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

October 13, 2020

Abstract

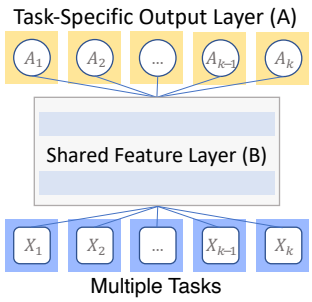
Hard parameter sharing for multi-task learning is widely used in empirical research despite the fact that their generalization properties have not been well established in many cases. This paper studies a fundamental question to better understand this approach: How does hard parameter sharing work given 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 become increasingly large in 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 the variance but hurts eventually due to the increasing bias. This suggests progressively adding data for optimizing hard parameter sharing, and we validate its efficiency in text classification tasks.

1 Introduction

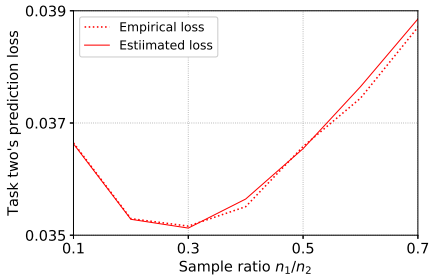
Hard parameter sharing (HPS) for multi-task learning is widely used in empirical research and goes back to the seminal work of Caruana (1997). Recent work has revived interest in this approach because it improves performance and reduces the cost of collecting labeled data (Liu et al., 2019; Zamir et al., 2018). 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. (i) It reduces model parameters since all tasks use the same feature space. (ii) It reduces the amount of labeled data needed from each task by augmenting the entire training dataset.

Hard parameter sharing offers great intuitive appeal as an inductive transfer mechanism. It reduces overfitting by acting as a regularizer (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 impacts HPS has not been established. Part of the challenge may be that HPS' generalization performance depends intricately on the sample size ratios and covariate shifts between tasks, not amenable to standard concentration results. Previous results based on Rademacher complexity or VC dimensions have considered when all tasks' sample sizes are equal to logarithm factors of 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 establish 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 at a fixed ratio. This is motivated by many multi-task learning applications, where the amount of labeled data from each dataset is usually insufficient for learning a single task. For example, this is the case if a dataset's sample size is only a small constant factor of the feature dimension. The high-dimensional setting is challenging but is crucial for understanding how datasets' sample sizes impact generalization performance.



(a) A hard parameter sharing architecture



(b) Varying sample size ratio

Figure 1: An illustrative example of our result: Consider the prediction loss of hard parameter sharing (left) for task two, given two linear regression tasks. Increasing task one's sample size decreases task two's prediction loss initially, but increases afterwards. This phenomenon occurs due to different bias-variance tradeoffs as the sample size ratio 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

We assume that there are multiple datasets that all follow (potentially different) linear models. In each dataset, suppose the feature vector is $x = \Sigma^{1/2} z$, where $z \in \mathbb{R}^p$ has i.i.d entries with zero mean and unit variance, and the population covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ is deterministic and positive semidefinite.

We study a hard parameter sharing architecture to learn jointly from multiple datasets. In this architecture, there is a shared feature representation layer $B \in \mathbb{R}^{p \times r}$ for all tasks and a separate output layer $A_i \in \mathbb{R}^r$ for every task i. Suppose we have t datasets. For each dataset i from 1 to t, let n_i denote its sample size, $X^{(i)} \in \mathbb{R}^{n_i \times p}$ denote its feature covariates, and $Y^{(i)} \in \mathbb{R}^{n_i}$ denote all the labels. Let $A = [A_1, A_2, \ldots, A_t] \in \mathbb{R}^{r \times t}$. We study the following optimization objective.

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

Given a solution $(\hat{A}, \hat{B})$ to the above optimization problem, let $\hat{\beta}_i^{\text{HPS}} = \hat{B}\hat{A}_i$ denote the hard parameter sharing (HPS) estimator for task i. Some 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 the multi-label prediction setting where all datasets have the same features, and we want to make multiple predictions (see e.g. (Hsu et al., 2009)). We analyze the global minimizer of $f(A, B)$, and provide a sharp generalization bound on its (out-of-sample) prediction loss for any task. This case 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 $X^{(i)}, Y^{(i)}$). We show that the prediction loss of hard parameter sharing admits a clean bias-variance decomposition. Our results imply that hard parameter sharing helps by reducing variance compared to single-task learning, but hurts by increasing bias. Our second result (Theorem 3.1) applies to two tasks with arbitrarily different sample size ratios and population covariance matrices. We analyze the local minimizers of $f(A, B)$ and provide a sharp generalization bound on both tasks' prediction loss. Despite its simplicity, we show several interesting phenomena by varying sample sizes and covariate shifts in this setting. See Figure 1 for an illustration.

Consequently, using our precise loss estimates, we observe qualitative properties of hard parameter sharing for varying datasets' properties.

(i) *Sample efficiency (Example 2.4)*: One advantage of combining multiple datasets is that the requirement for labeled data reduces compared to STL, a phenomenon that Zamir et al. (2018) has observed empirically. Our results further imply that HPS's sample efficiency depends on model-specific variance

vs. noise variance. It is generally high when the noise variance is large compared to model-specific variance across different tasks.

(ii) *Sample size ratio (Example 3.2)*: Increasing one task's sample size does not always help to reduce another task's loss. In a simplified setting, we find that the task loss either decreases first before increasing afterwards, or decreases monotonically depending on how fast the bias increases. These two trends result from different tradeoffs between increasing bias and decreasing variance.

(iii) *Covariate shift (Example 3.4)*: In addition to sample sizes, variance also scales with the covariate shift between different datasets. 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.

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 minimum of $f(A, B)$ and its population version. The proof of our second result is significantly more involved because of different sample sizes and covariate shifts. Using recently developed techniques from random matrix theory (Knowles and Yin, 2016), 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(i)). One limitation of our analysis is that for two tasks with different sample sizes, there is an extra error term in the prediction loss of hard parameter sharing (cf. equation (3.8)), which can be large for very small n_1. This requires studying the spectrum of non-symmetric random matrices, which is an intriguing open question (see Section 6 for more discussion).

Finally, we discuss the practical implications of 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.

1.2 Related Work

There is a large body of both classical and recent works on multi-task learning. We focus our discussion on theoretical works, and refer interested readers to several excellent surveys for general references (Pan and Yang, 2009; Ruder, 2017; Zhang and Yang, 2017; Vandenhende et al., 2020). The early work of Baxter (2000); Ben-David and Schuller (2003); Maurer (2006) have sought for a study of 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 bound that combines 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). Moreover, these techniques do not apply to the high-dimensional setting, because the results usually require a sample size at least $p \log p$.

Our proof techniques use the so-called local law of random matrices (Erdos and Yau, 2017), which is a recent development in the random matrix theory literature. 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 setting. These techniques provide almost sharp convergence rates to the asymptotic limit compared to other techniques such as free probability (Nica and Speicher, 2006). To the best of our knowledge, we are not aware of any previous results about the inverse of the sum of two sample covariance matrices 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) provides minimax optimal rates for predicting 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. Wu et al. (2020) (and an earlier work by Kumar and Daumé III (2012)) observed that the shared layer size r in hard parameter sharing plays a critical role of regularization.

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 present our technical results that describe how varying sample sizes and covariate shifts impact hard parameter sharing using random matrix theory. In Sections 4 and 5, we validate our theory in both simulations and a real world classification task. In Section 6, we conclude the paper and describe several open questions.

Notations. For an $n \times p$ matrix X, let $\lambda_{\min}(X)$ denote its smallest singular value and $\|X\|$ denote its spectral norm. Let $\lambda_1(X) \geqslant \lambda_2(X) \geqslant \cdots \geqslant \lambda_{p\wedge n}(X)$ denote the singular values of X. Let X^+ denote the Moore-Penrose psuedoinverse. 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 sample size and features.

Setting. Suppose we have t datasets whose samples size are all equal to n and whose features are all denoted by $X \in \mathbb{R}^{n\times p}$, and the label (vector) of the i-th task follows a linear model $Y^{(i)} = X\beta^{(i)} + \varepsilon^{(i)}$. Let $\Sigma \in \mathbb{R}^{p\times p}$ denote the population covariance matrix of all tasks. We assume that:

- $X = Z\Sigma^{1/2}$ and every entry of $Z \in \mathbb{R}^{n\times p}$ is drawn independently from a one dimensional distribution with mean zero, unit variance, and constant φ-th moment for a fixed $\varphi > 4$;
- every entry of $\varepsilon^{(i)} \in \mathbb{R}^{n\times t}$ is drawn indepdently from a one dimensional ditribution with mean zero, 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

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

Our main result shows that hard parameter sharing essentially approximates all tasks through a rank-r subspace. To formalize this intuition, we introduce several notations. Let $B^\star := [\beta_1, \beta_2, \ldots, \beta_t] \in \mathbb{R}^{p\times t}$. Let $A^\star A^{\star\top}$ denote the best rank-r subspace approximation of ${B^\star}^\top \Sigma B^\star$:[2]

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

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 hard parameter sharing decomposes to a bias term $L(B^\star a_i^\star)$ measures the error of $B^\star a_i^\star$, and a variance term that scales with $\|a_i^\star\|^2$.

Theorem 2.1. *Assume that $n > \rho p$ for a fixed constant $\rho > 1$. Let c_φ be any fixed value within $(0, \frac{\varphi-4}{2\varphi})$. Let $L(B^\star a_i^\star) := \|\Sigma^{1/2}(B^\star a_i^\star - \beta^{(i)})\|^2$. Let $\hat{A}, \hat{B}$ be a global minimum of $f(A, B)$. For any task $i = 1, 2, \ldots, t$, the prediction loss of $\hat{\beta}_i^{\mathrm{HPS}} = \hat{B}\hat{A}_i$ satisfies that with high probability,*

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

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 single-task learning, 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.

[1] More precisely, 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 $\lambda_r({B^\star}^\top\Sigma B^\star)$.

How does hard parameter sharing scale with sample size n? Obviously, the concentration error decreases with n. How about the variance $\mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right]$? 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 dimension p, i.e. $n/p \to \rho \in (\tau, 1/\tau)$ as p goes to infinity.

Under the above assumption, the following result is well-known.

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

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

Applying 2.3 to Theorem 2.1, we obtain that hard parameter sharing's variance is

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

As a remark, Fact 2.3 has a rich 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 a non-Gaussian random matrix, this result can be obtained using the well-known Stieltjes transform method (cf. Lemma 3.11 of Bai and Silverstein (2010)).

Next, we consider the bias $L(B^\star a_i^\star)$. We illustrate through a random-effects model, which has been studied for a single task case (Dobriban and Sheng, 2020). Suppose every β_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 the task-specific component be a random Gaussian vector with i.i.d. entries of mean zero and variance $p^{-1}d^2$. Thus, for any two different β_i and β_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-effects model described above, we further assume that Σ is isotropic for illustration. We show that when the rank of hard parameter sharing is one, the bias $L(B^\star a_i^\star)$ satisfies that with high probability,*

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

Since $A^\star A^{\star\top}$ is the best rank-1 approximation of $B^{\star\top}\Sigma B^\star = B^{\star\top} B^\star$, and the (i,j)-th entry of this matrix is roughly equal to

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

which follows from the definition of the random-effects model. Note that $\|\beta_0\|^2$ is roughly κ^2. One can verify that the top eigenvalue of $B^{\star\top} B^\star$ is approximately $t\kappa^2 + d^2$ and the rest of its eigenvalues are all equal to d^2. Therefore, by taking a rank-1 approximation of $B^{\star\top} B^\star$, we get the average prediction loss as above.

