{Tr}\left[\Sigma^{(2)}(a_1\Sigma^{(1)} + a_2\Sigma^{(2)})^{-1}\right], \end{aligned}$$

where we note that $M^\top M = (\Sigma^{(2)})^{-1/2}\Sigma^{(1)}(\Sigma^{(2)})^{-1/2}$ and its SVD is equal to $V^\top\Lambda^2 V$.

Bias asymptotic limit. For the bias limit in equation (C.1), we show that it is governed by the derivative of $(W - z\operatorname{Id})^2$ with respect to z at $z = 0$. First, we can express the empirical bias term in equation (C.1) as

$$(n_1+n_2)^2\hat{\Sigma}^{-1}\Sigma^{(2)}\hat{\Sigma}^{-1} = {\Sigma^{(2)}}^{-1/2} V W^{-2} V^\top {\Sigma^{(2)}}^{-1/2}. \tag{C.8}$$

Let $\mathcal{G}(z) := (W - z\operatorname{Id})^{-1}$ denote the resolvent of W. Our key observation is that $\frac{\mathrm{d}\mathcal{G}(z)}{\mathrm{d}z} = \mathcal{G}^2(z)$. Hence, provided that the limit of $(W - z\operatorname{Id})^{-1}$ is $(a_1(z)\Lambda^2 + (a_2(z) - z)\operatorname{Id})^{-1}$ near $z = 0$, the limit of $\frac{\mathrm{d}\mathcal{G}(0)}{\mathrm{d}z}$ satisfies that

$$\frac{\mathrm{d}\mathcal{G}(0)}{\mathrm{d}z} \to \frac{-\frac{\mathrm{d}a_1(0)}{\mathrm{d}z}\Lambda^2 - (\frac{\mathrm{d}a_2(0)}{\mathrm{d}z} - 1)\operatorname{Id}}{(a_1(0)\Lambda^2 + a_2(0)\operatorname{Id}_p)^2}. \tag{C.9}$$

To find the derivatives of $a_1(z)$ and $a_2(z)$, we take the derivatives on both sides of the system of equations (C.7). Let $a_3 = -\frac{\mathrm{d}a_1(0)}{\mathrm{d}z}$ and $a_4 = -\frac{\mathrm{d}a_2(0)}{\mathrm{d}z}$. One can verify that a_3 and a_4 satisfy the self-consistent equations in (C.2) (details omitted). Applying equation (C.9) to equation (C.8), we obtain the asymptotic limit of the bias term.

As a remark, in order for $\frac{\mathrm{d}\mathcal{G}(z)}{\mathrm{d}z}$ to stay close to its limit at $z = 0$, we not only need to find the limit of $\mathcal{G}(0)$, but also the limit of $\mathcal{G}(z)$ within a small neighborhood of 0. This is why we consider W's resolvent for a general z.

How to derive the matrix limit? We begin with a warm up analysis when the entries of $Z^{(1)}$ and $Z^{(2)}$ are drawn i.i.d. from an isotropic Gaussian distribution. By the rotational invariance of the multivariate Gaussian distribution, we have that the entries of $Z^{(1)}U$ and $Z^{(2)}V$ also follow an isotropic Gaussian distribution. Hence it suffices to consider the following resolvent

$$G(z) = \begin{pmatrix} -z\,\mathrm{Id}_{p\times p} & (n_1+n_2)^{-1/2}\Lambda(Z^{(1)})^\top & (n_1+n_2)^{-1/2}(Z^{(2)})^\top \\ (n_1+n_2)^{-1/2}Z^{(1)}\Lambda & -\mathrm{Id}_{n_1\times n_1} & 0 \\ (n_1+n_2)^{-1/2}Z^{(2)} & 0 & -\mathrm{Id}_{n_2\times n_2} \end{pmatrix}^{-1}. \tag{C.10}$$

We show how to derive the matrix limit $\mathfrak{G}(z)$ and the self-consistent equation system (C.7). We first introduce several useful notations. We define $n := n_1 + n_2$ and the following index sets

$$\mathcal{I}_0 := [\![1, p]\!], \quad \mathcal{I}_1 := [\![p+1, p+n_1]\!], \quad \mathcal{I}_2 := [\![p+n_1+1, p+n_1+n_2]\!], \quad \mathcal{I} := \mathcal{I}_0 \cup \mathcal{I}_1 \cup \mathcal{I}_2.$$

We will study the following partial traces of the resolve $G(z)$:

$$\begin{aligned} m(z) &:= \frac{1}{p}\sum_{i\in\mathcal{I}_0} G_{ii}(z), \quad m_0(z) := \frac{1}{p}\sum_{i\in\mathcal{I}_0} \lambda_i^2 G_{ii}(z), \\ m_1(z) &:= \frac{1}{n_1}\sum_{\mu\in\mathcal{I}_1} G_{\mu\mu}(z), \quad m_2(z) := \frac{1}{n_2}\sum_{\nu\in\mathcal{I}_2} G_{\nu\nu}(z). \end{aligned} \tag{C.11}$$

To deal with the matrix inverse, we consider the following resolvent minors of $G(z)$.

Definition C.2 (Resolvent minors). Let $X \in \mathbb{R}^{(p+n_1+n_2)\times(p+n_1+n_2)}$ and $i = 1, 2, \ldots, p+n_1+n_2$. The minor of X after removing the i-th row and column of X is denoted by $X^{(i)} := [X_{a_1a_2} : a_1, a_2 \in \mathcal{I} \setminus \{i\}]$ as a square matrix with dimension $p+n_1+n_2-1$. For the indices of $X^{(i)}$, we use $X^{(i)}_{a_1a_2}$ to denote $X_{a_1a_2}$ when a_1 and a_2 are both not equal to i, and $X^{(i)}_{a_1a_2} = 0$ when $a_1 = i$ or $a_2 = i$. The resolvent minor of $G(z)$ after removing the i-th row and column is defined as

$$G^{(i)}(z) := \left[\begin{pmatrix} -z\,\mathrm{Id}_{p\times p} & n^{-1/2}\Lambda(Z^{(1)})^\top & n^{-1/2}(Z^{(2)})^\top \\ n^{-1/2}Z^{(1)}\Lambda & -\mathrm{Id}_{n_1\times n_1} & 0 \\ n^{-1/2}Z^{(2)} & 0 & -\mathrm{Id}_{n_2\times n_2} \end{pmatrix}^{(i)}\right]^{-1}.$$

As a remark, we define the partial traces $m^{(i)}(z)$, $m_0^{(i)}(z)$, $m_1^{(i)}(z)$, and $m_2^{(i)}(z)$ by replacing $G(z)$ with $G^{(i)}(z)$ in equation (C.11).

Self-consistent equations. We briefly describe the ideas for deriving the system of self-consistent equations (C.7). A complete proof can be found in Lemma C.17. We show that with high probability, the following equations hold approximately:

$$\begin{aligned} m_1^{-1}(z) &= -1 + \frac{1}{n}\sum_{i=1}^{p} \frac{\lambda_i^2}{z + \lambda_i^2 \frac{n_1}{n_1+n_2} m_1(z) + \frac{n_2}{n_1+n_2} m_2(z) + \mathrm{o}(1)} + \mathrm{o}(1), \\ m_2^{-1}(z) &= -1 + \frac{1}{n}\sum_{i=1}^{p} \frac{1}{z + \lambda_i^2 \frac{n_1}{n_1+n_2} m_1(z) + \frac{n_2}{n_1+n_2} m_2(z) + \mathrm{o}(1)} + \mathrm{o}(1). \end{aligned} \tag{C.12}$$

With algebraic calculations, it is not hard to verify that these equations reduce to the self-consistent equations that we stated in equation (C.7) up to a small error o(1). More precisely, we have that $m_1(z)$ is approximately equal to $-\frac{n_1+n_2}{n_1}a_1(z)$ and $m_2(z)$ is approximately equal to $-\frac{n_1+n_2}{n_2}a_2(z)$.

The core idea is to study $G(z)$ using the Schur complement formula. First, we consider the diagonal entries of

$G(z)$ for each block in $\mathcal{I}_0$, $\mathcal{I}_1$, and $\mathcal{I}_2$. For any i in $\mathcal{I}_0$, any μ in $\mathcal{I}_1$, and any ν in $\mathcal{I}_2$, we have that

$$\begin{aligned}
G_{ii}^{-1}(z) &= -z - \frac{\lambda_i^2}{n} \sum_{\mu,\nu\in\mathcal{I}_1} Z_{\mu i}^{(1)} Z_{\nu i}^{(1)} G_{\mu\nu}^{(i)}(z) - \frac{1}{n} \sum_{\mu,\nu\in\mathcal{I}_2} Z_{\mu i}^{(2)} Z_{\nu i}^{(2)} G_{\mu\nu}^{(i)}(z) - \frac{2\lambda_i}{n} \sum_{\mu\in\mathcal{I}_1,\nu\in\mathcal{I}_2} Z_{\mu i}^{(1)} Z_{\nu i}^{(2)} G_{\mu\nu}^{(i)}(z) \\
G_{\mu\mu}^{-1}(z) &= -1 - \frac{1}{n} \sum_{i,j\in\mathcal{I}_0} \lambda_i\lambda_j Z_{\mu i}^{(1)} Z_{\mu j}^{(1)} G_{ij}^{(\mu)}(z) \\
G_{\nu\nu}^{-1}(z) &= -1 - \frac{1}{n} \sum_{i,j\in\mathcal{I}_0} Z_{\nu i}^{(2)} Z_{\nu j}^{(2)} G_{ij}^{(\nu)}(z).
\end{aligned}$$

For the first equation, we expand the Schur complement formula $G_{ii}^{-1}(z) = -z - H_i G^{(i)}(z) H_i^\top$, where H_i is the i-th row of H with the (i,i)-th entry removed. The second and third equations follow by similar calculations.

Next, we apply standard concentration bounds to simplify the above results. For $G_{ii}^{-1}(z)$, recall that the resolvent minor $G^{(i)}$ is defined such that it is independent of the i-th row and column of $Z^{(1)}$ and $Z^{(2)}$. Hence by standard concentration inequalities, we have that the cross terms are approximately zero. As shown in Lemma C.17, we have that with high probability the following holds

$$\begin{aligned}
G_{ii}^{-1}(z) &= -z - \frac{\lambda_i^2}{n} \sum_{\mu\in\mathcal{I}_1} G_{\mu\mu}^{(i)} - \frac{1}{n} \sum_{\mu\in\mathcal{I}_2} G_{\mu\mu}^{(i)} + \mathrm{o}(1) \\
&= -z - \frac{\lambda_i^2 \cdot n_1}{n_1+n_2} m_1^{(i)}(z) - \frac{n_2}{n_1+n_2} m_2^{(i)}(z) + \mathrm{o}(1),
\end{aligned}$$

by our definition of the partial traces $m_1^{(i)}(z)$ and $m_2^{(i)}(z)$. Since we have removed only one column and one row from $H(z)$, $m_1^{(i)}(z)$ and $m_2^{(i)}(z)$ should be approximately equal to $m_1(z)$ and $m_2(z)$. Hence we obtain that

$$G_{ii}(z) = -\left(z + \frac{\lambda_i^2 \cdot n_1}{n_1+n_2} m_1(z) + \frac{n_2}{n_1+n_2} m_2(z) + \mathrm{o}(1)\right)^{-1}. \tag{C.13}$$

For the other two blocks $\mathcal{I}_1$ and $\mathcal{I}_2$, using similar ideas we obtain the following equations with high probability:

$$G_{\mu\mu}(z) = -\left(1 + \frac{p}{n_1+n_2} m_0(z) + \mathrm{o}(1)\right)^{-1}, \quad G_{\nu\nu}(z) = -\left(1 + \frac{p}{n_1+n_2} m(z) + \mathrm{o}(1)\right)^{-1}.$$

By averaging the above results over $\mu \in \mathcal{I}_1$ and $\nu \in \mathcal{I}_2$, we obtain that with high probability

$$\begin{aligned}
m_1(z) &= \frac{1}{n_1} \sum_{\mu\in\mathcal{I}_1} G_{\mu\mu}(z) = -\left(1 + \frac{p}{n_1+n_2} m_0(z) + \mathrm{o}(1)\right)^{-1}, \\
m_2(z) &= \frac{1}{n_2} \sum_{\nu\in\mathcal{I}_2} G_{\nu\nu}(z) = -\left(1 + \frac{p}{n_1+n_2} m(z) + \mathrm{o}(1)\right)^{-1}.
\end{aligned}$$

Furthermore, we obtain that for $\mu \in \mathcal{I}_1$ and $\nu \in \mathcal{I}_2$, with high probability $G_{\mu\mu}(z) = m_1(z) + \mathrm{o}(1)$ and $G_{\nu\nu}(z) = m_2 + \mathrm{o}(1)$. In other words, both block matrices within $\mathcal{I}_1$ and $\mathcal{I}_2$ are approximately a scaling of the identity matrix. The above results for $m_1(z)$ and $m_2(z)$ imply that

$$m_1^{-1}(z) = -1 - \frac{1}{n} \sum_{i=1}^{p} \lambda_i^2 G_{ii}(z) + \mathrm{o}(1), \quad m_2^{-1}(z) = -1 - \frac{1}{n} \sum_{i=1}^{p} G_{ii}(z) + \mathrm{o}(1).$$

where we used the definitions of $m(z)$ and $m_0(z)$. By applying equation (C.13) for $G_{ii}(z)$ to these two equations, we obtain the system of self-consistent equations (C.12). In Lemma C.11, we show that the self-consistent equations are stable, that is, a small perturbation of the equations leads to a small perturbation of the solution.

Matrix limit. Finally, we derive the matrix limit $\mathfrak{G}(z)$. We have shown that $m_1(z)$ is approximately equal to $-\frac{n_1+n_2}{n_1} a_1(z)$ and $m_2(z)$ is approximately equal to $-\frac{n_1+n_2}{n_2} a_2(z)$ because we know that (C.12) holds. Inserting $m_1(z)$ and $m_2(z)$ into equation (C.13), we get that for i in $\mathcal{I}_0$, $G_{ii}(z) = (-z + \lambda_i^2 a_1(z) + a_2(z) + \mathrm{o}(1))^{-1}$ with high probability. For μ in $\mathcal{I}_1$ and ν in $\mathcal{I}_2$, by $G_{\mu\mu}(z) = m_1(z) + \mathrm{o}(1)$ and $G_{\nu\nu}(z) = m_2 + \mathrm{o}(1)$, we have that $G_{\mu\mu}(z) = -\frac{n_1+n_2}{n_1} a_1(z) + \mathrm{o}(1)$ and $G_{\nu\nu}(z) = -\frac{n_1+n_2}{n_2} a_2(z) + \mathrm{o}(1)$ with high probability. Hence we have derived the diagonal entries of $\mathfrak{G}(z)$. In Lemma C.16, we show that the off-diagonal entries are close to zero. For example, for $i \neq j \in \mathcal{I}_0$, by Schur complement, we have that

$$G_{ij}(z) = -G_{ii}(z) \cdot n^{-1/2} \Big(\lambda_i \sum_{\mu \in \mathcal{I}_1} Z^{(1)}_{\mu i} G^{(i)}_{\mu j}(z) + \sum_{\mu \in \mathcal{I}_2} Z^{(2)}_{\mu i} G^{(i)}_{\mu j}(z) \Big).$$

Using standard concentration inequalities, we can show that $\sum_{\mu \in \mathcal{I}_1} Z^{(1)}_{\mu i} G^{(i)}_{\mu j}(z)$ and $\sum_{\mu \in \mathcal{I}_2} Z^{(2)}_{\mu i} G^{(i)}_{\mu j}(z)$ are both close to zero. The other off-diagonal entries are bounded similarly.