For the variance term, since $A^\star$ has rank-1, we have that $\sum_{i=1}^t \|a_i^\star\|^2 = 1$. Moreover, we can use Fact 2.3 to calculate $\mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right]$. Combined together, the average prediction loss of hard parameter sharing is as follows

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

We compare hard parameter sharing to single-task learning. Using Fact 2.3 above, the average prediction loss of single-task learning is $\sigma^2 \cdot \mathrm{Tr}\left[\Sigma(X^\top X)^{-1}\right] = \frac{\sigma^2 p}{n-p} \pm \mathrm{O}(n^{-c_\varphi}\sigma^2)$ with high probability. Suppose n is sufficiently large so that the error is negligible.

(i) The prediction loss of hard parameter sharing is less than single-task learning if and only if $d^2 < \frac{\sigma^2 p}{n-p}$, that is, the "task-specific variance" of β_i is smaller than the "noise variance".

(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 d^2, but variance increases by $\frac{\sigma^2 p}{n-1}$ (details omitted).

(iii) Hard parameter sharing requires at most $p + \frac{n-p}{t-(t-1)\frac{d^2(n-p)}{\sigma^2 p}} < n$ samples to get comparable loss to single-task learning. This follows by using this sample size in the average prediction loss equation above.

Proof overview. The key idea for proving Theorem 2.1 is a characterization of $f(A, B)$'s global minimizer. Since all tasks have the same covariates, the optimization objective (1.1) becomes

$$f(A, B) = \sum_{j=1}^{t} \left\| XBA_j - Y^{(j)} \right\|^2, \tag{2.1}$$

where we recall that $B \in \mathbb{R}^{p\times r}$ and $A_1, A_2, \ldots, A_t \in \mathbb{R}^r$. Let $Y = [Y^{(1)}, Y^{(2)}, \ldots, Y^{(t)}]$. Using the local optimality condition over B, that is, $\frac{\partial f}{\partial B} = 0$, we obtain $\hat{B}$ as a function of the output layers as follows

$$\hat{B}(A) := (X^\top X)^{-1} X^\top \left(\sum_{j=1}^{t} Y^{(j)} A_j^\top \right) (AA^\top)^+ = (X^\top X)^{-1} X^\top Y A^\top (AA^\top)^+. \tag{2.2}$$

Here we have used that $X^\top X$ is invertible since $n > \rho p$ (cf. Fact D.3). Plugging $\hat{B}(A)$ into equation (2.1), we obtain the following objective that only depends on the output layer:

$$g(A) := \sum_{j=1}^{t} \left\| X(X^\top X)^{-1} X^\top Y A^\top (AA^\top)^+ A_j - Y^{(j)} \right\|^2. \tag{2.3}$$

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 next claim shows that the subspace spanned by the rows of $\hat{A}$ is close to that of $A^\star$.

Claim 2.5. *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*

$$\left\| U_{\hat{A}} U_{\hat{A}}^\top - A^\star A^{\star\top} \right\|_F^2 \leqslant p^{-c_\varphi} \cdot C_1.$$

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

Claim 2.6. *In the setting of Theorem 2.1, we have that*

$$\mathop{\mathbb{E}}_{\{\varepsilon^{(j)}\}_{j=1}^t, X} [g(A)] = n \left\| \Sigma^{1/2} B^\star \left(A^\top (AA^\top)^\dagger A - \mathrm{Id}_{t\times t} \right) \right\|_F^2 + \sigma^2 (n \cdot t - p \cdot r). \tag{2.4}$$

As a result, the minimum of $\mathbb{E}[g(A)]$, 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 minimizer 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 2.7. *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*

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

The proof of Claim 2.5, Claim 2.6, and Claim 2.7 can be found in Appendix A. Provided with these results, we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. Using Claim 2.7, we get that the prediction loss of $\hat{\beta}_i^{\mathrm{HPS}}$ is within an $\mathrm{O}(n^{-1/4})$ fraction of $L(B^\star \hat{a}_i) + \sigma^2 \|\hat{a}_i\|^2 \cdot \mathrm{Tr}\left[\Sigma (X^\top X)^{-1}\right]$. Moreover, Claim 2.5 gives directly an upper bound on $\|\hat{a}_i - a_i^\star\|^2$. With this estimate, we can bound the difference

$$L(B^\star \hat{a}_i) + \sigma^2 \|\hat{a}_i\|^2 \cdot \mathrm{Tr}\left[\Sigma (X^\top X)^{-1}\right] - L(B^\star a_i^\star) - \sigma^2 \|a_i^\star\|^2 \cdot \mathrm{Tr}\left[\Sigma (X^\top X)^{-1}\right].$$

Combined together, our proof is complete. We omit the details. □

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 studies how different sample sizes and different features impact hard parameter sharing. The setting where features differ across tasks is often also called "covariate shift".

Unlike the previous section, we no longer have an optimal solution for $f(A, B)$. This is because $f(A, B)$ is in general non-convex. Instead, our result implies tight generalization bounds for any *local minimum* of $f(A, B)$. We focus on the two-task case to better understand the impact of having different sample sizes and different covariates. Suppose that $X^{(1)} = Z^{(1)}(\Sigma^{(1)})^{1/2} \in \mathbb{R}^{n_1 \times p}$ and $X^{(2)} = Z^{(2)}(\Sigma^{(2)})^{1/2} \in \mathbb{R}^{n_2 \times p}$, 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 $\varphi > 4$. The matrices $\Sigma^{(1)} \in \mathbb{R}^{p \times p}$ 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 different sample covariance matrices. To motivate our study, consider a special case where $A_1 = A_2 = 1$. Without loss of generality, we consider task two's prediction loss and the same result applies to task one. By Proposition 1 of Wu et al. (2020), we consider the case of $r = 1 < t = 2$, since when $r > 1$, the optimal solution of $f(A, B)$ is equivalent to single-task learning. When $r = 1$, B is a vector and A_1, A_2 are both scalars. Hence we can write down a closed form equation for any local minimum of $f(A, B)$. By solving B in equation (1.1), we obtain the hard parameter sharing estimator $\hat{\beta}_2^{\mathrm{HPS}} = BA_2 = B$ as follows:

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

The matrix $\hat{\Sigma}$ adds up both tasks' sample covariances, 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 its expectation similar to Claim 2.7):

$$\mathop{\mathbb{E}}_{\mathcal{E}}\left[L(\hat{\beta}_2^{\mathrm{HPS}}) \mid X^{(1)}, X^{(2)}\right] = \left\|{\Sigma^{(2)}}^{1/2}\hat{\Sigma}^{-1}{X^{(1)}}^\top X^{(1)}(\beta^{(1)} - \beta^{(2)})\right\|^2 \tag{3.2}$$

$$+ \sigma^2 \operatorname{Tr}\left[\Sigma^{(2)}\hat{\Sigma}^{-1}\right]. \tag{3.3}$$

Equation (3.2) and (3.3) correspond to the bias and variance of hard parameter sharing for two tasks, respectively. Our main result in this section characterizes the asymptotic bias-variance limits under the high-dimensional assumption. 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 $U\Lambda V^\top$ denote the SVD of M and let $\lambda_1, \lambda_2, \ldots, \lambda_p$ denote M's singular values in descending order. Our main result is stated as follows.

Theorem 3.1. *Let c_φ be a fixed value in $(0, \frac{\varphi-4}{2\varphi})$. 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 the following limits.*
(i) The variance equation (3.3) *satisfies the following estimate with high probability:*

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

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

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

(ii) The bias equation (3.2) *satisfies the following limit with high probability for any fixed vector* $w \in \mathbb{R}^p$ *with unit norm,*

$$\left| w^\top \Sigma^{(1)} \left((n_1+n_2)^2 \hat{\Sigma}^{-1} \Sigma^{(2)} \hat{\Sigma}^{-1} - {\Sigma^{(2)}}^{-1/2} V \frac{a_3 \Lambda^2 + (a_4+1)\,\mathrm{Id}}{(a_1\Lambda^2 + a_2\,\mathrm{Id})^2} V^\top {\Sigma^{(2)}}^{-1/2} \right) \Sigma^{(1)} w \right| \leqslant p^{-c_\varphi}, \tag{3.6}$$

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

$$a_3 + a_4 = \frac{1}{n_1+n_2} \sum_{i=1}^p \frac{1}{\lambda_i^2 a_1 + a_2}, \quad a_3 + \frac{1}{n_1+n_2} \sum_{i=1}^p \frac{\lambda_i^2 (a_2 a_3 - a_1 a_4)}{(\lambda_i^2 a_1 + a_2)^2} = \frac{1}{n_1+n_2} \sum_{i=1}^p \frac{\lambda_i^2 a_1}{(\lambda_i^2 a_1 + a_2)^2}. \tag{3.7}$$

Part (i) of 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 equation (3.5) and obtain $a_1 = 0$ and $a_2 = (n_2 - p)/n_2$, and apply these solutions to equation (3.4).

How does hard parameter sharing scale with sample size and covariate shift? One can see that the bias-variance limits depend intricately on both tasks' samples sizes and covariate shift. Next, we illustrate how varying them impact prediction loss, respectively.

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

Applying Theorem 3.1 to the above setting, we first solve the self-consistent equations (3.5) *by using* $\lambda_i = 1$ *for all* $1 \leqslant i \leqslant p$*. This gives us the variance limit*

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

since $a_1 + a_2 = 1 - \frac{p}{n_1+n_2}$*. Similarly, for the bias limit, we solve the self-consistent equations* (3.7) *to get* a_3 *and* a_4 *after we have obtained* a_1, a_2*. 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:* $d^2 \leqslant p^{-c}\kappa^2$*. Let* $\varepsilon = (1 + \sqrt{p/n_1})^4 - 1$*, which decreases as* n_1 *increases. The prediction loss of hard parameter sharing for task two satisfies that*

$$\left| L(\hat{\beta}_2^{\mathrm{HPS}}) - \frac{d^2 n_1^2 (n_1+n_2)}{(n_1+n_2-p)^3} - \frac{\sigma^2 p}{n_1+n_2-p} \right| \leqslant \varepsilon \cdot \frac{d^2 n_1^2 (n_1+n_2)}{(n_1+n_2-p)^3} + \mathrm{O}(p^{-c/2}). \tag{3.8}$$

The above result provides a more concrete interpretation of the bias-variance decomposition, since it depends explicitly on datasets' properties of interest. The proof of Corollary 3.3 can be found in Appendix C. As a remark, in equation (3.8), the predication loss $L(\hat{\beta}_2^{\mathrm{HPS}})$ was obtained using the global minimizer $\hat{A}$ and $\hat{B}$. By combining the bias and variance limits, we can also obtain a generalization bound for any local minimizer of $f(A, B)$. The proof is similar to Corollary 3.3, so we omit the details.

Next, we illustrate an interesting phenomenon where adding task one's samples helps task two initially, but hurts eventually. Consider the limiting estimate on the left hand side of equation (3.8). We vary sample ratio n_1/n_2 by fixing n_2 and increasing n_1. The variance term always reduces as n_1 increases. The bias term always increases as n_1 increases, which can be verified by calculating the derivative of the bias term. Furthermore, by calculating the derivative of the bias-variance limits with respect to n_1, we obtain the following dichotomy (details omitted).

(i) When $\frac{d^2}{\sigma^2} < \frac{p}{2n_2 - 3p}$, the prediction loss decreases monotonically as n_1 increases. Intuitively, this regime of d^2 always helps task two.

(ii) When $\frac{d^2}{\sigma^2} > \frac{p}{2n_2 - 3p}$, the limiting estimate 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 that bias increases less than the variance decreases, and there is only one critical point for the derivative being 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)*, which shows that*

$$\operatorname{Tr}\left[\Sigma^{(2)}\hat{\Sigma}^{-1}\right] \to \frac{1}{n_1+n_2}\operatorname{Tr}\left[\Sigma^{(2)}(a_1\Sigma^{(1)}+a_2\Sigma^{(2)})^{-1}\right] = \frac{1}{n_1+n_2}\operatorname{Tr}\left[(a_1M^\top M + a_2\,\mathrm{Id})^{-1}\right].$$

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.*

- *If* $n_1 \geqslant n_2$*, then the variance limit is smallest when there is no covariate shift.*
- *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 estimate for this example is equal to

$$\frac{p}{2(n_1+n_2)}f(\lambda), \quad \text{where} \quad f(\lambda) = (\lambda^{-2}a_1+a_2)^{-1} + (\lambda^2 a_1+a_2)^{-1}.$$

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

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

where $g(\lambda, a_1) > 0$ is a fixed function and can be derived from algebraic calculations (details omitted). We now show that $a_1 \geqslant \frac{n_1+n_2-p}{2(n_1+n_2)}$ if and only if $n_1 \geqslant n_2$, and hence explain the dichonomy. In fact, if $a_1 > a_2$, then the equations in (3.5) give that $a_1 > \frac{n_1+n_2-p}{2(n_1+n_2)}$, and one can verify that

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

This implies $n_1 > n_2$. Similarly, if $a_1 < a_2$, equations in (3.5) give that $a_1 < \frac{n_1+n_2-p}{2(n_1+n_2)}$ and $n_1 < n_2$. Thus, we conclude that $f(\lambda) \geqslant f(1)$ if and only if $n_1 \geqslant n_2$.

Overview of Stieltjes transform. 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 $\hat{\Sigma}^{-1}$. 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

$$W := \frac{1}{n_1+n_2}(\Lambda U^\top (Z^{(1)})^\top Z^{(1)} U\Lambda + V^\top (Z^{(2)})^\top Z^{(2)} V). \tag{3.9}$$

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.

Our proof uses the Stieltjes transform or the resolvent method in random matrix theory. We briefly describe the key ideas and refer 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 given by

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

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 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

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

Above, the matrix $(W - z\,\mathrm{Id})^{-1}$ is known as W's resolvent or Green's function. We will prove the convergence of W's resolvent using the so-called "local laws" with a sharp convergence rate (Bloemendal et al., 2014; Erdos and Yau, 2017; Knowles and Yin, 2016). We say that $(W - z\,\mathrm{Id})^{-1}$ converges to a deterministic $p \times p$ matrix limit $R(z)$ if for any sequence of deterministic vectors $v \in \mathbb{R}^p$ with unit norm,

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

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

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

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

$$H := \begin{pmatrix} 0 & F \\ F^\top & 0 \end{pmatrix}. \tag{3.10}$$

For this block matrix, we define its resolvent as

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

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

$$G(z) = \begin{pmatrix} (W - z\,\mathrm{Id})^{-1} & (W - z\,\mathrm{Id})^{-1}F \\ F^\top(W - z\,\mathrm{Id})^{-1} & z(F^\top F - z\,\mathrm{Id})^{-1} \end{pmatrix}.$$

Variance asymptotic limit. In Theorem B.4, we will show that for z in a small neighborhood around 0, when p goes to infinity, $G(z)$ converges to the following limit

$$\mathfrak{G}(z) := \begin{pmatrix} (a_1(z)\Lambda^2 + (a_2(z) - z)\,\mathrm{Id}_{p\times p})^{-1} & 0 & 0 \\ 0 & -\frac{n_1+n_2}{n_1}a_1(z)\,\mathrm{Id}_{n_1\times n_1} & 0 \\ 0 & 0 & -\frac{n_1+n_2}{n_2}a_2(z)\,\mathrm{Id}_{n_2\times n_2}\end{pmatrix}, \tag{3.11}$$

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

$$\begin{aligned} a_1(z) + a_2(z) &= 1 - \frac{1}{n_1+n_2}\left(\sum_{i=1}^{p}\frac{\lambda_i^2 a_1(z) + a_2(z)}{\lambda_i^2 a_1(z) + a_2(z) - z}\right), \\ a_1(z) + \frac{1}{n_1+n_2}&\left(\sum_{i=1}^{p}\frac{\lambda_i^2 a_1(z)}{\lambda_i^2 a_1(z) + a_2(z) - z}\right) = \frac{n_1}{n_1+n_2}. \end{aligned} \tag{3.12}$$

The existence and uniqueness of solutions to the above system are shown in Lemma B.7. Given this result, we 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 equation (3.12) reduces to equation (3.5). Second, since W^{-1} is the upper-left block matrix of $G(0)$, we have that W^{-1} converges to $(a_1\Lambda^2 + a_2\,\mathrm{Id})^{-1}$. Using the fact that $\operatorname{Tr}[\Sigma^{(2)}\hat{\Sigma}^{-1}] = \operatorname{Tr}\left[W^{-1}\right]/(n_1+n_2)$, we get that when p goes to infinity,

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

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

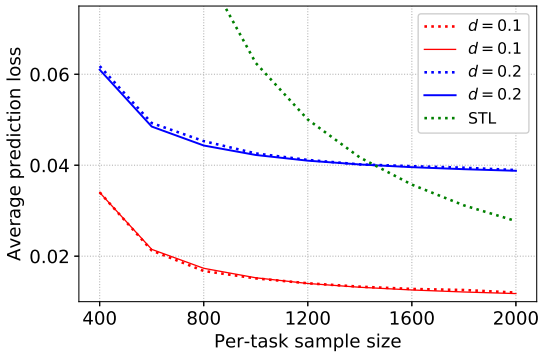


Figure 2: Validating Example 2.4 in Section 2 for 10 tasks: our estimated loss (solid line) matches the empirical loss (dotted line) accurately for various task-specific variance d^2 and sample size n settings. The feature dimension p is 200, and noise variance σ^2 is 1/4.

Bias asymptotic limit. For the bias limit in equation (3.6), we show that it is governed by the derivative of $(W - z\,\mathrm{Id})^2$ with respect to z at $z = 0$. First, we reduce the empirical bias term in equation (3.6) to W

$$(n_1 + n_2)^2 \hat{\Sigma}^{-1} \Sigma^{(2)} \hat{\Sigma}^{-1} = {\Sigma^{(2)}}^{-1/2} V W^{-2} V^\top {\Sigma^{(2)}}^{-1/2}. \tag{3.13}$$

Let $\mathcal{G}(z) := (W - z\,\mathrm{Id})^{-1}$ 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 limit of $(W - z\,\mathrm{Id})^{-1}$ is $(a_1(z)\Lambda^2 + (a_2(z) - z)\,\mathrm{Id})^{-1}$ near $z = 0$, the limit of $\frac{\mathrm{d}\mathcal{G}(0)}{\mathrm{d}z}$ satisfies that

$$\frac{\mathrm{d}\mathcal{G}(0)}{\mathrm{d}z} \to \frac{-\frac{\mathrm{d}a_1(0)}{\mathrm{d}z}\Lambda^2 - (\frac{\mathrm{d}a_2(0)}{\mathrm{d}z} - 1)\,\mathrm{Id}}{(a_1(0)\Lambda^2 + a_2(0)\,\mathrm{Id}_p)^2}. \tag{3.14}$$

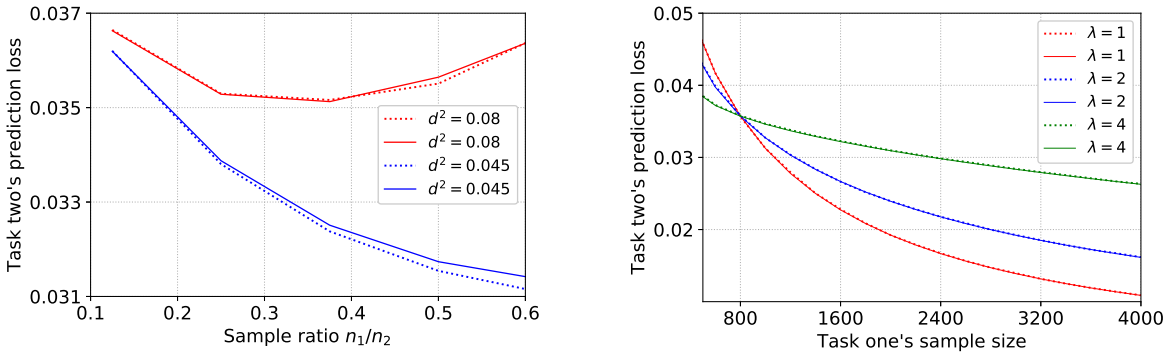
To find the derivatives of $a_1(z)$ and $a_2(z)$, we take the derivatives on both sides of equation (3.12). Let $a_3 = -\frac{\mathrm{d}a_1(0)}{\mathrm{d}z}$ and $a_4 = -\frac{\mathrm{d}a_2(z)}{\mathrm{d}z}$. One can verify that a_3 and a_4 satisfy the self-consistent equations (3.7) (details omitted). Applying equation (3.14) to equation (3.13), 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.

Schur complement and self-consistent equations. We briefly describe how to derive the matrix limit $\mathfrak{G}(z)$. First, we consider the special case where $Z^{(1)}$ and $Z^{(2)}$ are both multivariate Gaussian random matrices. By rotational invariance, we have that $Z^{(1)}U$ and $Z^{(2)}V$ are still multivariate Gaussian random matrices. Next, we use the Schur complement formula to deal with the resolvent $G(z)$. We show that $G(z)$'s diagonal entries satisfy a set of self-consistent equations in the limit, leading to equation (3.12). On the other hand, $G(z)$'s off-diagonal entries are approximately zero using standard concentration bounds. Finally, we extend our result to general random matrices under the finite φ-th moment condition. We prove an anisotropic local law using recent developments in random matrix theory (Erdos and Yau, 2017; Knowles and Yin, 2016). A complete proof of Theorem 3.1 can be found in Appendix B.