Notations. We introduce several useful notations for the proof of Theorem 3.1. We say that an event Ξ holds with overwhelming probability if for any constant $D > 0$, $\mathbb{P}(\Xi) \geqslant 1 - n^{-D}$ for large enough n. Moreover, we say Ξ holds with overwhelming probability in an event Ω if for any constant $D > 0$, $\mathbb{P}(\Omega \setminus \Xi) \leqslant n^{-D}$ for large enough n. The following notion of stochastic domination, which was first introduced in Erdős et al. (2013a), is commonly used in the study of random matrices.

Definition C.3 (Stochastic domination). Let $\xi \equiv \xi^{(n)}$ and $\zeta \equiv \zeta^{(n)}$ be two n-dependent random variables. We say that ξ is stochastically dominated by ζ, denoted by $\xi \prec \zeta$ or $\xi = \mathrm{O}_\prec(\zeta)$, if for any small constant $c > 0$ and any large constant $D > 0$, there exists a function $n_0(c, D)$ such that for all $n > n_0(c, D)$,

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

In case $\xi(u)$ and $\zeta(u)$ is a function of u supported in $\mathcal{U}$, then we say $\xi(u)$ is stochastically dominated by $\zeta(u)$ uniformly in $\mathcal{U}$ if

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

We make several remarks. First, since we allow an n^c factor in stochastic domination, we can ignore log factors without loss of generality since $(\log n)^C \prec 1$ for any constant $C > 0$. Second, given a random variable ξ whose moments exist up to any order, we have that $|\xi| \prec 1$. This is because by Markov's inequality, let k be larger than D/c, then we have that

$$\mathbb{P}(|\xi| \geqslant n^c) \leqslant n^{-kc} \mathbb{E}|\xi|^k \leqslant n^{-D}.$$

As a special case, this implies that a Gaussian random variable ξ with unit variance satisfies that $|\xi| \prec 1$.

The following fact collects several basic properties that are often used in the proof. Roughly speaking, it shows that the stochastic domination "$\prec$" behaves in the same way as "$<$" in some sense.

Fact C.4 (Lemma 3.2 in Bloemendal et al. (2014)). Let ξ and ζ be two families of nonnegative random variables depending on some parameters $u \in \mathcal{U}$ or $v \in \mathcal{V}$.

(i) Suppose that $\xi(u, v) \prec \zeta(u, v)$ uniformly in $u \in \mathcal{U}$ and $v \in \mathcal{V}$. If $|\mathcal{V}| \leqslant n^C$ for some constant $C > 0$, then $\sum_{v \in \mathcal{V}} \xi(u, v) \prec \sum_{v \in \mathcal{V}} \zeta(u, v)$ uniformly in u.

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

(iii) Suppose that $\Psi(u) \geqslant n^{-C}$ is a family of deterministic parameters, 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.

Next, we introduce the bounded support assumption for a random matrix. We say that a random matrix $Z \in \mathbb{R}^{n \times p}$ satisfies the *bounded support condition* with Q or Z has support Q if

$$\max_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p} |Z_{ij}| \prec Q. \tag{C.14}$$

As shown in the example above, if the entries of Z have finite moments up to any order, then Z has bounded support 1. More generally, if the entries of Z have finite φ-th moment, then using Markov's inequality and a simple union bound we get that

$$\mathbb{P}\left(\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}|Z_{ij}| \geqslant (\log n)n^{\frac{2}{\varphi}}\right) \leqslant \sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{P}\left(|Z_{ij}| \geqslant (\log n)n^{\frac{2}{\varphi}}\right)$$
$$\leqslant \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{C(\varphi)}{\left[(\log n)n^{\frac{2}{\varphi}}\right]^{\varphi}} = \mathrm{O}((\log n)^{-\varphi}). \tag{C.15}$$

In other words, Z has bounded support $Q = n^{\frac{2}{\varphi}}$ with high probability.

The following resolvent identities are important tools for our proof. Recall that the resolvent minors have been defined in Definition C.2, and matrix F is given in equation (C.59).

Lemma C.5. *We have the following resolvent identities.*

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

$$\frac{1}{G_{ii}} = -z - \left(FG^{(i)}F^\top\right)_{ii}, \quad \frac{1}{G_{\mu\mu}} = -1 - \left(F^\top G^{(\mu)}F\right)_{\mu\mu}. \tag{C.16}$$

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

$$G_{ia} = -G_{ii}\left(FG^{(i)}\right)_{ia}, \quad G_{\mu b} = -G_{\mu\mu}\left(F^\top G^{(\mu)}\right)_{\mu b}. \tag{C.17}$$

(iii) For $a \in \mathcal{I}$ and $a_1, a_2 \in \mathcal{I}\setminus\{a\}$, we have

$$G^{(a)}_{a_1a_2} = G_{a_1a_2} - \frac{G_{a_1a}G_{aa_2}}{G_{aa}}. \tag{C.18}$$

The above result can be proved using Schur's complement formula, cf. Knowles and Yin (2016, Lemma 4.4).

The following lemma gives sharp concentration bounds for linear and quadratic forms of bounded random variables. We recall that the stochastic domination "$\prec$" has been defined in Definition C.3.

Lemma C.6 (Lemma 3.8 of (Erdős et al., 2013c) and Theorem B.1 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 and y_j have variance at most 1, and satisfy the bounded support condition (C.14) for a deterministic parameter Q. Then we have the following results:*

$$\Big|\sum_{i=1}^{n}A_ix_i\Big| \prec Q\max_i|A_i| + \Big(\sum_i|A_i|^2\Big)^{1/2}, \quad \Big|\sum_{i,j=1}^{n}x_iB_{ij}y_j\Big| \prec Q^2B_d + Qn^{1/2}B_o + \Big(\sum_{i\neq j}|B_{ij}|^2\Big)^{1/2}, \tag{C.19}$$

$$\Big|\sum_{i=1}^{n}(|x_i|^2 - \mathbb{E}|x_i|^2)B_{ii}\Big| \prec Qn^{1/2}B_d, \quad \Big|\sum_{1\leqslant i\neq j\leqslant n}\bar{x}_iB_{ij}x_j\Big| \prec Qn^{1/2}B_o + \Big(\sum_{i\neq j}|B_{ij}|^2\Big)^{1/2}, \tag{C.20}$$

where we denote $B_d := \max_i|B_{ii}|$ and $B_o := \max_{i\neq j}|B_{ij}|$. Moreover, if the moments of x_i and y_j exist up to any order, then we have the following stronger results:

$$\Big|\sum_i A_ix_i\Big| \prec \Big(\sum_i|A_i|^2\Big)^{1/2}, \quad \Big|\sum_{i,j}x_iB_{ij}y_j\Big| \prec \Big(\sum_{i,j}|B_{ij}|^2\Big)^{1/2}, \tag{C.21}$$

$$\Big|\sum_i(|x_i|^2 - \mathbb{E}|x_i|^2)B_{ii}\Big| \prec \Big(\sum_i|B_{ii}|^2\Big)^{1/2}, \quad \Big|\sum_{i\neq j}\bar{x}_iB_{ij}x_j\Big| \prec \Big(\sum_{i\neq j}|B_{ij}|^2\Big)^{1/2}. \tag{C.22}$$

C.1 Limit of the Resolvent

We now state the main random matrix result—Theorem C.7—which gives an almost optimal estimate on the resolvent $G(z)$ of H. It is conventionally called the *anisotropic local law* (Knowles and Yin, 2016). We define a domain of the spectral parameter z as

$$\mathbf{D} := \left\{z = E + \mathrm{i}\eta \in \mathbb{C}_+ : |z| \leqslant (\log n)^{-1}\right\}. \tag{C.23}$$

Theorem C.7. *In the setting of Theorem 3.1, let q be equal to $n^{-\frac{\varphi-4}{2\varphi}}$. We have that the resolvent $G(z)$ converges to the matrix limit $\mathfrak{G}(z)$: for any deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{p+n_1+n_2}$, the following estimate*

$$\max_{z\in\mathbf{D}} \left|\mathbf{u}^\top (G(z) - \mathfrak{G}(z))\mathbf{v}\right| \prec q \tag{C.24}$$

holds on the high probability event

$$\left\{ \max_{1\leqslant i\leqslant n_1, 1\leqslant j\leqslant p} |Z^{(1)}_{ij}| \leqslant (\log n) n^{\frac{2}{\varphi}}, \max_{1\leqslant i\leqslant n_2, 1\leqslant j\leqslant p} |Z^{(2)}_{ij}| \leqslant (\log n) n^{\frac{2}{\varphi}} \right\}. \tag{C.25}$$

The above result can be derived using the following lemma, which holds under a more general bounded support assumption on the random matrices.

Lemma C.8. *In the setting of Theorem C.7, assume that $Z^{(1)}$ and $Z^{(2)}$ satisfy the bounded support condition* (C.14) *with $Q = \sqrt{n}q = n^{\frac{2}{\varphi}}$. Then we have that the anisotropic local law in equation* (C.24) *holds for any deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{p+n_1+n_2}$.*

Remark C.9. The reason why we say the bounded support assumption is more general is because it provides greater flexibility in dealing with bounded moments. For example, we can also replace equation (C.15) with

$$\mathbb{P}\left(\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p} |Z_{ij}| \geqslant n^{\frac{2}{\varphi}+\delta} \right) = \mathrm{O}(n^{-\varphi\delta})$$

for a small constant $\delta > 0$. Hence we can replace event (C.25) with

$$\left\{ \max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p} |Z^{(1)}_{ij}| \leqslant n^{\frac{2}{\varphi}+\delta}, \max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p} |Z^{(2)}_{ij}| \leqslant n^{\frac{2}{\varphi}+\delta} \right\},$$

which holds with higher probability. But on this event we need to take a larger $q = n^{-\frac{\varphi-4}{2\varphi}+\delta}$, which means a worse convergence rate. In general, with Lemma C.8 one can determine the most suitable trade-off between probability and convergence rate depending on one's need.

Using the above result, we prove Theorem C.7 using a simple cutoff argument.

Proof of Theorem C.7. We introduce the truncated matrices $\widetilde{Z}^{(1)}$ and $\widetilde{Z}^{(2)}$ with entries

$$\widetilde{Z}^{(1)}_{\mu i} := \mathbf{1}\left(n^{-1/2}|Z^{(1)}_{\mu i}| \leqslant q\log n \right) \cdot Z^{(1)}_{\mu i}, \quad \widetilde{Z}^{(2)}_{\nu i} := \mathbf{1}\left(n^{-1/2}|Z^{(2)}_{\nu i}| \leqslant q\log n \right) \cdot Z^{(2)}_{\nu i},$$

for $q = n^{-\frac{\varphi-4}{2\varphi}}$. By equation (C.15), we have

$$\mathbb{P}(\widetilde{Z}^{(1)} = Z^{(1)}, \widetilde{Z}^{(2)} = Z^{(2)}) = 1 - \mathrm{O}((\log n)^{-\varphi}). \tag{C.26}$$

By definition, we have

$$\mathbb{E}\widetilde{Z}^{(1)}_{\mu i} = -\mathbb{E}\left[\mathbf{1}\left(|Z^{(1)}_{\mu i}| > qn^{1/2}\log n \right) Z^{(1)}_{\mu i} \right], \quad \mathbb{E}|\widetilde{Z}^{(1)}_{\mu i}|^2 = 1 - \mathbb{E}\left[\mathbf{1}\left(|Z^{(1)}_{\mu i}| > qn^{1/2}\log n \right) |Z^{(1)}_{\mu i}|^2 \right]. \tag{C.27}$$

Using the formula for expectation in terms of the tail probabilities, we can check that

$$\begin{aligned}
\mathbb{E}\left| \mathbf{1}\left(|Z^{(1)}_{\mu i}| > qn^{1/2}\log n \right) Z^{(1)}_{\mu i} \right| &= \int_0^\infty \mathbb{P}\left(\left| \mathbf{1}\left(|Z^{(1)}_{\mu i}| > qn^{1/2}\log n \right) Z^{(1)}_{\mu i} \right| > s \right) \mathrm{d}s \\
&= \int_0^{qn^{1/2}\log n} \mathbb{P}\left(|Z^{(1)}_{\mu i}| > qn^{1/2}\log n \right) \mathrm{d}s + \int_{qn^{1/2}\log n}^\infty \mathbb{P}\left(|Z^{(1)}_{\mu i}| > s \right) \mathrm{d}s \\
&\lesssim \int_0^{qn^{1/2}\log n} \left(qn^{1/2}\log n \right)^{-\varphi} \mathrm{d}s + \int_{qn^{1/2}\log n}^\infty s^{-\varphi} \mathrm{d}s \leqslant n^{-2(\varphi-1)/\varphi},
\end{aligned}$$

where in the third step we used the finite φ-th moment condition of $Z^{(1)}_{\mu i}$ and Markov's inequality. Similarly, we can obtain that

$$\begin{aligned}\mathbb{E}\left|\mathbf{1}\left(|Z^{(1)}_{\mu i}| > qn^{1/2}\log n\right)Z^{(1)}_{\mu i}\right|^2 &= 2\int_0^\infty s\mathbb{P}\left(\left|\mathbf{1}\left(|Z^{(1)}_{\mu i}| > qn^{1/2}\log n\right)Z^{(1)}_{\mu i}\right| > s\right)\mathrm{d}s \\ &= 2\int_0^{qn^{1/2}\log n} s\mathbb{P}\left(|Z^{(1)}_{\mu i}| > qn^{1/2}\log n\right)\mathrm{d}s + 2\int_{qn^{1/2}\log n}^\infty s\mathbb{P}\left(|Z^{(1)}_{\mu i}| > s\right)\mathrm{d}s \\ &\lesssim \int_0^{qn^{1/2}\log n} s\left(qn^{1/2}\log n\right)^{-\varphi}\mathrm{d}s + \int_{qn^{1/2}\log n}^\infty s^{-\varphi+1}\mathrm{d}s \leqslant n^{-2(\varphi-2)/\varphi}.\end{aligned}$$

Plugging the above two estimates into equation (C.27) and using $\varphi > 4$, we get that

$$|\mathbb{E}\widetilde{Z}^{(1)}_{\mu i}| = \mathrm{O}(n^{-3/2}), \quad \mathbb{E}|\widetilde{Z}^{(1)}_{\mu i}|^2 = 1 + \mathrm{O}(n^{-1}). \tag{C.28}$$

From the first estimate in equation (C.28), we can also get a bound on the operator norm:

$$\|\mathbb{E}\widetilde{Z}^{(1)}\| = \mathrm{O}(n^{-1/2}). \tag{C.29}$$

Similar estimates also hold for $\widetilde{Z}^{(2)}$. Then we can centralize and rescale $\widetilde{Z}^{(1)}$ and $\widetilde{Z}^{(2)}$ as

$$\widehat{Z}^{(1)} := \frac{\widetilde{Z}^{(1)} - \mathbb{E}\widetilde{Z}^{(1)}}{\left(\mathbb{E}|\widetilde{Z}^{(1)}_{\mu i}|^2\right)^{1/2}}, \quad \widehat{Z}^{(2)} := \frac{\widetilde{Z}^{(2)} - \mathbb{E}\widetilde{Z}^{(2)}}{\left(\mathbb{E}|\widetilde{Z}^{(2)}_{\mu i}|^2\right)^{1/2}}.$$

Now $\widehat{Z}^{(1)}$ and $\widehat{Z}^{(2)}$ satisfy the assumptions of Lemma C.8 with bounded support $\sqrt{n}q = n^{\frac{2}{\varphi}}$, so we get that

$$\left|\mathbf{u}^\top(G(\widehat{Z}^{(1)}, \widehat{Z}^{(2)}, z) - \mathfrak{G}(z))\mathbf{v}\right| \prec q, \tag{C.30}$$

where $G(\widehat{Z}^{(1)}, \widehat{Z}^{(2)}, z)$ is defined in the same way as $G(z)$, but with $(Z^{(1)}, Z^{(2)})$ replaced by $(\widehat{Z}^{(1)}, \widehat{Z}^{(2)})$.