4 Simulation Studies

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



(a) Example 3.2

(b) Example 3.4

Figure 3: Validating Example 3.2 and Example 3.4 in Section 3 for two tasks: in both experiments, our estimated losses (solid line) match the empirical losses (dotted line) accurately. In Figure 3a, we discover several interesting phenomena by fixing task two's sample size and increasing task one's sample size. Depending on how large d^2 is, task two's prediction loss decreases initially before increasing again, or decreases monotonically. In Figure 3b, 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 larger than task one. Otherwise, having covariate shift (surprisingly) decreases task two's prediction loss.

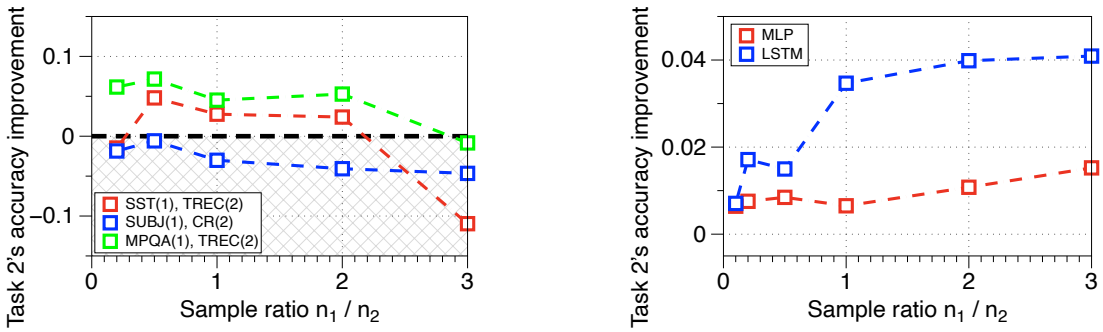
Sample efficiency. First, we study the setting of Example 2.4. Figure 2 shows the average prediction loss as we increase the number of samples per-task from 400 to 2000. In all the parameter settings, our results estimate the empirical losses accurately. We also observe a trend where the average prediction loss increases as we increase 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 at the same prediction loss level, we find that hard parameter sharing requires fewer samples per-task than STL.

Sample ratio. Second, we show how varying sample ratio impacts hard parameter sharing in the setting of Example 3.2. Figure 3a shows task two's prediction loss as we increase the sample ratio n_1/n_2 from 1/10 to 7/10. Again, our estimates are reasonably accurate compared to the empirical losses. 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 is $d^2 = 0.08$, the noise variance is $\sigma^2 = 0.3^2$, 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 increases when the sample ratio is above certain level. On the other hand, when $d^2 = 0.045$, task two's prediction loss decreases monotonically. Our work explains this trend as discussed below Corollary 3.3. ≪HZ notes: check≫

Covariate shift. Finally, we show how varying covariate shift impacts hard parameter sharing in the setting of Example 3.4. Figure 3b 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 more covariate shift. Our estimates match the empirical losses accurately, and indeed we observe the dichotomy in Example 3.4. The noise variance σ^2 is 1/4.

5 Further Studies on Text Classification Tasks

Our results and their implications 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, for predicting whether a sentence has a positive or a negative sentiment. Our datasets include a review sentiment dataset (MR) (Pang and Lee, 2005), a sentence subjectivity dataset (SUBJ) (Pang and Lee, 2004), a customer



(a) HPS vs. STL

(b) HPS vs. covariance alignment

Figure 4: Comparing hard parameter sharing (HPS) to single-task learning (STL) and a covariance alignment approach proposed by Wu et al. (2020): In Figure 4a, we observe that for multiple example task pairs, increasing task one's sample size improves task two's prediction accuracy initially, but hurts eventually – a phenomenon similar to Figure 3a. In Figure 4b, we observe that as task one's sample size increases, covariance alignment improves more over HPS.

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. In Figure 4a, 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 one's sample size at 1000 and increase task two's sample size from 100 to 3000. These examples and the one in Figure 3a suggest a natural progressive training schedule, where we add samples progressively until performance drops:

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

For example, if we apply this procedure to the settings of Figure 4a and 3a, it will terminate until reaching the optimal sample ratio. The advantage of this procedure is that it reduces the computational cost compared to standard round-robin training schedules.

We evaluate the progressive training procedure on the six text classification datasets. First, we conduct multi-task training over all 15 pairs from the six datasets. We find that adding task one's samples progressive requires only 45% of the computational cost to achieve the same test accuracy for task two than standard round-robin training schedules. Second, we conduct multi-task training on all six datasets. We find that adding samples progressively from all datasets requires less than 35% of the computational cost to achieve the same test accuracy averaged over the six datasets than standard round-robin training schedules.

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 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.

[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.

We conduct multi-task training on all 15 task pairs from the six datasets. In Figure 4b, 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 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 $3,000$.

6 Conclusions and Discussions

This work proposed a formal study of a widely used hard paramaeter sharing approach in high-dimensional linear regression, where the sample size and feature dimension increase at a fixed ratio. We provided tight generalization bounds that scale with dataset properties such as their sample sizes and covariate shifts. Based on these bounds, we analyzed the impact of varying sample sizes and covariate shifts on the prediction loss of hard parameter sharing. Our work rigorously explains several empirical phenomena such as sample efficiency and negative transfer related to these dataset properties. We validated our results and conducted further studies on a real world classification task.

We describe several open questions for future work. First, our bound in Corollary 3.3 involves an error term that scales down with n_1. Tightening this error bound requires showing the limit of $\|({Z^{(1)}}^\top Z^{(1)} + {Z^{(2)}}^\top Z^{(2)})^{-1} {Z^{(1)}}^\top Z^{(1)}\|_F^2$ for two isotropic random matrices. This requires studying the asymptotic distribution of the singular values of the non-symmetric matrix $({Z^{(1)}}^\top Z^{(1)})^{-1} {Z^{(2)}}^\top Z^{(2)} + \mathrm{Id}$, which is still an open problem in random matrix theory. The eigenvalue distribution of this matrix, which has been obtained in Zheng et al. (2017), might be helpful towards resolving this problem. Second, it would be interesting to extend our results to classification problems. Several recent work have made remarkable progress for logistic regression in the high-dimensional setting, e.g. Sur and Candès (2019). An interesting question is to study logistic regression in a multiple-sample setting.

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 Ré. 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 Missing Proof of Theorem 2.1

Proof of Claim 2.6. To facilitate the analysis, we consider the following matrix notations. Denote

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

For any $j = 1, 2, \ldots, t$, 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. We will use matrix notations in the proof since they are more compact. 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,

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

where $U_A U_A^\top \in \mathbb{R}^{t\times t}$ denotes the subspace $A^\top (AA^\top)^+ A$. Taking expectation of equation (A.1) over X, we get

$$\left\|\Sigma^{1/2}(B^\star U_A U_A^\top - B^\star)\right\|^2.$$

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

$$\sum_{j=1}^{t} \langle XH_j, E_j\rangle = \langle X(B^\star U_A U_A^\top - B^\star), \mathcal{W}A - \mathcal{E}\rangle = -\langle X(B^\star U_A U_A^\top - B^\star), \mathcal{E}\rangle, \tag{A.2}$$

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

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

$$\sum_{j=1}^{t} \|E_j\|^2 = \|\mathcal{W}A - \mathcal{E}\|_F^2 = \|\mathcal{E}\|_F^2 - \langle \mathcal{W}A, \mathcal{E}\rangle, \tag{A.3}$$

which is because $\|\mathcal{W}A\|^2 = \langle \mathcal{W}A, \mathcal{E}\rangle$ by algebraic calculation. Hence, it suffices to show that the expectation of equation (A.3) 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, notice that for $U_X U_X^\top := X(X^\top X)^{-1}X^\top$ and $1\leqslant i,j\leqslant t$,

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

which implies that $\mathbb{E}_{\mathcal{E}}\left[\mathcal{E}^\top U_X U_X^\top \mathcal{E}\right] = p\sigma^2\cdot \mathrm{Id}_{t\times t}$. In the above derivation, the second step used the fact that $\mathbb{E}(\varepsilon_k^{(i)}\varepsilon_l^{(i)}) = \sigma^2\cdot\delta_{ij}\delta_{kl}$ for $1\leqslant k,l\leqslant n$, and the third step used the fact that $\mathrm{Tr}(U_X U_X^\top) = \mathrm{Tr}(U_X^\top U_X) = p$. Therefore, we have that

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

because $U_A U_A^\top$ has rank r. Hence the proof is complete.

Proof of Claim 2.5. Corresponding to the right-hand side of (2.4), we define the function

$$h(A) := n\left\|\Sigma^{1/2}B^\star\left(A^\top(AA^\top)^\dagger A - \mathrm{Id}_{t\times t}\right)\right\|_F^2 + \sigma^2(n\cdot t - p\cdot r). \tag{A.4}$$

In order to show that the $U_{\hat{A}}U_{\hat{A}}^\top$ is close to $A^\star {A^\star}^\top$, we first show that $g(A)$ is close to $h(A)$ as follows:

$$|g(A) - h(A)| \lesssim n^{-c_\varphi}\cdot n\left\|\Sigma^{1/2}B^\star(U_A U_A^\top - \mathrm{Id}_{t\times t})\right\|^2 + n^{-c_\infty}\cdot\sigma^2\cdot n\cdot t, \tag{A.5}$$

for any fixed values c_φ within $(0, \frac{\varphi-4}{2\varphi})$ and c_∞ within $(0, 1/2)$. To show the above result, we consider the concentration error of each part of $g(A)$.

For equation (A.1), applying Corollary D.5 to $XH_j = Z\Sigma^{1/2}H_j$, we obtain that $\left\|Z\Sigma^{1/2}H_j\right\|^2 = n\|\Sigma^{1/2}H_j\|^2\cdot(1+\mathrm{O}(p^{-c_\varphi}))$ with high probability. This implies that

$$\left|\sum_{j=1}^{t}\|XH_j\|^2 - \sum_{j=1}^{t} n\|\Sigma^{1/2}H_j\|^2\right| \lesssim n^{-c_\varphi}\cdot n\left\|\Sigma^{1/2}B^\star(U_A U_A^\top - \mathrm{Id}_{t\times t})\right\|^2. \tag{A.6}$$

For equation (A.2), using Corollary D.6, we obtain that the following holds with high probability for any small constant $c > 0$:

$$\begin{aligned}
|\langle XB^\star(U_AU_A^\top - \mathrm{Id}), \mathcal{E}\rangle| &\leqslant n^c \cdot \sigma \cdot \left\|XB^\star(U_AU_A^\top - \mathrm{Id})\right\|_F \\
&\leqslant n^c \cdot \sigma \cdot \|Z\| \cdot \left\|\Sigma^{1/2}B^\star(U_AU_A^\top - \mathrm{Id})\right\|_F \\
&\lesssim n^c \cdot \sigma \cdot \sqrt{n} \cdot \sqrt{t}\left\|\Sigma^{1/2}B^\star(U_AU_A^\top - \mathrm{Id})\right\| \\
&\leqslant n^{c_\infty + c} \cdot \left\|\Sigma^{1/2}B^\star(U_AU_A^\top - \mathrm{Id}_{t\times t})\right\|^2 + n^{-c_\infty} \cdot \sigma^2 \cdot n \cdot t
\end{aligned} \tag{A.7}$$

which is bounded by the right hand side of equation (A.5) as long as c is taken sufficiently small. In the second step, we used the fact that $X = Z\Sigma^{1/2}$. In the third step, we used Fact D.3 (ii) to bound the operator norm $\|Z\|$ by $\mathrm{O}(\sqrt{n})$, and used $\left\|\Sigma^{1/2}B^\star(U_AU_A^\top - \mathrm{Id})\right\|_F \leqslant \sqrt{t}\left\|\Sigma^{1/2}B^\star(U_AU_A^\top - \mathrm{Id})\right\|$ since the matrix $\Sigma^{1/2}B^\star(U_AU_A^\top - \mathrm{Id})$ has rank at most t. In the fourth step, we used AM-GM inequality.