Note that by equations (C.28) and (C.29), we can bound that for $k = 1, 2$,

$$\|\widehat{Z}^{(k)} - \widetilde{Z}^{(k)}\| \lesssim n^{-1}\|\widetilde{Z}^{(k)}\| + \|\mathbb{E}\widetilde{Z}^{(k)}\| \lesssim n^{-1/2}$$

with overwhelming probability, where we also used Fact E.1(ii) to bound the operator norm of $\widetilde{Z}^{(k)}$. Together with estimate (C.55) below, this bound implies that

$$\left|\mathbf{u}^\top(G(\widehat{Z}^{(1)}, \widehat{Z}^{(2)}, z) - G(Z^{(1)}, Z^{(2)}, z))\mathbf{v}\right| \lesssim n^{-1/2}\|\widehat{Z}^{(1)} - \widetilde{Z}^{(1)}\| + n^{-1/2}\|\widehat{Z}^{(2)} - \widetilde{Z}^{(2)}\| \lesssim n^{-1},$$

with overwhelming probability on the event $\{\widetilde{Z}^{(1)} = Z^{(1)}, \widetilde{Z}^{(2)} = Z^{(2)}\}$. Combining this estimate with equation (C.30), we obtain that estimate (C.24) holds for $G(z)$ on the event $\{\widetilde{Z}^{(1)} = Z^{(1)}, \widetilde{Z}^{(2)} = Z^{(2)}\}$, which concludes the proof by equation (C.26). □

Now we are ready to complete the proof of Theorem 3.1 and Theorem C.1 using Theorem C.7.

Proof of Theorem 3.1. With the definition of matrix W in equation (3.8), we can express $\Sigma^{(2)}\hat{\Sigma}^{-1}$ as

$$\Sigma^{(2)}\hat{\Sigma}^{-1} = n^{-1}(\Sigma^{(2)})^{1/2}V\mathcal{G}(0)V^\top(\Sigma^{(2)})^{-1/2},$$

where we recall that $\mathcal{G}(z) = (W - z\,\mathrm{Id})^{-1}$ is the resolvent of W. Then by Theorem C.7, for any $1 \leqslant i \leqslant p$ we have that

$$\begin{aligned}\left|\left[\Sigma^{(2)}\hat{\Sigma}^{-1} - n^{-1}(\Sigma^{(2)})^{1/2}V\mathfrak{G}(0)V^\top(\Sigma^{(2)})^{-1/2}\right]_{ii}\right| &= n^{-1}\left|\mathbf{e}_i^\top(\Sigma^{(2)})^{1/2}V\left(\mathcal{G}(0) - \mathfrak{G}(0)\right)V^\top(\Sigma^{(2)})^{-1/2}\mathbf{e}_i\right| \\ &\prec n^{-1}q\|V^\top(\Sigma^{(2)})^{-1/2}\mathbf{e}_i\| \cdot \|V^\top(\Sigma^{(2)})^{1/2}\mathbf{e}_i\| \lesssim n^{-1}q, \quad \text{(C.31)}\end{aligned}$$

on the event (C.25), where $q = n^{-\frac{\varphi-4}{2\varphi}}$ and $\mathbf{e}_i$ denotes the standard basis vector along the i-th direction. Next, we can verify that

$$n^{-1}(\Sigma^{(2)})^{1/2}V\mathfrak{G}(0)V^\top(\Sigma^{(2)})^{-1/2} = n^{-1}\Sigma^{(2)}(a_1\Sigma^{(1)} + a_2\Sigma^{(2)})^{-1}.$$

Together with equation (C.31), this identity implies that

$$\operatorname{Tr}\left[\Sigma^{(2)}\hat{\Sigma}^{-1}\right] = \sum_{i=1}^{p}\left(\Sigma^{(2)}\hat{\Sigma}^{-1}\right)_{ii} = n^{-1}\operatorname{Tr}\left[\Sigma^{(2)}(a_1\Sigma^{(1)} + a_2\Sigma^{(2)})^{-1}\right] + \mathrm{O}_\prec(q)$$

on the event (C.25), where we used Fact C.4 (i) in the second step. This concludes equation (3.4) using Definition C.3 and the fact that c_φ is any fixed value within $(0, \frac{\varphi-4}{2\varphi})$. □

Proof of Theorem C.1. Recall that in the setting of Theorem 3.1, we have equation (C.8). For simplicity, we denote the vector $\mathbf{v} := V^\top(\Sigma^{(2)})^{-1/2}\Sigma^{(1)}w$. By Corollary C.7, we have that

$$\max_{z\in\mathbb{C}:|z|=(\log n)^{-1}} |\mathbf{v}^\top(G(z) - \mathfrak{G}(z))\mathbf{v}| \prec q\|\mathbf{v}\|^2,$$

on the event (C.25) with $q := n^{-\frac{\varphi-4}{2\varphi}}$. Now combining this estimate with Cauchy's integral formula, we get that

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

where $\mathcal{C}$ is the contour $\{z\in\mathbb{C} : |z| = (\log n)^{-1}\}$. We can calculate the derivative $\mathbf{v}^\top\mathfrak{G}'(0)\,\mathbf{v}$ as

$$\mathbf{v}^\top\mathfrak{G}'(0)\,\mathbf{v} = \mathbf{v}^\top\frac{a_3\Lambda^2 + (1+a_4)\operatorname{Id}_p}{(a_1\Lambda^2 + a_2\operatorname{Id}_p)^2}\mathbf{v}, \tag{C.33}$$

where we recall equation (C.9) and that $a_3 = -\frac{\mathrm{d}a_1(0)}{\mathrm{d}z}$ and $a_4 = -\frac{\mathrm{d}a_2(0)}{\mathrm{d}z}$. Taking the derivatives of the system of equations (C.7), we can derive equation (C.2) for (a_3, a_4). This concludes the proof together with equation (C.32). □

C.2 Self-Consistent Equations

The rest of this section is devoted to the proof of Lemma C.8. In this section, we show that the limiting equation (C.7) has a unique solution $(a_1(z), a_2(z))$ for any $z\in\mathbf{D}$ in equation (C.23). Otherwise, Lemma C.8 will be a vacuous statement.

When $z = 0$, the system of equations (C.7) reduces to equations (3.5) and (3.6), from which we can derive an equation of a_1 only:

$$f(a_1) = \frac{n_1}{n_1+n_2}, \quad \text{with} \quad f(a_1) := a_1 + \frac{1}{n_1+n_2}\sum_{i=1}^{p}\frac{\lambda_i^2 a_1}{\lambda_i^2 a_1 + (1 - \frac{p}{n_1+n_2} - a_1)}. \tag{C.34}$$

It is not hard to see that f is strictly increasing on $[0, 1-\frac{p}{n_1+n_2}]$. Moreover, we have $f(0) = 0 < 1$, $f(1-\frac{p}{n_1+n_2}) = 1 > \frac{n_1}{n_1+n_2}$, and $f(\frac{n_1}{n_1+n_2}) > \frac{n_1}{n_1+n_2}$ if $\frac{n_1}{n_1+n_2} \leqslant 1 - \frac{p}{n_1+n_2}$. Hence by mean value theorem, there exists a unique solution a_1 satisfying

$$0 < a_1 < \min\left(1 - \frac{p}{n_1+n_2}, \frac{n_1}{n_1+n_2}\right).$$

Moreover, it is easy to check that $f'(x) = \mathrm{O}(1)$ for $x \in [0, 1-\frac{p}{n_1+n_2}]$. Hence there exists a constant $\tau > 0$, such that

$$\frac{n_1}{n_1+n_2}\tau \leqslant a_1 \leqslant \min\left\{1 - \frac{p}{n_1+n_2} - \frac{n_1}{n_1+n_2}\tau, \frac{n_1}{n_1+n_2}(1-\tau)\right\}, \quad \tau < a_2 \leqslant 1 - \frac{p}{n_1+n_2} - \frac{n_1}{n_1+n_2}\tau. \tag{C.35}$$

Next, we prove the existence and uniqueness of the solution to the self-consistent equation (C.7) for a general $z \in \mathbf{D}$. We denote

$$M_1(z) := -\frac{n_1+n_2}{n_1} a_1(z), \quad M_2(z) := -\frac{n_1+n_2}{n_2} a_2(z), \tag{C.36}$$

which are the asymptotic limits of $m_1(z)$ and $m_2(z)$ in equation (C.12). Then, it is not hard to verify that the system of equations (C.7) can be rewritten as the following system of equations:

$$\frac{1}{M_1} = \frac{\gamma_n}{p} \sum_{i=1}^p \frac{\lambda_i^2}{z + \lambda_i^2 r_1 M_1 + r_2 M_2} - 1, \quad \frac{1}{M_2} = \frac{\gamma_n}{p} \sum_{i=1}^p \frac{1}{z + \lambda_i^2 r_1 M_1 + r_2 M_2} - 1, \tag{C.37}$$

where, for simplicity of notations, we introduced the following ratios

$$\gamma_n := \frac{p}{n_1+n_2}, \quad r_1 := \frac{n_1}{n_1+n_2}, \quad r_2 := \frac{n_2}{n_1+n_2}. \tag{C.38}$$

One can compare equation (C.37) for $(M_1(z), M_2(z))$ to equation (C.12) for $(m_1(z), m_2(z))$. In the following proof, we shall focus on the system of equations (C.37) because it is more suitable than equation (C.7) for our purpose of showing that $(m_1(z), m_2(z))$ converges to the asymptotic limit $(M_1(z), M_2(z))$.

Now we claim the following lemma, which gives the existence and uniqueness of the solution $(M_1(z), M_2(z))$ to the system of equations (C.37).

Lemma C.10. *There exist constants $c_0, C_0 > 0$ depending only on τ in Assumption 2.2 and equation* (C.35) *such that the following statements hold. There exists a unique solution to equation* (C.37) *under the conditions*

$$|z| \leqslant c_0, \quad |M_1(z) - M_1(0)| + |M_2(z) - M_2(0)| \leqslant c_0. \tag{C.39}$$

Moreover, the solution satisfies

$$|M_1(z) - M_1(0)| + |M_2(z) - M_2(0)| \leqslant C_0 |z|. \tag{C.40}$$

Proof. The proof is a standard application of the contraction principle. For reader's convenience, we give more details. First, it is easy to check that equation (C.37) is equivalent to

$$r_1 M_1 = -(1-\gamma_n) - r_2 M_2 - z\left(M_2^{-1} + 1\right), \quad g_z(M_2(z)) = 1, \tag{C.41}$$

where

$$g_z(M_2) := -M_2 + \frac{\gamma_n}{p} \sum_{i=1}^p \frac{M_2}{z - \lambda_i^2(1-\gamma_n) + (1-\lambda_i^2) r_2 M_2 - \lambda_i^2 z \left(M_2^{-1} + 1\right)}.$$

We first show that there exists a unique solution $M_2(z)$ to the equation $g_z(M_2(z)) = 1$ under the conditions in equation (C.39). We abbreviate $\delta(z) := M_2(z) - M_2(0)$. From equation (C.41), we obtain that

$$0 = [g_z(M_2(z)) - g_0(M_2(0)) - g_z'(M_2(0))\delta(z)] + g_z'(M_2(0))\delta(z),$$

which gives that

$$\delta(z) = -\frac{g_z(M_2(0)) - g_0(M_2(0))}{g_z'(M_2(0))} - \frac{g_z(M_2(0) + \delta(z)) - g_z(M_2(0)) - g_z'(M_2(0))\delta(z)}{g_z'(M_2(0))}.$$

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

$$\delta^{(k+1)} = -\frac{g_z(M_2(0)) - g_0(M_2(0))}{g_z'(M_2(0))} - \frac{g_z(M_2(0) + \delta^{(k)}) - g_z(M_2(0)) - g_z'(M_2(0))\delta^{(k)}}{g_z'(M_2(0))}. \tag{C.42}$$

Then equation (C.42) defines a mapping $h_z : \mathbb{C} \to \mathbb{C}$, which maps $\delta^{(k)}$ to $\delta^{(k+1)} = h_z(\delta^{(k)})$.

With direct calculation, we obtain that

$$g_z'(M_2(0)) = -1 - \frac{\gamma_n}{p} \sum_{i=1}^p \frac{\lambda_i^2(1-\gamma_n) - z\left[1 - \lambda_i^2\left(2M_2^{-1}(0) + 1\right)\right]}{\left[z - \lambda_i^2(1-\gamma_n) + (1-\lambda_i^2) r_2 M_2(0) - \lambda_i^2 z \left(M_2^{-1}(0) + 1\right)\right]^2}.$$

Then it is not hard to check that there exist constants $\widetilde{c}, \widetilde{C} > 0$ depending only on τ such that the following estimates hold: for all z, δ_1 and δ_2 such that $|z| \leqslant \widetilde{c}$, $|\delta_1| \leqslant \widetilde{c}$ and $|\delta_2| \leqslant \widetilde{c}$,

$$\left|\frac{1}{g_z'(M_2(0))}\right| \leqslant \widetilde{C}, \quad \left|\frac{g_z(M_2(0)) - g_0(M_2(0))}{g_z'(M_2(0))}\right| \leqslant \widetilde{C}|z|, \tag{C.43}$$

and

$$\left|\frac{g_z(M_2(0)+\delta_1) - g_z(M_2(0)+\delta_2) - g_z'(M_2(0))(\delta_1 - \delta_2)}{g_z'(M_2(0))}\right| \leqslant \widetilde{C}|\delta_1 - \delta_2|^2. \tag{C.44}$$

Using equations (C.43) and (C.44), it is not hard to show that there exists a sufficiently small constant $c_1 > 0$ depending only on $\widetilde{C}$, such that $h_z : B_d \to B_d$ is a self-mapping on the ball $B_d := \{\delta \in \mathbb{C} : |\delta| \leqslant d\}$, as long as $d \leqslant c_1$ and $|z| \leqslant c_1$. Now it suffices to prove that h restricted to B_d is a contraction, which then implies that $\delta := \lim_{k\to\infty} \delta^{(k)}$ exists and $M_2(0) + \delta(z)$ is a unique solution to equation $g_z(M_2(z)) = 1$ subject to the condition $\|\delta\|_\infty \leqslant d$.