For equation (A.3), using Corollary D.6, we obtain that with high probability,

$$\begin{aligned}
\left|\|\mathcal{E}\|_F^2 - \sigma^2 \cdot n \cdot t\right| = \left|\mathrm{Tr}\left(\mathcal{E}^\top \mathrm{Id}_{n\times n}\,\mathcal{E}\right) - \sigma^2 \cdot n \cdot t\right| &\leqslant n^c \cdot \sigma^2 \|\,\mathrm{Id}_{n\times n}\,\|_F \\
&= n^{1/2+c} \cdot \sigma^2 \leqslant n^{-c_\infty} \cdot \sigma^2 \cdot n \cdot t,
\end{aligned} \tag{A.8}$$

as long as c is taken sufficiently small. For the inner product between $\mathcal{W}A$ and $\mathcal{E}$, we have that with high probability,

$$\begin{aligned}
\left|\langle \mathcal{W}A, \mathcal{E}\rangle - \sigma^2 \cdot p \cdot r\right| &= \left|\mathrm{Tr}\left[\left(\mathcal{E}^\top U_X U_X^\top \mathcal{E} - p\sigma^2 \cdot \mathrm{Id}_{t\times t}\right) U_AU_A^\top\right]\right| \\
&\leqslant \left\|U_AU_A^\top\right\|_F \cdot \left\|\mathcal{E}^\top U_X U_X^\top \mathcal{E} - p\sigma^2 \cdot \mathrm{Id}_{t\times t}\right\| \\
&\leqslant \sqrt{r} \cdot (n^c \cdot \sigma^2 \cdot \|U_X U_X^\top\|_F) \\
&\leqslant n^{1/2+c} \cdot \sigma^2 \cdot \sqrt{r} \leqslant n^{-c_\infty} \cdot \sigma^2 \cdot n \cdot t,
\end{aligned} \tag{A.9}$$

as long as c is taken sufficiently small. Here in the third step, we applied Corollary D.6 to $\left\|\mathcal{E}^\top U_X U_X^\top \mathcal{E} - p\sigma^2 \cdot \mathrm{Id}_{t\times t}\right\|$ and used that $\left\|U_AU_A^\top\right\|_F = \sqrt{r}$ because U_A is of rank r. In the fourth step, we used that $\|U_XU_X^\top\|_F = \sqrt{p} \leqslant \sqrt{n}$ because U_X is of rank p.

Combining equations (A.6), (A.7), (A.8), and (A.9), we obtain equation (A.5).

Next, we use equation (A.5) to prove the claim. Using triangle inequality, we upper bound the gap between $g(A^\star)$ and $g(\hat{A})$:

$$\begin{aligned}
h(\hat{A}) - h(A^\star) &\leqslant |g(A^\star) - h(A^\star)| + (g(\hat{A}) - g(A^\star)) + \left|g(\hat{A}) - h(\hat{A})\right| \\
&\leqslant |g(A^\star) - h(A^\star)| + \left|g(\hat{A}) - h(\hat{A})\right| \\
&\lesssim n^{-c_\varphi} \cdot n\left\|\Sigma^{1/2}B^\star\right\|_F^2 + n^{-c_\infty} \cdot \sigma^2 \cdot n \cdot t.
\end{aligned} \tag{A.10}$$

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 (A.5) and the fact that the spectral norm of $U_AU_A^\top - \mathrm{Id}$ is at most one. Using equation (A.4), we can obtain 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 $B^{\star\top}\Sigma B^\star$, and v_i, $1 \leqslant i \leqslant t$, be the corresponding eigenvectors. Then we have $A^\star A^{\star\top} = \sum_{i=1}^r v_i v_i^\top$, and

$$\begin{aligned}
h(\hat{A}) - h(A^\star) = n\sum_{i=1}^r \lambda_i - n\sum_{i=1}^t \lambda_i \|U_{\hat{A}}^\top v_i\|^2 = n\sum_{i=1}^r \lambda_i\left(1 - \|U_{\hat{A}}^\top v_i\|^2\right) - n\sum_{i=r+1}^t \lambda_i \|U_{\hat{A}}^\top v_i\|^2 \\
\geqslant n(\lambda_r - \lambda_{r+1})\sum_{i=r+1}^t \|U_{\hat{A}}^\top v_i\|^2,
\end{aligned} \tag{A.11}$$

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

$$\begin{aligned}\|A^\star A^{\star\top}-U_{\hat{A}}U_{\hat{A}}^{\top}\|_F^2&=\sum_{i=1}^{t}v_i^{\top}(A^\star A^{\star\top}-U_{\hat{A}}U_{\hat{A}}^{\top})^2v_i\\&=\sum_{i=1}^{r}\left(1-\|U_{\hat{A}}^{\top}v_i\|^2\right)+\sum_{i=r+1}^{t}\|U_{\hat{A}}^{\top}v_i\|^2=2\sum_{i=r+1}^{t}\|U_{\hat{A}}^{\top}v_i\|^2.\end{aligned}$$

Thus from (A.11), we obtain that

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

This completes the proof of Claim 2.5.

Proof of Claim 2.7. The proof is similar to the one for equation (A.5). The prediction loss of hard parameter sharing for task i is equal to

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

where we denote $E_i=(X^{\top}X)^{-1}X^{\top}\mathcal{E}\hat{a}_i$. We divide the prediction loss to 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. Using Corollary D.6, the concentration error can be bounded as

$$\begin{aligned}\left|\langle\Sigma^{1/2}(B^\star\hat{a}_i-\beta_i),\Sigma^{1/2}E_i\rangle\right|&=\left|\langle\hat{a}_i^{\top}X(X^{\top}X)^{-1}\Sigma(B^\star\hat{a}_i-\beta_i),\mathcal{E}\rangle\right|\\&\leqslant n^c\sigma\cdot\left\|\hat{a}_i^{\top}X(X^{\top}X)^{-1}\Sigma(B^\star\hat{a}_i-\beta_i)\right\|_F\\&\leqslant n^c\sigma\cdot\|\hat{a}_i\|\cdot\|\Sigma^{1/2}(B^\star\hat{a}_i-\beta_i)\|\cdot\|(Z^{\top}Z)^{-1}\|^{1/2}\\&\lesssim n^{-1/2+c}\sigma\cdot\|\hat{a}_i\|\cdot\left\|\Sigma^{1/2}(B^\star\hat{a}_i-\beta_i)\right\|\\&\leqslant n^{-1/2+c}\sigma^2\cdot\|\hat{a}_i\|^2+n^{-1/2+c}L(B^\star\hat{a}_i),\end{aligned}$$

with high probability for any small constant $c>0$. Here in the third step we used

$$\begin{aligned}\left\|\hat{a}_i^{\top}X(X^{\top}X)^{-1}\Sigma(B^\star\hat{a}_i-\beta_i)\right\|_F&=\|\hat{a}_i\|\cdot\left[(B^\star\hat{a}_i-\beta_i)^{\top}\Sigma(X^{\top}X)^{-1}X^{\top}X(X^{\top}X)^{-1}\Sigma(B^\star\hat{a}_i-\beta_i)\right]^{1/2}\\&\leqslant\|\hat{a}_i\|\cdot\|\Sigma^{1/2}(B^\star\hat{a}_i-\beta_i)\|\cdot\left\|\Sigma^{1/2}(X^{\top}X)^{-1}\Sigma^{1/2}\right\|^{1/2}\\&=\|\hat{a}_i\|\cdot\|\Sigma^{1/2}(B^\star\hat{a}_i-\beta_i)\|\cdot\left\|(Z^{\top}Z)^{-1}\right\|^{1/2}\end{aligned}$$

because $X=Z\Sigma^{1/2}$. In the fourth step, we used Fact D.3 (ii) to bound the operator norm $\left\|(Z^{\top}Z)^{-1}\right\|$ by $\mathrm{O}(n^{-1})$. In the last step we used AM-GM inequality.

Part 3: The final part is $\|E_i\|^2$. We rewrite it as

$$\begin{aligned}\|\Sigma^{1/2}E_i\|^2 &= \left\|\sum_{j=1}^{t}\hat{a}_j\Sigma^{1/2}(X^\top X)^{-1}X^\top\varepsilon^{(j)}\right\|^2 \\ &= \sum_{1\leqslant j,k\leqslant t}\hat{a}_i(j)\hat{a}_i(k){\varepsilon^{(j)}}^\top X(X^\top X)^{-1}\Sigma(X^\top X)^{-1}X^\top\varepsilon^{(k)}. \end{aligned} \tag{A.12}$$

If $j\neq k$, using Corollary D.6 we obtain that

$$\begin{aligned}\left|{\varepsilon^{(j)}}^\top X(X^\top X)^{-1}\Sigma(X^\top X)^{-1}X^\top\varepsilon^{(k)}\right| &\leqslant \sigma^2\cdot\left\|X(X^\top X)^{-1}\Sigma(X^\top X)^{-1}X^\top\right\|_F \\ &= \sigma^2\cdot\left\|\Sigma^{1/2}(X^\top X)^{-1}\Sigma^{1/2}\right\|_F \\ &\leqslant \sigma^2\cdot p^{1/2}\cdot\left\|(Z^\top Z)^{-1}\right\| \lesssim \sigma^2\cdot n^{-1/2+c}\end{aligned} \tag{A.13}$$

with high probability for any small constant $c>0$. On the other hand, if $j=k$, we notice that

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

Then using Corollary D.6, we obtain that

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

Plugging (A.13) and (A.14) into (A.12), we obtain that

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

Finally, combining the three parts together, we conclude the proof of Claim 2.7.