From the iteration relation (C.42), using equation (C.44) one can readily check that

$$\delta^{(k+1)} - \delta^{(k)} = h_z(\delta^{(k)}) - h_z(\delta^{(k-1)}) \leqslant \widetilde{C}|\delta^{(k)} - \delta^{(k-1)}|^2. \tag{C.45}$$

Hence as long as d is chosen to be sufficiently small such that $2d\widetilde{C} \leqslant 1/2$, then h is indeed a contraction mapping on B_d. This proves both the existence and uniqueness of the solution $M_2(z) = M_2(0) + \delta(z)$, if we choose c_0 in equation (C.39) as $c_0 = \min\{c_1, d\}$. After obtaining $M_2(z)$, we can then find $M_1(z)$ using the first equation in (C.41).

Note that with equation (C.43) and $\delta^{(0)} = 0$, we can obtain from equation (C.42) that $|\delta^{(1)}(z)| \leqslant \widetilde{C}|z|$. With the contraction mapping, we have the bound

$$|\delta| \leqslant \sum_{k=0}^{\infty} |\delta^{(k+1)} - \delta^{(k)}| \leqslant 2\widetilde{C}|z| \;\Rightarrow\; |M_2(z) - M_2(0)| \leqslant 2\widetilde{C}|z|. \tag{C.46}$$

Then using the first equation in equation (C.41), we immediately obtain the bound $r_1|M_1(z) - M_1(0)| \leqslant C|z|$ for some constant $C > 0$, which concludes equation (C.40) as long as if $r_1 \gtrsim 1$. To deal with the $r_1 = \mathrm{o}(1)$ case, we go back to the first equation in (C.37) and treat $M_1(z)$ as the solution to the following equation:

$$\widetilde{g}_z(M_1(z)) = 1, \quad \widetilde{g}_z(M_1) := -M_1 + \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{\lambda_i^2 M_1}{z + \lambda_i^2 r_1 M_1 + r_2 M_2(z)}.$$

(Note that we have found the solution $M_2(z)$, so this is an equation of M_1 only.) Then with a similar argument as above (i.e. the proof between equation (C.41) and equation (C.46)), we can conclude $|M_2(z) - M_2(0)| = \mathrm{O}(|z|)$, which further concludes equation (C.40) together with equation (C.46). We omit the details. □

As a byproduct of the above contraction mapping argument, we also obtain the following stability result that will be used in the proof of Lemma C.8. Roughly speaking, it states that if two complex functions $m_1(z)$ and $m_2(z)$ satisfy the self-consistent equation (C.37) approximately up to some small errors, then $m_1(z)$ and $m_2(z)$ will be close to the solutions $M_1(z)$ and $M_2(z)$. Later this result will be applied to equation (C.12) to show that the averaged resolvents $m_1(z)$ and $m_2(z)$ indeed converge to $M_1(z)$ and $M_2(z)$, respectively.

Lemma C.11. *There exist constants $c_0, C_0 > 0$ depending only on τ in Assumption 2.2 and equation* (C.35) *such that the self-consistent equations in equation* (C.37) *are stable in the following sense. Suppose $|z| \leqslant c_0$, and m_1 and m_2 are analytic functions of z such that*

$$|m_1(z) - M_1(0)| + |m_2(z) - M_2(0)| \leqslant c_0. \tag{C.47}$$

Moreover, assume that (m_1, m_2) satisfies the system of equations

$$\frac{1}{m_1} + 1 - \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{\lambda_i^2}{z + \lambda_i^2 r_1 m_1 + r_2 m_2} = \mathcal{E}_1, \quad \frac{1}{m_2} + 1 - \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{1}{z + \lambda_i^2 r_1 m_1 + r_2 m_2} = \mathcal{E}_2, \tag{C.48}$$

for some (deterministic or random) errors such that $|\mathcal{E}_1| + |\mathcal{E}_2| \leqslant \theta(z)$, where $\theta(z)$ is a deterministic function of z satisfying that $\theta(z) \leqslant (\log n)^{-1}$. Then we have

$$|m_1(z) - M_1(z)| + |m_2(z) - M_2(z)| \leqslant C_0\delta(z). \tag{C.49}$$

Proof. Under condition (C.47), we can obtain equation (C.41) approximately up to some small error:

$$r_1 m_1 = -(1-\gamma_n) - r_2 m_2 - z\left(m_2^{-1}+1\right) + \widetilde{\mathcal{E}}_1(z), \quad g_z(m_2(z)) = 1 + \widetilde{\mathcal{E}}_2(z), \tag{C.50}$$

where the errors satisfy that $|\widetilde{\mathcal{E}}_1(z)| + |\widetilde{\mathcal{E}}_2(z)| = \mathrm{O}(\theta(z))$. Then we subtract equation (C.41) from equation (C.50), and consider the contraction principle for the function $\delta(z) := m_2(z) - M_2(z)$. The rest of the proof is exactly the same as the one for Lemma C.10, so we omit the details. □

C.3 Beyond Multivariate Gaussian Random Matrices: an Anisotropic Local Law

In this section, we prove Lemma C.8 by extending from the Gaussian random matrices to general random matrices. The main difficulty in the proof is due to the fact that the entries of $Z^{(1)}U\Lambda$ and $Z^{(2)}V$ are not independent. When the entries of $Z^{(1)}$ and $Z^{(2)}$ are sampled i.i.d. from an isotropic Gaussian distribution, $Z^{(1)}U$ and $Z^{(2)}V$ still obey the Gaussian distribution. In this case, the problem is reduced to proving the anisotropic local law for $G(z)$ with $U = \mathrm{Id}$ and $V = \mathrm{Id}$, such that the entries of $Z^{(1)}\Lambda$ and $Z^{(2)}$ are independent. For this case, we use the standard resolvent methods in Bloemendal et al. (2014); Yang (2019); Pillai and Yin (2014) and prove the following result.

Proposition C.12. *In the setting of Lemma C.8, assume further that the entries of $Z^{(1)}$ and $Z^{(2)}$ are i.i.d. Gaussian random variables. Suppose U and V are identity. Then, the estimate* (C.24) *holds for all $z \in \mathbf{D}$ with $q = n^{-1/2}$.*

Note that if the entries of $Z^{(1)}$ and $Z^{(2)}$ are Gaussian, then we have $\varphi = \infty$, which gives $q = n^{-\frac{\varphi-4}{2\varphi}} = n^{-1/2}$.

Next we briefly describe how to extend Lemma C.8 from the Gaussian case to the case with general $Z^{(1)}$ and $Z^{(2)}$ satisfying the bounded support condition (C.14) with $Q = \sqrt{n}q = n^{\frac{2}{\varphi}}$. With Proposition C.12, it suffices is to prove that for $Z^{(1)}$ and $Z^{(2)}$ satisfying the assumptions in Lemma C.8, we have

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

for any deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{p+n_1+n_2}$ and $z \in \mathbf{D}$, where we abbreviated that

$$Z := \begin{pmatrix} Z^{(1)} \\ Z^{(2)} \end{pmatrix}, \quad \text{and} \quad Z^{\mathrm{Gauss}} := \begin{pmatrix} (Z^{(1)})^{\mathrm{Gauss}} \\ (Z^{(2)})^{\mathrm{Gauss}} \end{pmatrix}.$$

We will prove the above statement using a continuous comparison argument developed in Knowles and Yin (2016). Since the proof is almost the same as the ones in Sections 7 and 8 of Knowles and Yin (2016) and Section 6 of Yang (2019), we only describe the main ideas without writing down all the details.

We define the following continuous sequence of interpolating matrices between Z^{Gauss} and Z.

Definition C.13 (Interpolation). We denote $Z^0 := Z^{\mathrm{Gauss}}$ and $Z^1 := Z$. Let $\rho^0_{\mu i}$ and $\rho^1_{\mu i}$ be the laws of $Z^0_{\mu i}$ and $Z^1_{\mu i}$, respectively, for $i \in \mathcal{I}_0$ and $\mu \in \mathcal{I}_1 \cup \mathcal{I}_2$. For any $\theta \in [0,1]$, we define the interpolated law $\rho^\theta_{\mu i} := (1-\theta)\rho^0_{\mu i} + \theta\rho^1_{\mu i}$. We shall work on the probability space consisting of triples (Z^0, Z^θ, Z^1) of independent $n \times p$ random matrices, where the matrix $Z^\theta = (Z^\theta_{\mu i})$ has law

$$\prod_{i\in\mathcal{I}_0}\prod_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2} \rho^\theta_{\mu i}(\mathrm{d}Z^\theta_{\mu i}). \tag{C.51}$$

For $\lambda \in \mathbb{R}$, $i \in \mathcal{I}_0$ and $\mu \in \mathcal{I}_1 \cup \mathcal{I}_2$, we define the matrix $Z^{\theta,\lambda}_{(\mu i)}$ through

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

that is, it replaces the (μ, i)-th entry of Z^θ with λ.

Proof of Lemma C.8. We shall prove equation (C.24) through interpolation matrices Z^θ between Z^0 and Z^1. We have seen that equation (C.24) holds for $G(Z^0, z)$ by Proposition C.12. Using equation (C.51) and fundamental calculus, we get the following basic interpolation formula: for differentiable $F : \mathbb{R}^{n\times p} \to \mathbb{C}$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}F(Z^\theta) = \sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}\left[\mathbb{E}F\left(Z^{\theta,Z^1_{\mu i}}_{(\mu i)}\right) - \mathbb{E}F\left(Z^{\theta,Z^0_{\mu i}}_{(\mu i)}\right)\right], \tag{C.52}$$

provided all the expectations exist. We shall apply equation (C.52) to the function $F(Z) := F^s_{\mathbf{u}\,\mathbf{v}}(Z,z)$ for any fixed $s \in 2\mathbb{N}$, where

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

The main part of the proof is to show the following self-consistent estimate for the right-hand side of equation (C.52): for any fixed $s \in 2\mathbb{N}$, any constant $c > 0$ and all $\theta \in [0,1]$,

$$\sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}\left[\mathbb{E}F^s_{\mathbf{u}\,\mathbf{v}}\left(Z^{\theta,Z^1_{\mu i}}_{(\mu i)},z\right) - \mathbb{E}F^s_{\mathbf{u}\,\mathbf{v}}\left(Z^{\theta,Z^0_{\mu i}}_{(\mu i)},z\right)\right] \leqslant (n^c q)^s + C\mathbb{E}F^s_{\mathbf{u}\,\mathbf{v}}\left(Z^\theta, z\right), \tag{C.53}$$

for some constant $C > 0$. If equation (C.53) holds, then combining equation (C.52) with Grönwall's inequality we obtain that for any fixed $s \in 2\mathbb{N}$ and constant $c > 0$,

$$\mathbb{E}\left|\mathbf{u}^\top\left(G(Z^1,z) - \Pi(z)\right)\mathbf{v}\right|^s \lesssim (n^c q)^s.$$

Finally applying Markov's inequality and noticing that c can be chosen arbitrarily small, we conclude equation (C.24). Underlying the proof of the estimate (C.53) is an expansion approach, which is the same as the ones for Lemma 7.10 of Knowles and Yin (2016) and Lemma 6.11 of Yang (2019). So we omit the details. □

Now it remains to prove Proposition C.12, whose proof is based on the following entrywise local law.

Lemma C.14. *Under the assumptions of Proposition C.12, the following estimate holds uniformly in $z \in \mathbf{D}$:*

$$\max_{a,b\in\mathcal{I}}\left|G_{ab}(z) - \mathfrak{G}_{ab}(z)\right| \prec n^{-1/2}. \tag{C.54}$$

With Lemma C.14, we can complete the proof of Proposition C.12.

Proof of Proposition C.12. With estimate (C.54), one can use the polynomialization method in Section 5 of Bloemendal et al. (2014) to get the anisotropic local law (C.24) with $q = n^{-1/2}$. The proof is exactly the same, except for some minor differences in notations. Hence we omit the details. □

C.4 An Entrywise Local Law

Finally, this subsection is devoted to the proof of Lemma C.14. First, we claim the following a priori estimate on the resolvent $G(z)$ for $z \in \mathbf{D}$.

Lemma C.15. *In the setting of Lemma C.8, there exists a constant $C > 0$ such that the following estimates hold uniformly in $z, z' \in \mathbf{D}$ with overwhelming probability:*

$$\|G(z)\| \leqslant C, \tag{C.55}$$

and

$$\|G(z) - G(z')\| \leqslant C|z - z'|. \tag{C.56}$$

Proof. Our proof is a simple application of the spectral decomposition of G. Recall the matrix F defined in equation (C.3). Let

$$F = \sum_{k=1}^{p}\sqrt{\mu_k}\,\xi_k\zeta_k^\top, \quad \mu_1 \geqslant \mu_2 \geqslant \cdots \geqslant \mu_p \geqslant 0 = \mu_{p+1} = \ldots = \mu_n, \tag{C.57}$$

be a singular value decomposition of A, where $\{\xi_k\}_{k=1}^p$ are the left-singular vectors and $\{\zeta_k\}_{k=1}^n$ are the right-singular vectors. Then using equation (C.5), we get that for $i, j \in \mathcal{I}_1$ and $\mu, \nu \in \mathcal{I}_1 \cup \mathcal{I}_2$,

$$G_{ij} = \sum_{k=1}^{p}\frac{\xi_k(i)\xi_k^\top(j)}{\mu_k - z}, \quad G_{\mu\nu} = z\sum_{k=1}^{n}\frac{\zeta_k(\mu)\zeta_k^\top(\nu)}{\mu_k - z}, \quad G_{i\mu} = G_{\mu i} = \sum_{k=1}^{p}\frac{\sqrt{\mu_k}\,\xi_k(i)\zeta_k^\top(\mu)}{\mu_k - z}. \tag{C.58}$$

By Fact E.1 (ii), we have that with overwhelming probability $\mu_p \geqslant \lambda_p(n^{-1}(Z^{(2)})^\top Z^{(2)}) \geqslant c_\tau$ for some constant $c_\tau > 0$ depending only on τ. This further implies that

$$\inf_{z\in\mathbf{D}}\min_{1\leqslant k\leqslant p}|\mu_k - z| \geqslant c_\tau - (\log n)^{-1}.$$

Combining this bound with equation (C.58), we can easily conclude the estimates (C.55) and (C.56). □

For the rest of this subsection, we present the proof of Lemma C.14, which is the most technical part of the whole proof of Lemma C.8.