B Proof of Theorem 3.1

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 $Z^{(2)}$ also follow an isotropic Gaussian distribution. Hence it suffices to consider the following resolvent

$$G(z) = \begin{pmatrix} -z\operatorname{Id}_p & n^{-1/2}\Lambda(Z^{(1)})^\top & n^{-1/2}(Z^{(2)})^\top \\ n^{-1/2}Z^{(1)}\Lambda & -\operatorname{Id}_{n_1} & 0 \\ n^{-1/2}Z^{(2)} & 0 & -\operatorname{Id}_{n_2}\end{pmatrix}^{-1}. \tag{B.1}$$

We show how to derive the matrix limit $\mathfrak{G}(z)$ and the self-consistent equation system (3.12). We first introduce several notations. Define the 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 $G(z)$:

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

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

Definition B.1 (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 $X^{(i)} := [X_{a_1,a_2} : a_1, a_2 \in \mathcal{I} \setminus \{i\}]$ as a square matrix with dimension $p+n_1+n_2-1$. For the indices of $X^{(i)}$, we use $X^{(i)}_{a_1,a_2}$ to denote X_{a_1,a_2} when a_1 and a_2 both not equal to i, and $X^{(i)}_{a_1,a_2}$ to be zero 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

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

As a remark, we denote the partial traces $m^{(i)}(z)$, $m_0^{(i)}(z)$, $m_1^{(i)}(z)$, and $m_2^{(i)}(z)$ by replacing $G(z)$ with $G(z)^{(i)}$ in equation (B.2).

Self-consistent equations. We briefly describe the ideas for deriving the equation set (3.12). A complete full can be found in Lemma B.14. We show that with high probability, the following equations hold approximately:

$$\begin{aligned} m_1^{-1}(z) &= -1 + \frac{1}{n}\sum_{i=1}^{p} \frac{\lambda_i^2}{z + \lambda_i^2 \frac{n_1}{n_1+n_2} m_1(z) + \frac{n_2}{n_1+n_2} m_2(z) + \mathrm{o}(1)} + \mathrm{o}(1), \\ m_2^{-1}(z) &= -1 + \frac{1}{n}\sum_{i=1}^{p} \frac{1}{z + \lambda_i^2 \frac{n_1}{n_1+n_2} m_1(z) + \frac{n_2}{n_1+n_2} m_2(z) + \mathrm{o}(1)} + \mathrm{o}(1). \end{aligned} \tag{B.3}$$

These are the self-consistent equations that we stated in equation (3.12). 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

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

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

Next, we apply standard concentration bounds to simplify the above results. For $G^{-1}_{i,i}(z)$, 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 B.14, we have that with high probability the following holds

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

by our definition of the partial traces $m_1^{(i)}(z)$ and $m_2^{(i)}(z)$ with respect to the resolvent minor $G^{(i)}(z)$. Since we 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

$$G_{i,i}(z) = -\left(z + \frac{\lambda_i^2 \cdot n_1}{n_1+n_2} m_1(z) + \frac{n_2}{n_1+n_2} m_2(z) + \mathrm{o}(1)\right)^{-1}. \tag{B.4}$$

For the other two blocks $\mathcal{I}_1$ and $\mathcal{I}_2$, using similar ideas we have that the following holds with high probability

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

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

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

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

$$m_1^{-1}(z) = -1 - \frac{1}{n_1+n_2}\sum_{i=1}^{p} \lambda_i^2 G_{i,i}(z) + \mathrm{o}(1)$$
$$m_2^{-1}(z) = -1 - \frac{1}{n_1+n_2}\sum_{i=1}^{p} G_{i,i}(z) + \mathrm{o}(1).$$

where we used the definition of $m(z)$ and $m_0(z)$. By applying equation (B.4) for $G_{i,i}(z)$, we obtain the self-consistent equations B.3. In Lemma B.8, 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 $\mathfrak{G}(z)$. Applying our solution for $m_1(z)$ and $m_2(z)$ into equation (B.4), we get that for i in $\mathcal{I}_0$, $G_{i,i}(z) = (-z + \lambda_i^2 a_1(z) + a_2(z) + \mathrm{o}(1))^{-1}$. Similarly, for μ in $\mathcal{I}_1$, we have that $G_{\mu,\mu}(z) = -\frac{n_1+n_2}{n_1} a_1(z) + \mathrm{o}(1)$, and for ν in $\mathcal{I}_2$, we have that $G_{\nu,\nu}(z) = -\frac{n_1+n_2}{n_2} a_2(z) + \mathrm{o}(1)$. Hence we have derived the diagonal entries of $\mathfrak{G}(z)$. In Lemma B.13, we show that the off-diagonal entries are close to zero. For example, for $i \neq j \in \mathcal{I}_1$, by Schur complement, we have that

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

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

Notations. We introduce several 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 B.2 (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 any $n > n_0(c, D)$,

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

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 greater than D/c, then we have that

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

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.

Fact B.3 (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

$$\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p} |Z_{i,j}| \prec q. \tag{B.5}$$

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

$$\begin{aligned}\mathbb{P}\left(\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p} |Z_{i,j}| \geqslant (\log n)n^{\frac{2}{\varphi}}\right) &\leqslant \sum_{i=1}^{n}\sum_{j=1}^{p} \mathbb{P}\left(|Z_{i,j}| \geqslant (\log n)n^{\frac{2}{\varphi}}\right) \\ &\leqslant \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{C}{\left[(\log n)n^{\frac{2}{\varphi}}\right]^{\varphi}} = \mathrm{O}((\log n)^{-\varphi}).\end{aligned} \tag{B.6}$$

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

B.1 Limit of the Resolvent

We now state the main random matrix result—Theorem B.5—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

$$\mathbf{D} := \left\{z = E + \mathrm{i}\eta \in \mathbb{C}_+ : |z| \leqslant (\log n)^{-1}\right\}. \tag{B.7}$$

Theorem B.4. *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*

$$\max_{z\in\mathbf{D}} \left|\mathbf{u}^\top(G(z) - \mathfrak{G}(z))\mathbf{v}\right| \prec q \tag{B.8}$$

holds on the high probability event

$$\left\{\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p} |Z^{(1)}_{i,j}| \leqslant (\log n)n^{\frac{2}{\varphi}}, \ \max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p} |Z^{(2)}_{i,j}| \leqslant (\log n)n^{\frac{2}{\varphi}}\right\}. \tag{B.9}$$

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

Lemma B.5. *In the setting of Theorem B.4, assume that $Z^{(1)}$ and $Z^{(2)}$ satisfy the bounded support condition* (B.5) *with $q = n^{-\frac{\varphi-4}{2\varphi}}$. Then we have that the anisotropic local law in equation* (B.8) *holds for any deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{p+n_1+n_2}$.*

Remark B.6. 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 (B.6) with

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

for a small constant $\delta > 0$. Hence we can replace event (B.9) with

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

which holds with higher probability. But on this event we need to take a larger support $q = n^{-\frac{\varphi-4}{2\varphi}+\delta}$, which means a worse convergence rate. In general, with Lemma B.5 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 B.4 using a simple cutoff argument.

Proof. 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 (B.6), we have

$$\mathbb{P}(\widetilde{Z}^{(1)} = Z^{(1)}, \widetilde{Z}^{(2)} = Z^{(2)}) = 1 - \mathrm{O}((\log n)^{-\varphi}). \tag{B.10}$$

By definition, we have

$$\mathbb{E}\widetilde{Z}^{(1)}_{\mu i} = -\mathbb{E}\left[\mathbf{1}\left(|Z^{(1)}_{\mu i}| > qn^{1/2}\log n\right)Z^{(1)}_{\mu i}\right], \quad \mathbb{E}|\widetilde{Z}^{(1)}_{\mu i}|^2 = 1 - \mathbb{E}\left[\mathbf{1}\left(|Z^{(1)}_{\mu i}| > qn^{1/2}\log n\right)|Z^{(1)}_{\mu i}|^2\right]. \tag{B.11}$$

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

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

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

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

Plugging the above two estimates into equation (B.11) and using $\varphi > 4$, we get that

$$|\mathbb{E}\widetilde{Z}^{(1)}_{\mu i}| = \mathrm{O}(n^{-3/2}), \quad \mathbb{E}|\widetilde{Z}^{(1)}_{\mu i}|^2 = 1 + \mathrm{O}(n^{-1}). \tag{B.12}$$

From the first estimate in equation (B.12), we can also get a bound on the operator norm:

$$\|\mathbb{E}\widetilde{Z}^{(1)}\| = \mathrm{O}(n^{-1/2}). \tag{B.13}$$

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 Theorem B.5 with bounded support q, so we get that

$$\left|\mathbf{u}^\top (G(\widehat{Z}^{(1)}, \widehat{Z}^{(2)}, z) - \mathfrak{G}(z))\mathbf{v}\right| \prec q, \tag{B.14}$$

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 (B.12) and (B.13), 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 D.3(ii) to bound the operator norm of $\widetilde{Z}^{(k)}$. Together with equation (B.39) 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 (B.14), we obtain that estimate (B.8) also holds for $G(z)$ on the event $\{\widetilde{Z}^{(1)} = Z^{(1)}, \widetilde{Z}^{(2)} = Z^{(2)}\}$, which concludes the proof by equation (B.10). □

Now we are ready to complete the proof of Theorem 3.1 using Theorem B.4.

Proof of Theorem 3.1, Part i). Following up our discussion in Section 3, by Theorem B.4, for any $1 \leqslant i \leqslant p$ we have that

$$\begin{aligned}
&\left|\left[\left((X^{(1)})^\top X^{(1)} + (X^{(2)})^\top X^{(2)}\right)^{-1}\Sigma - n^{-1}\Sigma_2^{-1/2} V \mathfrak{G}(0) V^\top \Sigma_2^{-1/2}\Sigma\right]_{ii}\right| \\
&= n^{-1}\left|\mathbf{e}_i^\top \Sigma_2^{-1/2} V \left(\mathcal{G}(0) - \mathfrak{G}(0)\right) V^\top \Sigma_2^{-1/2}\Sigma\mathbf{e}_i\right| \prec n^{-1} q \|V^\top \Sigma_2^{-1/2}\Sigma\mathbf{e}_i\| \lesssim n^{-1} q \|\Sigma\mathbf{e}_i\|,
\end{aligned} \tag{B.15}$$

on the event (B.9), where $q = n^{-\frac{\varphi-4}{2\varphi}}$ and $\mathbf{e}_i$ denotes the standard basis vector along the i-th direction. Next, recall from Section 3 that

$$n^{-1}\Sigma_2^{-1/2} V \mathfrak{G}(0) V^\top \Sigma_2^{-1/2}\Sigma = n^{-1}(a_1\Sigma_1 + a_2\Sigma_2)^{-1}\Sigma.$$

Together with equation (B.15), this identity implies that

$$\begin{aligned}
\mathrm{Tr}\left[\left((X^{(1)})^\top X^{(1)} + (X^{(2)})^\top X^{(2)}\right)^{-1}\Sigma\right] &= \sum_{i=1}^p \left[\left((X^{(1)})^\top X^{(1)} + (X^{(2)})^\top X^{(2)}\right)^{-1}\Sigma\right]_{ii} \\
&= n^{-1}\,\mathrm{Tr}\left[(a_1\Sigma_1 + a_2\Sigma_2)^{-1}\Sigma\right] + \mathrm{O}_\prec(q\|A\|)
\end{aligned}$$

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

Proof of Theorem 3.1, part ii). Recall that in the setting of Theorem 3.1, we have equation (3.13). For simplicity, we denote the vector $\mathbf{v} := V^\top \Sigma_2^{-1/2} \Sigma_1^{1/2} w$. Under the constant φ-th moment condition, using Corollary B.4 we obtain that

$$\max_{z\in\mathbb{C}:|z|=(\log n)^{-1}} |\mathbf{v}^\top (G(z) - \mathfrak{G}(z))\mathbf{v}| \prec q\|\mathbf{v}\|^2, \quad q := n^{-\frac{\varphi-4}{2\varphi}},$$

on an event Ω with $\mathbb{P}(\Omega) = 1 - \mathrm{o}(1)$. Now combining this estimate with Cauchy's integral formula, we get that on Ω,

$$\begin{aligned} \mathbf{v}^\top \mathcal{G}'(0)\,\mathbf{v} = \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{\mathbf{v}^\top \mathcal{G}(z)\,\mathbf{v}}{z^2} \mathrm{d}z &= \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{\mathbf{v}^\top \mathfrak{G}(z)\,\mathbf{v}}{z^2} \mathrm{d}z + \mathrm{O}_\prec(q\|\mathbf{v}\|^2) \\ &= \mathbf{v}^\top \mathfrak{G}'(0)\,\mathbf{v} + \mathrm{O}_\prec(q\|\mathbf{v}\|^2), \end{aligned} \tag{B.16}$$

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

$$\mathbf{v}^\top \mathfrak{G}'(0)\,\mathbf{v} = \mathbf{v}\,\frac{a_3\Lambda^2 + (1+a_4)\,\mathrm{Id}_p}{(a_1\Lambda^2 + a_2\,\mathrm{Id}_p)^2}\,\mathbf{v}, \tag{B.17}$$

where we recall (3.14) and that $a_3 = -a_1'(0)$ and $a_4 = -a_2'(0)$. It remains to derive equation (3.7) for (a_3, a_4). Taking implicit differentiation of equation (3.12), we obtain that

$$\rho_1 \frac{-a_1'(0)}{a_1^2(0)} = \frac{1}{p}\sum_{i=1}^p \frac{\lambda_i^2\left(1 - \lambda_i^2 a_1'(0) - a_2'(0)\right)}{(\lambda_i^2 a_1(0) + a_2(0))^2}, \quad \rho_2 \frac{-a_2'(0)}{a_2^2(0)} = \frac{1}{p}\sum_{i=1}^p \frac{1 - \lambda_i^2 a_1'(0) - a_2'(0)}{(\lambda_i^2 a_1(0) + a_2(0))^2}.$$

It is not hard to see that this equation is equivalent to equation (3.7). Finally, we can calculate $\mathbf{v}^\top \mathfrak{G}'(0)\,\mathbf{v}$ as in equation (3.14), which concludes equation (3.6) by equation (B.16). □

B.2 Self-Consistent Equations

The rest of this section is devoted to the proof of Theorem B.5. In this section, we show that the limiting equation (3.12) has a unique solution $(a_1(z), a_2(z))$ for any $z \in \mathbf{D}$ in equation (B.7). Otherwise, Theorem B.5 will be a vacuous statement.

First observe that when $z = 0$, equation (3.12) can be reduced to equation (3.5), from which we can derive an equation of a_1 only:

$$f(a_1) = \frac{\rho_1}{\rho_1 + \rho_2}, \quad \text{with} \quad f(a_1) := a_1 + \frac{1}{\rho_1 + \rho_2} \cdot \frac{1}{p}\sum_{i=1}^p \frac{\lambda_i^2 a_1}{\lambda_i^2 a_1 + (1 - (\rho_1+\rho_2)^{-1} - a_1)}. \tag{B.18}$$

It is not hard to see that f is strictly increasing on $[0, 1 - (\rho_1+\rho_2)^{-1}]$. Moreover, we have $f(0) = 0 < 1$, $f(1 - (\rho_1+\rho_2)^{-1}) = 1 > \rho_1/(\rho_1+\rho_2)$, and $f(\rho_1/(\rho_1+\rho_2)) > \rho_1/(\rho_1+\rho_2)$ if $\rho_1/(\rho_1+\rho_2) \leqslant 1 - (\rho_1+\rho_2)^{-1}$. Hence by mean value theorem, there exists a unique solution a_1 satisfying

$$0 < a_1 < \min(1 - (\rho_1+\rho_2)^{-1}, \rho_1/(\rho_1+\rho_2)).$$

Moreover, it is easy to check that $f'(x) = \mathrm{O}(1)$ for $x \in [0, 1 - (\rho_1+\rho_2)^{-1}]$. Hence there exists a constant $\tau > 0$, such that

$$\frac{\rho_1}{\rho_1+\rho_2}\tau \leqslant a_1 \leqslant \min\left\{(1-\gamma_n) - \frac{\rho_1}{\rho_1+\rho_2}\tau, \frac{\rho_1}{\rho_1+\rho_2}(1-\tau)\right\}, \quad \tau < a_1 \leqslant 1 - \frac{1}{\rho_1+\rho_2} - \frac{\rho_1}{\rho_1+\rho_2}\tau. \tag{B.19}$$

Next, we prove the existence and uniqueness of the solution to the self-consistent equation (3.12) for a general $z \in \mathbf{D}$. Denote by

$$M_1(z) := -\frac{\rho_1+\rho_2}{\rho_1} a_1(z), \quad M_2(z) := -\frac{\rho_1+\rho_2}{\rho_2} a_2(z), \tag{B.20}$$

which are the asymptotic limits of $m_1(z)$ and $m_2(z)$ in equation (B.3). Then, equation (3.12) can be written as the unique solution to the following system of equations

$$\frac{1}{M_1} = \frac{\gamma_n}{p}\sum_{i=1}^{p}\frac{\lambda_i^2}{z+\lambda_i^2 r_1 M_1 + r_2 M_2} - 1, \quad \frac{1}{M_2} = \frac{\gamma_n}{p}\sum_{i=1}^{p}\frac{1}{z+\lambda_i^2 r_1 M_1 + r_2 M_2} - 1, \tag{B.21}$$

such that $\operatorname{Im} M_1(z) > 0$ and $\operatorname{Im} M_2(z) > 0$ for $z \in \mathbb{C}_+$, where, for simplicity of notations, we introduced the following ratios

$$\gamma_n := \frac{p}{n} = \frac{1}{\rho_1+\rho_2}, \quad r_1 := \frac{n_1}{n} = \frac{\rho_1}{\rho_1+\rho_2}, \quad r_2 := \frac{n_2}{n} = \frac{\rho_2}{\rho_1+\rho_2}. \tag{B.22}$$

One can compare equation (B.21) for $(M_1(z), M_2(z))$ to equation (B.3) for $(m_1(z), m_2(z))$. First we prove the following lemma, which gives the existence and uniqueness of the solution $(M_1(z), M_2(z))$ to (B.21).

Lemma B.7. *There exist constants $c_0, C_0 > 0$ depending only on τ in Assumption 2.2 such that the following statements hold. There exists a unique solution to equation* (B.21) *under the conditions*

$$|z| \leqslant c_0, \quad |M_1(z) - M_1(0)| + |M_2(z) - M_2(0)| \leqslant c_0. \tag{B.23}$$

Moreover, the solution satisfies

$$|M_1(z) - M_1(0)| + |M_2(z) - M_2(0)| \leqslant C_0|z|. \tag{B.24}$$

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 (B.21) is equivalent to

$$r_1 M_1 = -(1-\gamma_n) - r_2 M_2 - z\left(M_2^{-1}+1\right), \quad g_z(M_2(z)) = 1, \tag{B.25}$$

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 (B.23). We abbreviate $\delta(z) := M_2(z) - M_2(0)$. From equation (B.25), 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

$$\delta^{(k+1)} = -\frac{g_z(M_2(0)) - g_0(M_2(0))}{g_z'(M_2(0))} - \frac{g_z(M_2(0)+\delta^{(k)}) - g_z(M_2(0)) - g_z'(M_2(0))\delta^{(k)}}{g_z'(M_2(0))}. \tag{B.26}$$

Then equation (B.26) defines a mapping $h_z : \mathbb{C} \to \mathbb{C}$, which maps $\delta^{(k)}$ to $\delta^{(k+1)} = h(\delta^{(k)})$.

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 $|\delta_2| \leqslant \widetilde{c}$,

$$\left|\frac{1}{g_z'(M_2(0))}\right| \leqslant \widetilde{C}, \quad \left|\frac{g_z(M_2(0)) - g_0(M_2(0))}{g_z'(M_2(0))}\right| \leqslant \widetilde{C}|z|, \tag{B.27}$$

and

$$\left|\frac{g_z(M_2(0)+\delta_1)-g_z(M_2(0)+\delta_2)-g_z'(M_2(0))(\delta_1-\delta_2)}{g_z'(M_2(0))}\right| \leqslant \widetilde{C}|\delta_1-\delta_2|^2. \tag{B.28}$$

By equations (B.27) and (B.28), 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 the second equation in (B.25) subject to the condition $\|\delta\|_\infty \leqslant d$.

From the iteration relation (B.26), using equation (B.27) one can readily check that

$$\delta^{(k+1)} - \delta^{(k)} = h_z(\delta^{(k)}) - h_z(\delta^{(k-1)}) \leqslant \widetilde{C}|\delta^{(k)} - \delta^{(k-1)}|^2. \tag{B.29}$$

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 (B.23) as $c_0 = \min\{c_1, d\}$. After obtaining $M_2(z)$, we can then find $M_1(z)$ using the first equation in (B.25).

Note that with equation (B.28) and $\delta^{(0)} = 0$, we get from equation (B.26) that $|\delta^{(1)}| \leqslant \widetilde{C}|z|$. With the contraction mapping, we have the bound

$$|\delta| \leqslant \sum_{k=0}^{\infty} |\delta^{(k+1)} - \delta^{(k)}| \leqslant 2\widetilde{C}|z| \;\Rightarrow\; |M_2(z) - M_2(0)| \leqslant 2\widetilde{C}|z|. \tag{B.30}$$

Using the first equation in equation (B.25), we immediately obtain the bound $r_1|M_1(z) - M_1(0)| \leqslant C|z|$ for some constant $C > 0$, which concludes equation (B.24) 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 (B.21) and treat $M_1(z)$ as the solution to the following equation:

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

(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 the one between equations (B.25) and (B.30), we can conclude $|M_2(z) - M_2(0)| = \mathrm{O}(|z|)$, which further concludes equation (B.24) together with equation (B.30). 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 Theorem B.5. Roughly speaking, it states that if two complex functions $m_1(z)$ and $m_2(z)$ satisfy the self-consistent equation (B.21) approximately up to some small error, then $m_1(z)$ and $m_2(z)$ will be close to the solutions $M_1(z)$ and $M_2(z)$. It will be applied to (B.3) 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 B.8. *There exist constants $c_0, C_0 > 0$ depending only on τ in Assumption 2.2 such that the self-consistent equations in* (B.21) *are stable in the following sense. Suppose $|z| \leqslant c_0$, and $m_1 : \mathbb{C} \mapsto \mathbb{C}$ and $m_2 : \mathbb{C} \mapsto \mathbb{C}$ are analytic functions of z such that*

$$|m_1(z) - M_1(0)| + |m_2(z) - M_2(0)| \leqslant c_0. \tag{B.31}$$

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

$$\frac{1}{m_1} + 1 - \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{\lambda_i^2}{z + \lambda_i^2 r_1 m_1 + r_2 m_2} = \mathcal{E}_1, \quad \frac{1}{m_2} + 1 - \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{1}{z + \lambda_i^2 r_1 m_1 + r_2 m_2} = \mathcal{E}_2, \tag{B.32}$$

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

$$|m_1(z) - M_1(z)| + |m_2(z) - M_2(z)| \leqslant C_0\delta(z). \tag{B.33}$$

Proof. Under condition (B.31), we can obtain equation (B.25) approximately up to some small error:

$$r_1 m_1 = -(1-\gamma_n) - r_2 m_2 - z\left(m_2^{-1}+1\right) + \widetilde{\mathcal{E}}_1(z), \quad g_z(m_2(z)) = 1 + \widetilde{\mathcal{E}}_2(z), \tag{B.34}$$

where the errors satisfy that $|\widetilde{\mathcal{E}}_1(z)| + |\widetilde{\mathcal{E}}_2(z)| = \mathrm{O}(\theta(z))$. Then we subtract equation (B.25) from equation (B.34), 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 B.7, so we omit the details. □

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

In this section, we prove Lemma B.5 by extending our proof 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 $V = \mathrm{Id}$, and the entries of $Z^{(1)}\Lambda$ and $Z^{(2)}$ are independent. For this case, we use the standard resolvent methods (Bloemendal et al., 2014; Yang, 2019; Pillai and Yin, 2014) and prove the following result.