Proof of Lemma C.14. Recall that under the assumptions of Lemma C.14, we have

$$F \overset{d}{=} n^{-1/2}[\Lambda (Z^{(1)})^\top, (Z^{(2)})^\top], \tag{C.59}$$

and it suffices to consider the resolvent in equation (C.10) throughout the whole proof. The proof is divided into three steps. For simplicity, we introduce the following notation: for two (deterministic or random) nonnegative quantities ξ and ζ, we write $\xi \sim \zeta$ if $\xi \lesssim \zeta$ and $\zeta \lesssim \xi$.

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

$$\mathcal{Z}_a := (1 - \mathbb{E}_a)\left[(G_{aa})^{-1} \right],$$

where $\mathbb{E}_a[\cdot] := \mathbb{E}[\cdot \mid H^{(a)}]$ denotes the partial expectation over the entries in the a-th row and column of H. Now using equation (C.16), we get that for $i \in \mathcal{I}_0$,

$$\mathcal{Z}_i = \frac{\lambda_i^2}{n} \sum_{\mu,\nu \in \mathcal{I}_1} G^{(i)}_{\mu\nu}\left(\delta_{\mu\nu} - Z^{(1)}_{\mu i} Z^{(1)}_{\nu i}\right) + \frac{1}{n} \sum_{\mu,\nu \in \mathcal{I}_2} G^{(i)}_{\mu\nu}\left(\delta_{\mu\nu} - Z^{(2)}_{\mu i} Z^{(2)}_{\nu i}\right) - 2\frac{\lambda_i}{n} \sum_{\mu \in \mathcal{I}_1, \nu \in \mathcal{I}_2} Z^{(1)}_{\mu i} Z^{(2)}_{\nu i} G^{(i)}_{\mu\nu}, \tag{C.60}$$

and for $\mu \in \mathcal{I}_1$ and $\nu \in \mathcal{I}_2$,

$$\mathcal{Z}_\mu = \frac{1}{n} \sum_{i,j \in \mathcal{I}_0} \lambda_i \lambda_j G^{(\mu)}_{ij}\left(\delta_{ij} - Z^{(1)}_{\mu i} Z^{(1)}_{\mu j}\right), \quad \mathcal{Z}_\nu = \frac{1}{n} \sum_{i,j \in \mathcal{I}_0} G^{(\nu)}_{ij}\left(\delta_{ij} - Z^{(2)}_{\nu i} Z^{(2)}_{\nu j}\right). \tag{C.61}$$

Moreover, we introduce the random error

$$\Lambda_o := \max_{a \ne b} \left| G_{aa}^{-1} G_{ab} \right|, \tag{C.62}$$

which controls the size of the off-diagonal entries. The following lemma gives the desired large deviation estimate on Λ_o and $\mathcal{Z}$ variables.

Lemma C.16. *Under the assumptions of Proposition C.12, the following estimate holds uniformly in all $z \in \mathbf{D}$:*

$$\Lambda_o + \max_{a \in \mathcal{I}} |\mathcal{Z}_a| \prec n^{-1/2}. \tag{C.63}$$

Proof. Note that for any $a \in \mathcal{I}$, $H^{(a)}$ and $G^{(a)}$ also satisfies the assumptions in Lemma C.15. Hence equations (C.55) and (C.56) also hold for $G^{(a)}$ with overwhelming probability. Now applying equations (C.21) and (C.22) to equation (C.60), we get that for any $i \in \mathcal{I}_0$,

$$\begin{aligned} |\mathcal{Z}_i| &\lesssim \frac{1}{n}\left| \sum_{\mu,\nu \in \mathcal{I}_1} G^{(i)}_{\mu\nu}\left(\delta_{\mu\nu} - Z^{(1)}_{\mu i} Z^{(1)}_{\nu i}\right)\right| + \frac{1}{n}\left| \sum_{\mu,\nu \in \mathcal{I}_2} G^{(i)}_{\mu\nu}\left(\delta_{\mu\nu} - Z^{(2)}_{\mu i} Z^{(2)}_{\nu i}\right)\right| + \frac{1}{n}\left| \sum_{\mu \in \mathcal{I}_1, \nu \in \mathcal{I}_2} Z^{(1)}_{\mu i} Z^{(2)}_{\nu i} G^{(i)}_{\mu\nu}\right| \\ &\prec \frac{1}{n}\left(\sum_{\mu,\nu \in \mathcal{I}_1 \cup \mathcal{I}_2} |G^{(i)}_{\mu\nu}|^2 \right)^{1/2} \prec n^{-1/2}. \end{aligned}$$

Here in the last step we used equation (C.55) to get that for any $\mu \in \mathcal{I}_1 \cup \mathcal{I}_2$,

$$\sum_{\nu \in \mathcal{I}_1 \cup \mathcal{I}_2} |G^{(i)}_{\mu\nu}|^2 \leqslant \sum_{a \in \mathcal{I}} |G^{(i)}_{\mu a}|^2 = \left[G^{(i)} (G^{(i)})^* \right]_{\mu\mu} = \mathrm{O}(1), \quad \text{with overwhelming probability,} \tag{C.64}$$

where $(G^{(i)})^*$ denotes the complex conjugate transpose of $G^{(i)}$. Similarly, applying equations (C.21) and (C.22) to $\mathcal{Z}_\mu$ and $\mathcal{Z}_\nu$ in equation (C.61) and using equation (C.55), we can obtain the same bound.

Next we prove the off-diagonal estimate on Λ_o. For $i \in \mathcal{I}_1$ and $a \in \mathcal{I} \setminus \{i\}$, using equations (C.17), (C.21) and (C.55), we can obtain that

$$\left|G_{ii}^{-1} G_{ia}\right| \leqslant n^{-1/2} \left|\lambda_i \sum_{\mu \in \mathcal{I}_1} Z_{\mu i}^{(1)} G_{\mu a}^{(i)}(z)\right| + n^{-1/2} \left|\sum_{\mu \in \mathcal{I}_2} Z_{\mu i}^{(2)} G_{\mu a}^{(i)}(z)\right| \prec n^{-1/2} \left(\sum_{\mu \in \mathcal{I}_1 \cup \mathcal{I}_2} |G_{\mu a}^{(i)}|^2\right)^{1/2} \prec n^{-1/2}.$$

We can get the same estimate for $\left|G_{\mu\mu}^{-1} G_{\mu b}\right|$, $\mu \in \mathcal{I}_1 \cup \mathcal{I}_2$ and $b \in \mathcal{I} \setminus \{\mu\}$, using a similar argument. Thus we obtain that $\Lambda_o \prec n^{-1/2}$. □

Note that combining $\max_a |G_{aa}| = \mathrm{O}(1)$ by equation (C.55) with equation (C.63), we immediately conclude equation (C.54) for the off-diagonal entries with $a \neq b$.

Step 2: Self-consistent equations. In this step, we derive the approximate self-consistent equations in (C.12) satisfied by $m_1(z)$ and $m_2(z)$ with more precise error rates. More precisely, we will show that $(m_1(z), m_2(z))$ satisfies equation (C.48) for some small errors satisfying $|\mathcal{E}_1| + |\mathcal{E}_2| \prec n^{-1/2}$. Later in Step 3, we will apply Lemma C.11 to show that $(m_1(z), m_2(z))$ is close to $(M_1(z), M_2(z))$.

We define the following z-dependent event

$$\Xi(z) := \left\{|m_1(z) - M_1(z)| + |m_2(z) - M_2(z)| \leqslant (\log n)^{-1/2}\right\}. \tag{C.65}$$

Note that by equation (C.40), we have that for $z \in \mathbf{D}$ the following estimates hold:

$$|M_1(z) - M_1(0)| = |M_1(z) + r_1^{-1} a_1| \lesssim (\log n)^{-1}, \quad |M_2(z) + M_2(0)| = |M_2 + r_2^{-1} a_2| \lesssim (\log n)^{-1}.$$

Together with the estimates in equation (C.35) and the assumption that the singular values λ_i are bounded, we obtain the following estimates

$$|M_1| \sim |M_2| \sim 1, \quad |z + \lambda_i^2 r_1 M_1 + r_2 M_2| \sim 1, \quad \text{uniformly in } z \in \mathbf{D}. \tag{C.66}$$

Moreover, using equation (C.37) we get

$$|1 + \gamma_n M(z)| = |M_2^{-1}(z)| \sim 1, \quad |1 + \gamma_n M_0(z)| = |M_1^{-1}(z)| \sim 1, \tag{C.67}$$

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

$$M(z) := -\frac{1}{p} \sum_{i=1}^{p} \frac{1}{z + \lambda_i^2 r_1 M_1(z) + r_2 M_2(z)}, \quad M_0(z) := -\frac{1}{p} \sum_{i=1}^{p} \frac{\lambda_i^2}{z + \lambda_i^2 r_1 M_1(z) + r_2 M_2(z)}. \tag{C.68}$$

In fact, $M(z)$ and $M_0(z)$ are the asymptotic limits of $m(z)$ and $m_0(z)$, respectively. Plugging equation (C.66) into equation (C.6), we get that

$$|\mathfrak{G}_{aa}(z)| \sim 1 \text{ uniformly in } z \in \mathbf{D} \text{ and } a \in \mathcal{I}. \tag{C.69}$$

Then we prove the following key lemma, which shows that $(m_1(z), m_2(z))$ satisfies equation (C.48) with some small errors $\mathcal{E}_1$ and $\mathcal{E}_2$.

Lemma C.17. *Under the assumptions of Proposition C.12, the following estimates hold uniformly in $z \in \mathbf{D}$:*

$$\mathbf{1}(\Xi) \left|\frac{1}{m_1} + 1 - \frac{\gamma_n}{p} \sum_{i=1}^{p} \frac{\lambda_i^2}{z + \lambda_i^2 r_1 m_1 + r_2 m_2}\right| \prec n^{-1/2}, \tag{C.70}$$

and

$$\mathbf{1}(\Xi) \left|\frac{1}{m_2} + 1 - \frac{\gamma_n}{p} \sum_{i=1}^{p} \frac{1}{z + \lambda_i^2 r_1 m_1 + r_2 m_2}\right| \prec n^{-1/2}. \tag{C.71}$$

Proof. By equations (C.16), (C.60) and (C.61), we obtain that

$$\frac{1}{G_{ii}} = -z - \frac{\lambda_i^2}{n}\sum_{\mu\in\mathcal{I}_1} G^{(i)}_{\mu\mu} - \frac{1}{n}\sum_{\mu\in\mathcal{I}_2} G^{(i)}_{\mu\mu} + \mathcal{Z}_i = -z - \lambda_i^2 r_1 m_1 - r_2 m_2 + \mathcal{E}_i, \quad \text{for } i \in \mathcal{I}_0, \tag{C.72}$$

$$\frac{1}{G_{\mu\mu}} = -1 - \frac{1}{n}\sum_{i\in\mathcal{I}_0} \lambda_i^2 G^{(\mu)}_{ii} + \mathcal{Z}_\mu = -1 - \gamma_n m_0 + \mathcal{E}_\mu, \quad \text{for } \mu \in \mathcal{I}_1, \tag{C.73}$$

$$\frac{1}{G_{\nu\nu}} = -1 - \frac{1}{n}\sum_{i\in\mathcal{I}_0} G^{(\nu)}_{ii} + \mathcal{Z}_\nu = -1 - \gamma_n m + \mathcal{E}_\nu, \quad \text{for } \nu \in \mathcal{I}_2, \tag{C.74}$$

where we denoted (recall equation (C.11) and Definition C.2)

$$\mathcal{E}_i := \mathcal{Z}_i + \lambda_i^2 r_1 \left(m_1 - m_1^{(i)}\right) + r_2\left(m_2 - m_2^{(i)}\right),$$

and

$$\mathcal{E}_\mu := \mathcal{Z}_\mu + \gamma_n(m_0 - m_0^{(\mu)}), \quad \mathcal{E}_\nu := \mathcal{Z}_\nu + \gamma_n(m - m^{(\nu)}).$$

Using equations (C.18), (C.62) and (C.63), we can bound that

$$|m_1 - m_1^{(i)}| \leqslant \frac{1}{n_1}\sum_{\mu\in\mathcal{I}_1}\left|\frac{G_{\mu i}G_{i\mu}}{G_{ii}}\right| \leqslant |\Lambda_o|^2 |G_{ii}| \prec n^{-1}.$$

where we also used bound (C.55) in the last step. Similarly, we also have that

$$|m_2 - m_2^{(i)}| \prec n^{-1}, \quad |m_0 - m_0^{(\mu)}| \prec n^{-1}, \quad |m - m^{(\nu)}| \prec n^{-1},$$

for any $i \in \mathcal{I}_0$, $\mu \in \mathcal{I}_1$ and $\nu \in \mathcal{I}_2$. Together with equation (C.63), we obtain the bound

$$\max_{i\in\mathcal{I}_0} |\mathcal{E}_i| + \max_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2} |\mathcal{E}_\mu| \prec n^{-1/2}. \tag{C.75}$$

With equation (C.66) and the definition of the event Ξ in (C.65), we get that

$$\mathbf{1}(\Xi)|z + \lambda_i^2 r_1 m_1 + r_2 m_2| \sim 1.$$

Combining it with equations (C.72) and (C.75), we obtain that

$$\mathbf{1}(\Xi) G_{ii} = \mathbf{1}(\Xi)\left[-\frac{1}{z + \lambda_i^2 r_1 m_1 + r_2 m_2} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right]. \tag{C.76}$$

Plugging (C.76) into the definitions of m and m_0 in equation (C.11), we get

$$\mathbf{1}(\Xi) m = \mathbf{1}(\Xi)\left[-\frac{1}{p}\sum_{i\in\mathcal{I}_0}\frac{1}{z + \lambda_i^2 r_1 m_1 + r_2 m_2} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right], \tag{C.77}$$

$$\mathbf{1}(\Xi) m_0 = \mathbf{1}(\Xi)\left[-\frac{1}{p}\sum_{i\in\mathcal{I}_0}\frac{\lambda_i^2}{z + \lambda_i^2 r_1 m_1 + r_2 m_2} + \mathrm{O}_\prec\left(n^{-1/2}\right)\right]. \tag{C.78}$$

As a byproduct, we obtain from these two equations and equation (C.68) that

$$|m(z) - M(z)| + |m_0(z) - M_0(z)| \lesssim (\log n)^{-1/2}, \quad \text{with overwhelming probability on } \Xi. \tag{C.79}$$

Together with equation (C.67), we get that

$$|1 + \gamma_n m(z)| \sim 1, \quad |1 + \gamma_n m_0(z)| \sim 1, \quad \text{with overwhelming probability on } \Xi. \tag{C.80}$$

Now combining equations (C.73), (C.74), (C.75) and (C.80), we obtain that for $\mu \in \mathcal{I}_1$ and $\nu \in \mathcal{I}_2$,

$$\mathbf{1}(\Xi)\left(G_{\mu\mu} + \frac{1}{1+\gamma_n m_0}\right) = \mathrm{O}_\prec\left(n^{-1/2}\right), \quad \mathbf{1}(\Xi)\left(G_{\nu\nu} + \frac{1}{1+\gamma_n m}\right) = \mathrm{O}_\prec\left(n^{-1/2}\right). \tag{C.81}$$

Taking average over $\mu \in \mathcal{I}_1$ and $\nu \in \mathcal{I}_2$, we get that

$$\mathbf{1}(\Xi)\left(m_1 + \frac{1}{1+\gamma_n m_0}\right) = \mathrm{O}_{\prec}\left(n^{-1/2}\right), \quad \mathbf{1}(\Xi)\left(m_2 + \frac{1}{1+\gamma_n m}\right) = \mathrm{O}_{\prec}\left(n^{-1/2}\right), \tag{C.82}$$

which further implies

$$\mathbf{1}(\Xi)\left(\frac{1}{m_1} + 1 + \gamma_n m_0\right) \prec n^{-1/2}, \quad \mathbf{1}(\Xi)\left(\frac{1}{m_2} + 1 + \gamma_n m\right) \prec n^{-1/2}. \tag{C.83}$$

Finally, plugging equations (C.77) and (C.78) into equation (C.83), we conclude equations (C.70) and (C.71). □

Step 3: Ξ holds with overwhelming probability. In this step, we show that the event $\Xi(z)$ in (C.65) actually holds with overwhelming probability for all $z \in \mathbf{D}$. Once we have proved this fact, applying Lemma C.11 to equations (C.70) and (C.71) immediately shows that $(m_1(z), m_2(z))$ is close to $(M_1(z), M_2(z))$ up to an error of order $\mathrm{O}_{\prec}(n^{-1/2})$.

We claim that it suffices to show that

$$|m_1(0) - M_1(0)| + |m_2(0) - M_2(0)| \prec n^{-1/2}. \tag{C.84}$$

In fact, notice that by equations (C.40) and (C.56) we have

$$|M_1(z) - M_1(0)| + |M_2(z) - M_2(0)| = \mathrm{O}((\log n)^{-1}), \quad |m_1(z) - m_1(0)| + |m_2(z) - m_2(0)| = \mathrm{O}((\log n)^{-1}),$$

with overwhelming probability for all $z \in \mathbf{D}$. Thus if equation (C.84) holds, we can obtain that

$$\sup_{z\in\mathbf{D}} \left(|m_1(z) - M_1(z)| + |m_2(z) - M_2(z)|\right) \lesssim (\log n)^{-1} \quad \text{with overwhelming probability,} \tag{C.85}$$

and

$$\sup_{z\in\mathbf{D}} \left(|m_1(z) - M_1(0)| + |m_2(z) - M_2(0)|\right) \lesssim (\log n)^{-1} \quad \text{with overwhelming probability.} \tag{C.86}$$

The equation (C.85) shows that Ξ holds with overwhelming probability, while the equation (C.86) verifies the condition (C.47) of Lemma C.11. Now applying Lemma C.11 to equations (C.70) and (C.71), we obtain that

$$|m_1(z) - M_1(z)| + |m_2(z) - M_2(z)| \prec n^{-1/2} \tag{C.87}$$

uniformly for all $z \in \mathbf{D}$. Together with equations (C.81) and (C.82), equation (C.87) implies that

$$\max_{\mu\in\mathcal{I}_1} |G_{\mu\mu}(z) - M_1(z)| + \max_{\nu\in\mathcal{I}_2} |G_{\nu\nu}(z) - M_2(z)| \prec n^{-1/2}. \tag{C.88}$$

Then plugging estimate (C.87) into equation (C.76) and recalling (C.36), we obtain that

$$\max_{i\in\mathcal{I}_1} |G_{ii}(z) - \mathfrak{G}_{ii}(z)| \prec n^{-1/2}.$$

Together with equation (C.88), it gives the diagonal estimate

$$\max_{a\in\mathcal{I}} |G_{aa}(z) - \Pi_{aa}(z)| \prec n^{-1/2}. \tag{C.89}$$

Combining equation (C.89) with the off-diagonal estimate on Λ_o in equation (C.63), we conclude the proof of Lemma C.14.

Finally, we give the proof of equation (C.84). By equation (C.58), we have that with overwhelming probability,

$$m(0) = \frac{1}{p}\sum_{i\in\mathcal{I}_0} G_{ii}(0) = \frac{1}{p}\sum_{i\in\mathcal{I}_0}\sum_{k=1}^{p} \frac{|\xi_k(i)|^2}{\mu_k} \geqslant \mu_1^{-1} \gtrsim 1,$$

where we used Fact E.1 in the last step to bound $\mu_1 \geqslant \lambda_1(n^{-1}(Z^{(2)})^\top Z^{(2)}) \gtrsim 1$ with overwhelming probability. Similarly, we can also get that $m_0(0)$ is positive and has size $m_0(0) \sim 1$. Hence we have the estimates

$$1 + \gamma_n m(0) \sim 1, \quad 1 + \gamma_n m_0(0) \sim 1. \tag{C.90}$$

Combining these estimates with equations (C.73), (C.74) and (C.75), we obtain that equation (C.82) holds at $z = 0$ even without the indicator function $\mathbf{1}(\Xi)$. Furthermore, we have that with overwhelming probability,

$$\left|\lambda_i^2 r_1 m_1(0) + r_2 m_2(0)\right| = \left|\frac{\lambda_i^2 r_1}{1+\gamma_n m_0(0)} + \frac{r_2}{1+\gamma_n m(0)} + \mathrm{O}_\prec(n^{-1/2})\right| \sim 1.$$

Then combining this estimate with equations (C.72) and (C.75), we obtain that equations (C.77) and (C.78) also hold at $z = 0$ even without the indicator function $\mathbf{1}(\Xi)$. Finally, plugging equations (C.77) and (C.78) into equation (C.83), we conclude that equations (C.70) and (C.71) hold at $z = 0$, that is,

$$\begin{aligned}\left|\frac{1}{m_1(0)} + 1 - \frac{1}{n}\sum_{i=1}^p \frac{\lambda_i^2}{\lambda_i^2 r_1 m_1(0) + r_2 m_2(0)}\right| &\prec n^{-1/2},\\ \left|\frac{1}{m_2(0)} + 1 - \frac{1}{n}\sum_{i=1}^p \frac{1}{\lambda_i^2 r_1 m_1(0) + r_2 m_2(0)}\right| &\prec n^{-1/2}.\end{aligned} \tag{C.91}$$

Denoting $y_1 = -m_1(0)$ and $y_2 = -m_2(0)$, by equation (C.82) we have

$$y_1 = \frac{1}{1+\gamma_n m_0(0)} + \mathrm{O}_\prec(n^{-1/2}), \quad y_2 = \frac{1}{1+\gamma_n m(0)} + \mathrm{O}_\prec(n^{-1/2}).$$

Hence by (C.90), there exists a constant $c > 0$ such that

$$c \leqslant y_1 \leqslant 1, \quad c \leqslant y_2 \leqslant 1, \quad \text{with overwhelming probability.} \tag{C.92}$$

Also one can verify from equation (C.91) that $(r_1 y_1, r_2 y_2)$ satisfies approximately the same system of equations as equation (3.6):

$$r_1 y_1 + r_2 y_2 = 1 - \gamma_n + \mathrm{O}_\prec(n^{-1/2}), \quad r_1^{-1} f(r_1 y_1) = 1 + \mathrm{O}_\prec(n^{-1/2}), \tag{C.93}$$

where recall that the function f was defined in equation (C.34). The first equation of (C.93) and equation (C.92) together imply that $y_1 \in [0, r_1^{-1}(1-\gamma_n)]$ with overwhelming probability. For the second equation of (C.93), we know that $y_1 = r_1^{-1} a_1$ is a solution. Moreover, it is easy to check that the function $g(y_1) := r_1^{-1} f(r_1 y_1)$ is strictly increasing and has bounded derivative on $[0, r_1^{-1}(1-\gamma_n)]$. So by basic calculus, we obtain that

$$|m_1(0) - M_1(0)| = |y_1 - r_1^{-1} a_1| \prec n^{-1/2}.$$

Plugging it into the first equation of equation (C.93), we get

$$|m_2(0) - M_2(0)| = |y_2 - r_2^{-1} a_2| \prec n^{-1/2}.$$

The above two estimates conclude equation (C.84). □

D Proof of Corollary 3.3

We follow a similar logic to the proof of Theorem 2.1. We first characterize the global minimizer of $f(A, B)$ in the random-effect model. Based on the characterization, we reduce the prediction loss of hard parameter sharing to the bias-variance asymptotic limits. Finally, we prove Corollary 3.3 based on these limiting estimates. We set up several notations. In the two-task case, the optimization objective $f(A, B)$ is equal to

$$f(A,B) = \left\|X^{(1)} B A_1 - Y^{(1)}\right\|^2 + \left\|X^{(2)} B A_2 - Y^{(2)}\right\|^2, \tag{D.1}$$

where $B \in \mathbb{R}^p$ and $A = [A_1, A_2] \in \mathbb{R}^2$ because the width of B is one. Without loss of generality, we assume that A_1 and A_2 are both nonzero. Otherwise, the problem reduces to STL. Using the local optimality condition $\frac{\partial f}{\partial B} = 0$, we obtain that $\hat{B}$ satisfies the following

$$\hat{B} := \left[A_1^2 (X^{(1)})^\top X^{(1)} + A_2^2 (X^{(2)})^\top X^{(2)}\right]^{-1} \left[A_1 (X^{(1))})^\top Y^{(1)} + A_2 (X^{(2)})^\top Y^{(2)}\right]. \tag{D.2}$$

We denote $\hat{\Sigma}(x) = x^2 (X^{(1)})^\top X^{(1)} + (X^{(2)})^\top X^{(2)}$. Applying $\hat{B}$ to equation (D.1), we obtain an objective that only depends on $x := A_1/A_2$ as follows

$$\begin{aligned} g(x) := &\Big\| X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} (x\beta^{(2)} - \beta^{(1)}) + \Big(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1 \times n_1} \Big) \varepsilon^{(1)} \\ &+ x X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top \varepsilon^{(2)} \Big\|^2 \\ &+ \Big\| X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} (x\beta^{(1)} - x^2\beta^{(2)}) + \Big(X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top - \mathrm{Id}_{n_2 \times n_2} \Big) \varepsilon^{(2)} \\ &+ x X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top \varepsilon^{(1)} \Big\|^2 . \end{aligned} \tag{D.3}$$

We have that the conditional expectation of $g(x)$ over $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ is

$$\begin{aligned} &\mathop{\mathbb{E}}_{\varepsilon^{(1)}, \varepsilon^{(2)}} \Big[g(x) \mid X_1, X_2, \beta^{(1)}, \beta^{(2)} \Big] \\ =&(\beta^{(1)} - x\beta^{(2)})^\top (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} (\beta^{(1)} - x\beta^{(2)}) + \sigma^2 (n_1 + n_2 - p). \end{aligned}$$

The calculation is tedious but rather straightforward, so we leave the details to the reader. In the random-effect model, recall that the entries of $(\beta^{(1)} - x\beta^{(2)}) \in \mathbb{R}^p$ are i.i.d. Gaussian random variables with mean zero and variance $p^{-1}[(x-1)^2\kappa^2 + (1+x^2)d^2]$. Hence, by further taking expectation over $\beta^{(1)}$ and $\beta^{(2)}$, we obtain

$$\begin{aligned} &\mathbb{E}\left[g(x) \mid X_1, X_2 \right] \\ =&((x-1)^2\kappa^2 + (x^2+1)d^2) p^{-1} \operatorname{Tr}\Big[(X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} \Big] + \sigma^2 (n_1 + n_2 - p), \end{aligned} \tag{D.4}$$

Part 1: characterizing the global minimum of $f(A, B)$. Let $\hat{x}$ denote the global minimizer of $g(x)$. We show that in the setting of Corollary 3.3, $\hat{x}$ is close to 1. This gives us the global minimum of $f(A, B)$, since $\hat{B}$ is given by $\hat{x}$ using local optimality conditions. First, we show that $g(x)$ and its expectation are close using standard concentration bounds.

Claim D.1. *In the setting of Corollary 3.3, for any x, we have that with high probability*

$$\Big| g(x) - \mathbb{E}\Big[g(x) \mid X^{(1)}, X^{(2)} \Big] \Big| \leqslant p^{1/2+c} \left(\sigma^2 + \kappa^2 + d^2 \right).$$

Proof. There are two terms in $g(A)$ from equation (D.3). We will focus on dealing with the concentration error of the first term. The second term is similar to the first and we omit the details. For the first term, we expand into several equations under various situations involving the random noise and the random-effect model.

$$\begin{aligned} &\Big\| X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} (x\beta^{(2)} - \beta^{(1)}) + \Big(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1 \times n_1} \Big) \varepsilon^{(1)} \\ &+ x X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top \varepsilon^{(2)} \Big\|^2 = h_1(x) + h_2(x) + h_3(x) + 2h_4(x) + 2h_5(x) + 2h_6(x), \end{aligned} \tag{D.5}$$

where

$$\begin{aligned} h_1(x) &:= (\beta^{(1)} - x\beta^{(2)})^\top (X^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} (\beta^{(1)} - x\beta^{(2)}), \\ h_2(x) &:= (\varepsilon^{(1)})^\top \Big(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1 \times n_1} \Big)^2 \varepsilon^{(1)}, \\ h_3(x) &:= x^2 (\varepsilon^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top \varepsilon^{(2)}, \\ h_4(x) &:= (\varepsilon^{(1)})^\top \Big(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1 \times n_1} \Big) X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} (x\beta^{(2)} - \beta^{(1)}), \\ h_5(x) &:= x (\varepsilon^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} (x\beta^{(2)} - \beta^{(1)}), \\ h_6(x) &:= x (\varepsilon^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top \Big(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1 \times n_1} \Big) \varepsilon^{(1)}. \end{aligned}$$

Next, we estimate each term using Lemma C.6 for random variables with bounded moment up to any order. We first state several facts that will be commonly used in the proof. By Fact E.1 (ii), we have that w.h.p. the operator norm of $X^{(1)}$ and $X^{(2)}$ are both bounded by $O(\sqrt{n})$. Furthermore, the operator norm of $\hat{\Sigma}(x)^{-1}$ is bounded by $(x^2+1)^{-1}O(n_1+n_2) = (x^2+1)^{-1}O(p)$.

For $h_1(x)$, using Lemma C.6 and the fact that the entries of $(\beta^{(1)} - x\beta^{(2)}) \in \mathbb{R}^p$ are i.i.d. Gaussian random variables with mean zero and variance $b = p^{-1}((x-1)^2\kappa^2 + (x^2+1)d^2)$, we obtain the following estimate w.h.p.

$$\begin{aligned}
&\left|h_1(x) - \mathop{\mathbb{E}}_{\beta^{(1)},\beta^{(2)}}[h_1(x) \mid X_1, X_2]\right| \\
\leqslant & p^c \cdot p^{-1} b \cdot \left\|(X^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)}\right\|_F \\
\leqslant & p^c \cdot p^{-1} b \cdot p^{1/2} \left\|(X^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)}\right\| \\
\lesssim & p^{1/2+c} \cdot \frac{b}{(x^2+1)^2} \lesssim p^{1/2+c}(\kappa^2 + d^2).
\end{aligned} \tag{D.6}$$

In the third step we use the operator norm bound of $X^{(1)}$, $X^{(2)}$, and $\hat{\Sigma}(x)^{-1}$.

For $h_2(x)$ and $h_3(x)$, since the entries of $\varepsilon^{(1)}, \varepsilon^{(2)} \in \mathbb{R}^p$ are i.i.d. Gaussian random variables with mean zero and variances σ^2, using Lemma C.6, we obtain w.h.p.

$$\left|h_2(x) - \mathop{\mathbb{E}}_{\varepsilon^{(1)}}[h_2(x) \mid X_1, X_2]\right| \lesssim p^{1/2+c}\sigma^2, \quad \left|h_3(x) - \mathop{\mathbb{E}}_{\varepsilon^{(2)}}[h_3(x) \mid X_1, X_2]\right| \lesssim p^{1/2+c}\sigma^2. \tag{D.7}$$

For $h_4(x)$, using Lemma C.6, we obtain w.h.p.:

$$\begin{aligned}
&|h_4(x)| \\
\leqslant & p^c \cdot \sigma \cdot \sqrt{b/p} \left\|\left(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1 \times n_1}\right) X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)}\right\|_F \\
\leqslant & p^c \cdot \sigma \sqrt{b/p} \cdot p^{1/2} \left\|\left(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1 \times n_1}\right) X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)}\right\| \\
\leqslant & p^c \cdot \sigma \sqrt{b} \cdot \left\|x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1 \times n_1}\right\| \cdot \left\|X^{(1)}\right\| \cdot \left\|\hat{\Sigma}(x)^{-1}\right\| \cdot \left\|(X^{(2)})^\top X^{(2)}\right\| \\
\lesssim & p^{1/2+c} \frac{\sigma\sqrt{b}}{x^2+1} \lesssim p^{1/2+c}(\sigma^2 + \kappa^2 + d^2).
\end{aligned} \tag{D.8}$$

Above, in the fourth step we use the operator norm of $x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}$ being at most one and the operator norm bound of $X^{(1)}$, $X^{(2)}$, and $\hat{\Sigma}(x)^{-1}$. In the last step we use AM-GM inequality. Using the same argument, we can show that $|h_5(x)| \leqslant p^{1/2+c}(\sigma^2+\kappa^2+d^2)$ and $|h_6(x)| \leqslant p^{1/2+c}\sigma^2$. Combining the concentration error bound for $h_1(x), h_2(x), \ldots, h_6(x)$, we complete the proof. The second term of $g(A)$ can be dealt in similar ways and we omit the details. □

Next, we show that the global minimizer $\hat{x}$ of $g(x)$ is close to 1.

Claim D.2. *Let c be a sufficiently small fixed constant. In the setting of Corollary 3.3, we have that with high probability,*

$$|\hat{x} - 1| \leqslant \frac{2d^2}{\kappa^2} + p^{-1/4+c}. \tag{D.9}$$

Proof. Corresponding to equation (D.4), we define the function

$$\begin{aligned}
h(x) =& [(x-1)^2\kappa^2 + (x^2+1)d^2] \cdot p^{-1} \operatorname{Tr}\left[(X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)}\right] \\
=& [(1-x^{-1})^2\kappa^2 + (1+x^{-2})d^2] \cdot p^{-1} \operatorname{Tr}\left[\left([(X^{(2)})^\top X^{(2)}]^{-1} + x^{-2}[(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1}\right].
\end{aligned}$$

Let $x^\star$ denote the global minimizer of $h(x)$. Our proof involves two steps. First, we will show that $|x^\star - 1| \leqslant d^2/\kappa^2$. Second, we will use Claim D.1 to show that the global minimizer of $g(x)$ and $h(x)$ are close to each other.

For the first step, it is easy to observe that $h(x) < h(-x)$ for any positive x. Hence the minimum of $h(x)$ is achieved when x is positive. Next, we consider the case where $x \geqslant 1$. Notice that the following function always increases when x increases in the positive orthant:

$$\operatorname{Tr}\left[\left([(X^{(2)})^\top X^{(2)}]^{-1} + x^{-2}[(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1}\right]$$

By taking the derivative of $h(x)$, we obtain that for any $x > 1 + d^2/\kappa^2$,

$$h'(x) \geqslant \left[2(1-x^{-1})\frac{\kappa^2}{x^2} - 2\frac{d^2}{x^3}\right] \cdot p^{-1} \operatorname{Tr}\left[\left([(X^{(2)})^\top X^{(2)}]^{-1} + x^{-2}[(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1}\right] > 0, \tag{D.10}$$

Finally, we consider the case where $x \leqslant 1$. Notice that the following function always decreases when x decreases from 1:

$$\operatorname{Tr}\left[\left(x^2[(X^{(2)})^\top X^{(2)}]^{-1} + [(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1}\right].$$

Hence, by taking derivative of $h(x)$, we obtain that for any $x \leqslant 1 - d^2/\kappa^2$,

$$h'(x) \leqslant [-2(1-x)\kappa^2 + 2xd^2] \cdot p^{-1} \operatorname{Tr}\left[\left(x^2[(X^{(2)})^\top X^{(2)}]^{-1} + [(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1}\right] < 0, \tag{D.11}$$

In summary, the global minimizer of $h(x)$ lies within $1 - d^2/\kappa^2$ and $1 + d^2/\kappa^2$.

For the second step, using Claim D.1, we have that $g(x)$ and $h(x)$ differ by at most $p^{1/2+c}(\sigma^2 + \kappa^2 + d^2)$. Therefore, our goal reduces to showing that if $\hat{x}$ deviates too far from $1 \pm d^2/\kappa^2$, it is no longer a global minimum of $g(x)$. We prove by contradiction. First, suppose that $\hat{x} \geqslant 1 + 2d^2/\kappa^2 + p^{-1/2+c}$. For any $x \geqslant 1 + 3d^2/(2\kappa^2)$, we can lower bound the derivative of $h(x)$ using equation (D.10) as follows

$$h'(x) \geqslant \frac{2(x-1)\kappa^2 - 2d^2}{x^3} \cdot p^{-1} \operatorname{Tr}\left[\left([(X^{(2)})^\top X^{(2)}]^{-1} + x^{-2}[(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1}\right] \gtrsim p\kappa^2 \cdot \frac{x-1}{x(1+x^2)}.$$

Therefore, the difference between $h(x)$ and $h(1)$ is at least the following

$$h(x) - h(1) \geqslant h(x) - h\left(1 + \frac{3d^2}{2\kappa^2}\right) \geqslant \int_{1+\frac{3d^2}{2\kappa^2}}^{x} h'(x)\mathrm{d}x \gtrsim p\kappa^2 \cdot \int_{1+\frac{3d^2}{2\kappa^2}}^{x} \frac{x-1}{x(1+x^2)}\mathrm{d}x.$$

When $\hat{x}$ is sufficiently far from 1 (e.g. $2d^2/\kappa^2 + p^{-1/4+c}$), one can verify that $h(x) - h(1)$ is at least $\mathrm{O}(p\frac{d^4}{\kappa^2} + p^{1/2+2c}\kappa^2) > \mathrm{O}(p^{1/2+c}(\sigma^2 + \kappa^2 + d^2))$, under the condition that $\sigma^2 = \mathrm{O}(\kappa^2)$ and $d^2 = \mathrm{o}(\kappa^2)$. On the other hand, by triangle inequality and Claim D.1 we have that

$$h(\hat{x}) - h(1) = g(\hat{x}) - g(1) + (h(\hat{x}) - g(\hat{x})) + (g(1) - h(1)) \leqslant \mathrm{O}\left(p^{1/2+c}\left(\sigma^2 + \kappa^2 + d^2\right)\right).$$

Hence, we have arrived at a contradition.

Second, suppose that $\hat{x} \leqslant 1 - 2d^2/\kappa^2 - p^{-1/2+c}$. Using equation (D.11), we obtain that for any $x \leqslant 1 - 3d^2/(2\kappa^2)$,

$$-h'(x) \geqslant [2(1-x)\kappa^2 - 2xd^2] \cdot p^{-1} \operatorname{Tr}\left[\left(x^2[(X^{(2)})^\top X^{(2)}]^{-1} + [(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1}\right] \gtrsim p\kappa^2 \cdot \frac{(1-x)x^2}{1+x^2}.$$

Using a similar argument to the first case, we get that the difference between $h(x)$ and $h(1)$ is at least the integral of the abvoe derivative. This implies that $\hat{x}$ cannot be too far from one. Hence we have completed the proof. □

Part 2: a reduction to the bias and variance limits. Recall that the hard parameter sharing estimator $\hat{\beta}_2^{\mathrm{HPS}}$ is equal to $\hat{B}\hat{A}_2$. Using the local optimility condition for $\hat{B}$, we obtain the predication loss of HPS as follows

$$\begin{aligned} L(\hat{\beta}_2^{\mathrm{HPS}}) &= \left\|(\Sigma^{(2)})^{1/2}\left(\hat{B}\hat{A}_2 - \beta^{(2)}\right)\right\| \\ &= \left\|(\Sigma^{(2)})^{1/2}\hat{\Sigma}(\hat{x})^{-1}\left[(X^{(1)})^\top X^{(1)}(\hat{x}\beta^{(1)} - \hat{x}^2\beta^{(2)}) + (X^{(2)})^\top \varepsilon^{(2)} + \hat{x}(X^{(1)})^\top \varepsilon^{(1)}\right]\right\|^2. \end{aligned} \tag{D.12}$$

Using Lemma D.2 and the concentration estimates in Lemma C.6, we simplify $L(\hat{\beta}_2^{\mathrm{HPS}})$ as follows.

Claim D.3. *Recall that $\hat{\Sigma}(1)$ is equal to $\hat{\Sigma}$ (cf. equation (3.1)). In the setting of Claim 3.3, we have the following estimate w.h.p.*

$$\begin{aligned}&\left|L(\hat{\beta}_2^{\mathrm{HPS}}) - \frac{2d^2}{p}\operatorname{Tr}\left[\hat{\Sigma}^{-2}\left((X^{(1)})^\top X^{(1)}\right)^2\right] - \sigma^2\operatorname{Tr}\left[\hat{\Sigma}^{-1}\right]\right|\\ \lesssim&\frac{d^4+\sigma^2 d^2}{\kappa^2} + p^{-1/2+2c}\kappa^2 + p^{-1/4+c}(\sigma^2+d^2).\end{aligned}$$

Proof. Our proof is divided into two steps. First, using Lemma C.6, we show that

$$\left|L(\hat{\beta}_2^{\mathrm{HPS}}) - \mathcal{L}(\hat{x})\right| \leqslant p^{-1/2+c}\left(\sigma^2+\kappa^2+d^2\right), \tag{D.13}$$

where $\mathcal{L}(\hat{x})$ is defined as

$$\mathcal{L}(\hat{x}) := \hat{x}^2\left[(\hat{x}-1)^2\kappa^2 + (\hat{x}^2+1)d^2\right]\cdot p^{-1}\operatorname{Tr}\left[(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(1)})^\top X^{(1)}\right] + \sigma^2\cdot\operatorname{Tr}\left[\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}\right].$$

Next, we further simplify $\mathcal{L}(\hat{x})$ since $\hat{x}$ is close to one and $\Sigma^{(1)}, \Sigma^{(2)}$ are both isotropic

$$\begin{aligned}&\left|\mathcal{L}(\hat{x}) - \frac{2d^2}{p}\operatorname{Tr}\left[\hat{\Sigma}^{-2}\left((X^{(1)})^\top X^{(1)}\right)^2\right] - \sigma^2\operatorname{Tr}\left[\hat{\Sigma}^{-1}\right]\right|\\ \lesssim&\frac{d^4+\sigma^2 d^2}{\kappa^2} + p^{-1/2+2c}\kappa^2 + p^{-1/4+c}(\sigma^2+d^2).\end{aligned}\tag{D.14}$$

Combining equation (D.13) and (D.14), we obtain the desired claim. We prove these two equations one by one as follows.

First, we prove equation (D.14). We can bound the left hand side of equation (D.14) as

$$\begin{aligned}&\left|\mathcal{L}(\hat{x}) - \frac{2d^2}{p}\operatorname{Tr}\left[\hat{\Sigma}^{-2}\left((X^{(1)})^\top X^{(1)}\right)^2\right] - \sigma^2\operatorname{Tr}\left(\hat{\Sigma}^{-1}\right)\right|\\ \lesssim&\left(|\hat{x}-1|^2\kappa^2 + |\hat{x}-1|d^2\right)\cdot p^{-1}\operatorname{Tr}\left[\hat{\Sigma}^{-2}\left((X^{(1)})^\top X^{(1)}\right)^2\right]\\ &+\frac{d^2}{p}\left|\operatorname{Tr}\left[\left(\hat{\Sigma}(\hat{x})^{-2}-\hat{\Sigma}^{-2}\right)\left((X^{(1)})^\top X^{(1)}\right)^2\right]\right| + \sigma^2\left|\operatorname{Tr}\left[\hat{\Sigma}(\hat{x})^{-1}-\hat{\Sigma}^{-1}\right]\right|.\end{aligned}$$

We deal with the trace terms in the above equation one by one. Using Claim D.2 and operator norm bound of $X^{(1)}$, $X^{(2)}$, and $\hat{\Sigma}(x)$, we have that w.h.p.

$$\|\hat{\Sigma}^{-1}-\hat{\Sigma}(\hat{x})^{-1}\| \leqslant |\hat{x}^2-1|\cdot\|\hat{\Sigma}^{-1}\|\cdot\|(X^{(1)})^\top X^{(1)}\|\cdot\|\hat{\Sigma}(\hat{x})^{-1}\| \lesssim p^{-1}\left(\frac{d^2}{\kappa^2}+p^{-1/4+c}\right). \tag{D.15}$$

Using similar arguments, we get that w.h.p.

$$\left\|\left(\hat{\Sigma}^{-2}-\hat{\Sigma}(\hat{x})^{-2}\right)\left((X^{(1)})^\top X^{(1)}\right)^2\right\| \lesssim \frac{d^2}{\kappa^2}+p^{-1/4+c}, \tag{D.16}$$

and

$$\operatorname{Tr}\left[\hat{\Sigma}^{-2}\left((X^{(1)})^\top X^{(1)}\right)^2\right] \leqslant p\left\|\hat{\Sigma}^{-2}\left((X^{(1)})^\top X^{(1)}\right)^2\right\| \lesssim p. \tag{D.17}$$

Applying the above results (D.15), (D.16), and (D.17) to the bound of $\mathcal{L}(\hat{x})$ above, we have shown that equation (D.14) holds.

Second, we prove equation (D.13). The proof is very similar to Claim D.1. The key difference is that $\hat{x}$ correlates with $\varepsilon^{(1)}$, $\varepsilon^{(2)}$, $\beta^{(1)}$, and $\beta^{(2)}$. Nevertheless, Lemma C.6 still applies for any arbitrary $\hat{x}$. We describe a proof sketch and omit the details. Recall that β_0 is the shared component of $\beta^{(1)}$ and $\beta^{(2)}$ with i.i.d. Gaussian entries of mean zero and variance $p^{-1}\kappa^2$. The task-specific components, denoted by $\widetilde{\beta}^{(1)}$ and $\widetilde{\beta}^{(2)}$, consist of i.i.d. Gaussian random variables with mean zero and variance $p^{-1}d^2$. We write $L(\hat{\beta}_2^{\mathrm{HPS}})$ from equation (D.12) as:

$$\begin{aligned} L(\hat{\beta}_2^{\mathrm{HPS}}) = \Big\|&(\Sigma^{(2)})^{1/2}\hat{\Sigma}(\hat{x})^{-1}\left[(X^{(1)})^\top X^{(1)}(\hat{x}-\hat{x}^2)\beta_0 + (X^{(1)})^\top X^{(1)}\hat{x}\widetilde{\beta}^{(1)} - (X^{(1)})^\top X^{(1)}\hat{x}^2\widetilde{\beta}^{(2)}\right] \\ &+(\Sigma^{(2)})^{1/2}\hat{\Sigma}(\hat{x})^{-1}\left[(X^{(2)})^\top\varepsilon^{(2)} + \hat{x}(X^{(1)})^\top\varepsilon^{(1)}\right]\Big\|^2. \end{aligned} \tag{D.18}$$

Similar to the analysis of $g(x)$, we expand $L(\hat{\beta}_2^{\mathrm{HPS}})$ into the sum of 15 terms, and bound the concentration error of each term similar to $h_1(x),\dots,h_6(x)$. For example, for the leading term $\hat{x}^2(\widetilde{\beta}^{(1)})^\top(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(1)})^\top X^{(1)}\widetilde{\beta}^{(1)}$, using Lemma C.6 and the operator norm bounds, we obtain the following estimate w.h.p.

$$\begin{aligned} &\left|(\widetilde{\beta}^{(1)})^\top(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(1)})^\top X^{(1)}\widetilde{\beta}^{(1)} - \frac{d^2}{p}\operatorname{Tr}\left[(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(1)})^\top X^{(1)}\right]\right| \\ &\leqslant p^{-1+c}d^2\cdot\left\|(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(1)})^\top X^{(1)}\right\|_F \\ &\leqslant p^{-1/2+c}d^2\cdot\left\|(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(1)})^\top X^{(1)}\right\| \lesssim p^{-1/2+c}d^2. \end{aligned}$$

For the cross term $\hat{x}(\widetilde{\beta}^{(1)})^\top(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(2)})^\top\varepsilon^{(2)}$, using Lemma C.6 and the operator norm bounds, we obtain the following estimate w.h.p.

$$\begin{aligned} \left|(\widetilde{\beta}^{(1)})^\top(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(2)})^\top\varepsilon^{(2)}\right| &\leqslant p^c\cdot\sigma\sqrt{p^{-1}d^2}\cdot\|(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(2)})^\top\|_F \\ &\lesssim p^c\cdot\sigma d\cdot\|(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(2)})^\top\| \\ &\lesssim p^{-1/2+c}\sigma d \leqslant p^{-1/2+c}(\sigma^2+d^2). \end{aligned}$$

The rest of the terms in the expansion of $L(\hat{\beta}_2^{\mathrm{HPS}})$ can be dealt with similarly, and we omit the details. □

Part 3: applying the bias-variance limits. Finally, we are ready to complete the proof of Corollary 3.3. We derive the variance term $\sigma^2\operatorname{Tr}[\hat{\Sigma}^{-1}]$ and the bias term $\frac{2d^2}{p}\operatorname{Tr}\left[\hat{\Sigma}^{-2}\left((X^{(1)})^\top X^{(1)}\right)^2\right]$ using our random matrix theory results.

Proof of Corollary 3.3. For the variance term, using equation (3.4), we obtain that

$$\operatorname{Tr}[\hat{\Sigma}^{-1}] = \operatorname{Tr}\left[\left((X^{(1)})^\top X^{(1)} + (X^{(2)})^\top X^{(2)}\right)^{-1}\right] = \operatorname{Tr}\left[\frac{(a_1+a_2)^{-1}\,\mathrm{Id}_{p\times p}}{n_1+n_2}\right] + \mathrm{O}(p^{-c_\varphi}) \tag{D.19}$$

with high probability. Solving equation (3.6) with $\lambda_i \equiv 1$, $1\leqslant i\leqslant p$, we get that

$$a_1 = \frac{n_1(n_1+n_2-p)}{(n_1+n_2)^2},\quad a_2 = \frac{n_2(n_1+n_2-p)}{(n_1+n_2)^2}. \tag{D.20}$$

Applying the above to equation (D.19), we obtain that

$$\operatorname{Tr}[\hat{\Sigma}^{-1}] = \frac{p}{n_1+n_2}\cdot\frac{n_1+n_2}{n_1+n_2-p} + \mathrm{O}(p^{-c_\varphi}) = \frac{p}{n_1+n_2-p} + \mathrm{O}(p^{-c_\varphi}) \tag{D.21}$$

with high probability.

For the bias term, since the spectrum of $(X^{(1)})^\top X^{(1)}$ is tightly concentrated by Fact E.1, we have that

$$\begin{aligned} \frac{(\sqrt{n_1}-\sqrt{p})^4\cdot(1-p^{-c_\varphi})}{p}\operatorname{Tr}\left[\hat{\Sigma}^{-2}\right] &\leqslant p^{-1}\operatorname{Tr}\left[\hat{\Sigma}^{-2}\left((X^{(1)})^\top X^{(1)}\right)^2\right] \\ &\leqslant \frac{(\sqrt{n_1}+\sqrt{p})^4\cdot(1+p^{-c_\varphi})}{p}\operatorname{Tr}\left[\hat{\Sigma}^{-2}\right]. \end{aligned} \tag{D.22}$$

Using the bias limit (C.1) with $\Sigma^{(1)} = \Sigma^{(2)} = \Lambda = V = \mathrm{Id}_{p\times p}$, and $w = e_i$ (the i-th coordinate vector), we have w.h.p. (via a union bound)

$$e_i^\top \hat{\Sigma}^{-2} e_i = \frac{1}{(n_1+n_2)^2}\left[\frac{a_3+a_4+1}{(a_1+a_2)^2} + \mathrm{O}(p^{-c_\varphi})\right], \text{ for all } i = 1,2,\dots,p.$$

We solve the self-consistent equations (C.2) given a_1, a_2, and obtain

$$a_3 = \frac{p\cdot n_1}{(n_1+n_2)(n_1+n_2-p)}, \quad a_4 = \frac{p\cdot n_2}{(n_1+n_2)(n_1+n_2-p)}.$$

Applying a_3, a_4 to the equation above, we obtain

$$e_i^\top \hat{\Sigma}^{-2} e_i = \frac{1}{(n_1+n_2)^2}\left[\frac{(n_1+n_2)^3}{(n_1+n_2-p)^3} + \mathrm{O}(p^{-c_\varphi})\right], \text{ for all } i = 1,2,\dots,p.$$

Applying the above result to equation (D.22), we get the desired result for the bias term. Combining the bias and variance estimates, we get that

$$\left|L(\hat{\beta}_2^{\mathrm{HPS}}) - \frac{2d^2 n_1^2(n_1+n_2)}{(n_1+n_2-p)^3} - \frac{\sigma^2 p}{n_1+n_2-p}\right| \leqslant \left[\left(1+\sqrt{\frac{p}{n_1}}\right)^4 - 1\right]\cdot\frac{2d^2n_1^2(n_1+n_2)}{(n_1+n_2-p)^3}$$
$$+ \mathrm{O}\left(p^{-c_\varphi}(\sigma^2+d^2) + \frac{d^4+\sigma^2 d^2}{\kappa^2} + p^{-1/2+2c}\kappa^2 + p^{-1/4+c}(\sigma^2+d^2)\right).$$

Since $\sigma^2 \lesssim \kappa^2$ and $d^2 \leqslant p^{-c}\kappa^2$ by our assumption, we obtain the desired result. The proof is complete. □

E Random Matrices with Bounded Moments

We state several concentration results for dealing with random matrices with bounded moments. Recall from Section 2 that we consider random matrices $Z \in \mathbb{R}^{n\times p}$ whose entries are i.i.d. with zero mean, unit variance, and bounded φ-th moment. The following facts are well-known.

Fact E.1. Suppose Assumption 2.2 holds. With probability $1 - 1/\mathrm{poly}(n)$ over the randomness of Z, we have:

(i) When n/p converges to a fixed $\rho > 1$, the sample covariance matrix $X^\top X/n$ has full rank p.

(ii) The singular values of $Z^\top Z$ are greater than $(\sqrt{n}-\sqrt{p})^2 - n\cdot p^{-c_\varphi}$ and less than $(\sqrt{n}+\sqrt{p})^2 + n\cdot p^{-c_\varphi}$, cf. Bloemendal et al. (2014, Theorem 2.10) and Ding and Yang (2018, Lemma 3.12).

Next, we state a concentration result for Z.

Corollary E.2. *For any deterministic vector $v \in \mathbb{R}^p$, we have that w.h.p.*

$$\left|\|Zv\|^2 - n\|v\|^2\right| \leqslant 2n^{1-c_\varphi}\|v\|^2. \tag{E.1}$$

Proof. Let $Q = n^{\frac{2}{\varphi}}\log n$. We introduce a truncated matrix $\widetilde{Z}$ with entries $\widetilde{Z}_{ij} := \mathbf{1}(|Z_{ij}| \leqslant Q)\cdot Z_{ij}$. Then $\widetilde{Z}$ is equal to Z when all the entries of Z are smaller than Q. Using Markov's inequality and a simple union bound (cf. equation (C.15)), this happens with probability

$$\mathbb{P}(\widetilde{Z} = Z) = 1 - \mathrm{O}((\log n)^{-\varphi}). \tag{E.2}$$

Furthermore, using the finite φ-th moment condition and the tail probabilities, we can show that the mean and variance of $Z - \widetilde{Z}$ are small, which gives that (cf. equation (C.28))

$$|\mathbb{E}\widetilde{Z}_{ij}| = \mathrm{O}(n^{-3/2}), \quad \mathbb{E}|\widetilde{Z}_{ij}|^2 = 1 + \mathrm{O}(n^{-1}). \tag{E.3}$$

We centralize and rescale $\widetilde{Z}$ as $\widehat{Z} := \frac{\widetilde{Z}-\mathbb{E}\widetilde{Z}}{(\mathbb{E}|\widetilde{Z}_{11}|^2)^{1/2}}$. Let c be a sufficiently small fixed constant. To prove the result, it suffices to show that

$$\left|\|\widehat{Z}v\|^2 - n\|v\|^2\right| \leqslant n^{1/2+c}Q\|v\|^2. \tag{E.4}$$

This is because provided with equation (E.3) and (E.4), we can get that

$$\left| \|\widetilde{Z}v\|^2 - n\|v\|^2 \right| \leqslant n^{1/2+c} Q \|v\|^2,$$

which implies the desired our (recall that $c_\varphi < 1/2 - c$). To prove equation (E.4), we first show that for any $i = 1, 2, \ldots, n$, $(\widehat{Z}v)_i = \sum_{1 \leqslant j \leqslant p} \widehat{Z}_{ij} v_j$ is at most $2n^c Q$. Since $\widehat{Z}v$ consists of i.i.d random variables with mean zero and variance $\|v\|^2$, using equation (C.19) from Lemma C.6, we get that with overwhelming probability

$$|(\widehat{Z}v)_i| \leqslant n^c Q \max_{1 \leqslant i \leqslant p} |v_i| + n^c \|v\| \leqslant 2n^c Q.$$

Hence, $\frac{(\widehat{Z}v)}{\|v\|}$ consists of i.i.d random variables with mean zero, unit variance, and bounded support $2n^c Q$. Applying equation (C.20), we get that

$$\left| \|\widehat{Z}v\|^2 - n\|v\|^2 \right| = \left| \sum_i \left(|(\widehat{Z}v)_i|^2 - \mathbb{E}|(\widehat{Z}v)_i|^2 \right) \right| \leqslant 2n^{1/2+2c} Q \|v\|^2.$$

Hence the proof is complete. □

Next, we provide several concentration results for $\mathcal{E} = [\varepsilon^{(1)}, \varepsilon^{(2)}, \cdots, \varepsilon^{(t)}] \in \mathbb{R}^{n \times t}$, which consists of i.i.d. random variables with mean zero, variance σ^2 and bounded moments up to any order.

Corollary E.3. *Let $c > 0$ be a sufficiently small fixed constant. For any deterministic vector $v \in \mathbb{R}^n$, we have that w.h.p.*

$$|v^\top \varepsilon^{(i)}| \leqslant n^c \cdot \sigma \|v\|, \text{ and } \|v^\top \mathcal{E}\| \leqslant n^c \cdot \sigma \|v\|. \tag{E.5}$$

For any deterministic matrix $B \in \mathbb{R}^{n \times n}$, we have that w.h.p.

$$\left| (\varepsilon^{(i)})^\top B \varepsilon^{(j)} - \delta_{ij} \cdot \sigma^2 \operatorname{Tr}(B) \right| \leqslant n^c \cdot \sigma^2 \|B\|_F, \text{ and } \left\| \mathcal{E}^\top B \mathcal{E} - \sigma^2 \operatorname{Tr}[B] \cdot \mathrm{Id}_{t \times t} \right\|_F \leqslant n^c \cdot \sigma^2 \|B\|_F. \tag{E.6}$$

Proof. Rescaling $\varepsilon^{(i)}$ by σ, we get a random vector with zero mean, unit variance, and bounded moments up to any order. Using the first estimate in equation (C.21) of Lemma C.6 (recall that the stochastic domination notation means the inequality holds with a multiplicative factor of p^c on the right), we obtain that equation (E.5) holds (the second result is a consequence of the first). Using the second estimate in equation (C.21), we obtain that for $i \neq j$,

$$\left| (\varepsilon^{(i)})^\top B \varepsilon^{(j)} \right| = \Big| \sum_{k,l=1}^n \varepsilon_k^{(i)} \varepsilon_l^{(j)} B_{kl} \Big| \leqslant p^c \cdot \sigma^2 \Big(\sum_{k,l=1}^n |B_{kl}|^2 \Big)^{1/2} = \sigma^2 \|B\|_F.$$

Using the two estimates in equation (C.22), we obtain that

$$\begin{aligned} \left| (\varepsilon^{(i)})^\top B \varepsilon^{(i)} - \sigma^2 \operatorname{Tr}[B] \right| &\leqslant \Big| \sum_{k=1} \left(|\varepsilon_k^{(i)}|^2 - \mathbb{E}|\varepsilon_k^{(i)}|^2 \right) B_{ii} \Big| + \Big| \sum_{k \neq l} \varepsilon_k^{(i)} \varepsilon_l^{(i)} B_{kl} \Big| \\ &\leqslant p^c \cdot \sigma^2 \Big(\sum_k |B_{kk}|^2 \Big) + p^c \cdot \sigma^2 \Big(\sum_{k \neq l} |B_{kl}|^2 \Big)^{1/2} \leqslant n^c \cdot \sigma^2 \|B\|_F. \end{aligned}$$

Hence, we have shown equation (E.6) (the second result is again a consequence of the first result). □