Proposition B.9. *In the setting of Lemma B.5, assume further that the entries of $n^{-1/2}Z^{(1)}$ and $n^{-1/2}Z^{(2)}$ are i.i.d. Gaussian random variables, which satisfy the bounded support condition with $q = n^{-1/2}$. Suppose U and V are identity. Then, the estimate* (B.8) *holds for all $z \in \mathbf{D}$.*

Next we briefly describe how to extend Theorem B.5 from the Gaussian case to the case with general $Z^{(1)}$ and $Z^{(2)}$ satisfying the bounded support condition (B.5) with $q \leqslant n^{-\phi}$ for some constant $\phi > 0$. With Proposition B.9, it suffices is to prove that for $Z^{(1)}$ and $Z^{(2)}$ satisfying the assumptions in Theorem B.5, we have

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

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 B.10 (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

$$\prod_{i\in\mathcal{I}_0}\prod_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2} \rho^\theta_{\mu i}(\mathrm{d}Z^\theta_{\mu i}). \tag{B.35}$$

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 B.5. We shall prove equation (B.8) through interpolation matrices Z^θ between Z^0 and Z^1. We have seen that equation (B.8) holds for Z^0 by Proposition B.9. Using the definition in (B.35) and fundamental calculus, we get the following basic interpolation formula: for differentiable $F : \mathbb{R}^{n\times p} \to \mathbb{C}$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}F(Z^\theta) = \sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}\left[\mathbb{E}F\left(Z^{\theta,Z^1_{\mu i}}_{(\mu i)}\right) - \mathbb{E}F\left(Z^{\theta,Z^0_{\mu i}}_{(\mu i)}\right)\right], \tag{B.36}$$

provided all the expectations exist. We shall apply equation (B.36) to $F(Z) := F^s_{\mathbf{u}\,\mathbf{v}}(Z,z)$ 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 (B.36): for any fixed $s \in 2\mathbb{N}$, any constant $c > 0$ and all $\theta \in [0,1]$,

$$\sum_{i\in\mathcal{I}_0}\sum_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2}\left[\mathbb{E}F^s_{\mathbf{u}\,\mathbf{v}}\left(Z^{\theta,Z^1_{\mu i}}_{(\mu i)},z\right) - \mathbb{E}F^s_{\mathbf{u}\,\mathbf{v}}\left(Z^{\theta,Z^0_{\mu i}}_{(\mu i)},z\right)\right] \leqslant (n^c q)^s + C\mathbb{E}F^s_{\mathbf{u}\,\mathbf{v}}\left(Z^\theta,z\right) \tag{B.37}$$

for some constant $C > 0$. If equation (B.37) holds, then combining equation (B.36) 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 (B.8). Underlying the proof of the estimate (B.37) is an expansion approach, which is very similar to the ones for Lemma 7.10 of (Knowles and Yin, 2016) and Lemma 6.11 of (Yang, 2019). So we omit the details. □

Now it remains to prove Proposition B.9, whose proof is based on the following entrywise local law, Lemma B.11.

Lemma B.11. *Under the assumptions of Proposition B.9, the following estimate holds uniformly in $z \in \mathbf{D}$:*

$$\max_{a,b\in\mathcal{I}}|G_{ab}(z) - \mathfrak{G}_{ab}(z)| \prec n^{-1/2}. \tag{B.38}$$

With Lemma B.11, we can complete the proof of the anisotropic local law (B.8) in Proposition B.9.

Proof of Proposition B.9. With estimate (B.38), one can use the polynomialization method in (Bloemendal et al., 2014, Section 5) to get the anisotropic local law (B.8) with $q = n^{-1/2}$. The proof is exactly the same, except for some minor differences in notations. Hence we omit the details. □

B.4 An Entrywise Local Law

Finally, this subsection is devoted to the proof of Lemma B.11. We first collect some preliminary results, Lemmas B.12-D.2, that will be used in the proof. We remark that theses results work under the general setting in Theorem B.5, that is, we do not require $Z^{(1)}$ and $Z^{(2)}$ to be Gaussian as in Lemma B.11.

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

Lemma B.12. *In the setting of Lemma B.5, there exists a constant $C > 0$ such that the following estimates hold uniformly in $z, z' \in \mathbf{D}$ with overwhelming probability:*

$$\|G(z)\| \leqslant C, \tag{B.39}$$

and

$$\|G(z) - G(z')\| \leqslant C|z - z'|. \tag{B.40}$$

Proof. Our proof is a simple application of the spectral decomposition of G. Recall the matrix A defined in Section 3. Let

$$A = \sum_{k=1}^{p} \sqrt{\mu_k}\xi_k\zeta_k^\top, \quad \mu_1 \geqslant \mu_2 \geqslant \ldots \geqslant \mu_p \geqslant 0 = \mu_{p+1} = \ldots = \mu_n, \tag{B.41}$$

be the 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 the definition of the resolve $G(z)$, we get that for $i, j \in \mathcal{I}_1$ and $\mu, \nu \in \mathcal{I}_1 \cup \mathcal{I}_2$,

$$G_{ij} = \sum_{k=1}^{p} \frac{\xi_k(i)\xi_k^\top(j)}{\mu_k - z}, \quad G_{\mu\nu} = z\sum_{k=1}^{n} \frac{\zeta_k(\mu)\zeta_k^\top(\nu)}{\mu_k - z}, \quad G_{i\mu} = G_{\mu i} = \sum_{k=1}^{p} \frac{\sqrt{\mu_k}\xi_k(i)\zeta_k^\top(\mu)}{\mu_k - z}. \tag{B.42}$$

By Fact D.3, we have that with overwhelming probability $\mu_p \geqslant \lambda_p((Z^{(2)})^\top Z^{(2)}) \geqslant c_\tau$ for some constant $c_\tau > 0$ depending only on τ. ≪HZ notes: to check≫ 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 (B.42), we can easily conclude estimates (B.39) and (B.40). □

For the rest of this subsection, we present the proof of Lemma B.11, which is the most technical part of the whole proof.

Proof of Lemma B.11. Recall that under the assumptions of Lemma B.11, we have

$$A \overset{d}{=} n^{-1/2}(\Lambda(Z^{(1)})^\top, (Z^{(2)})^\top), \tag{B.43}$$

and it suffices to consider the resolvent in equation (B.1) throughout the whole proof. The proof is divided into three steps. For simplicity, we introduce the following notations: for two (deterministic or random) nonnegative quantities ξ and ζ, we write $\xi \lesssim \zeta$ if there exists a constant $C > 0$ such that $\xi \leqslant C\zeta$, and 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 (Erdős et al., 2013c, Section 3) and (Knowles and Yin, 2016, Section 5), we introduce the $\mathcal{Z}$ variables

$$\mathcal{Z}_a := (1 - \mathbb{E}_a)\big(G_{aa}\big)^{-1},$$

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 (D.1), we get that for $i \in \mathcal{I}_0$,

$$\mathcal{Z}_i = \frac{\lambda_i^2}{n}\sum_{\mu,\nu\in\mathcal{I}_1} G_{\mu\nu}^{(i)}\left(\delta_{\mu\nu} - Z_{\mu i}^{(1)}Z_{\nu i}^{(1)}\right) + \frac{1}{n}\sum_{\mu,\nu\in\mathcal{I}_2} G_{\mu\nu}^{(i)}\left(\delta_{\mu\nu} - Z_{\mu i}^{(2)}Z_{\nu i}^{(2)}\right) - 2\frac{\lambda_i}{n}\sum_{\mu\in\mathcal{I}_1,\nu\in\mathcal{I}_2} Z_{\mu i}^{(1)}Z_{\nu i}^{(2)}G_{\mu\nu}^{(i)}, \tag{B.44}$$

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

$$\mathcal{Z}_\mu = \frac{1}{n}\sum_{i,j\in\mathcal{I}_0} \lambda_i\lambda_j G_{ij}^{(\mu)}\left(\delta_{ij} - Z_{\mu i}^{(1)}Z_{\mu j}^{(1)}\right), \quad \mathcal{Z}_\nu = \frac{1}{n}\sum_{i,j\in\mathcal{I}_0} G_{ij}^{(\nu)}\left(\delta_{ij} - Z_{\nu i}^{(2)}Z_{\nu j}^{(2)}\right). \tag{B.45}$$

Moreover, we introduce the random error

$$\Lambda_o := \max_{a\neq b}\left|G_{aa}^{-1}G_{ab}\right|, \tag{B.46}$$

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 B.13. *Under the assumptions of Proposition B.9, the estimate*

$$\Lambda_o + \max_{a\in\mathcal{I}} |\mathcal{Z}_a| \prec n^{-1/2} \tag{B.47}$$

holds uniformly in all $z \in \mathbf{D}$.

Proof. Note that for any $a \in \mathcal{I}$, $H^{(a)}$ and $G^{(a)}$ also satisfies the assumptions in Lemma B.12. Hence equations (B.39) and (B.40) also hold for $G^{(a)}$. Now applying bounds (D.6) and (D.7) to equations (B.44) and (B.45), we get that for any $i \in \mathcal{I}_0$,

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

where in the last step we used equation (B.39) to get that for any $\mu \in \mathcal{I}_1 \cup \mathcal{I}_2$,

$$\sum_{\nu\in\mathcal{I}_1\cup\mathcal{I}_2} |G^{(i)}_{\mu\nu}|^2 \leqslant \sum_{a\in\mathcal{I}} |G^{(i)}_{\mu a}|^2 = \left[G^{(i)}(G^{(i)})^*\right]_{\mu\mu} = \mathrm{O}(1), \quad \text{with overwhelming probability,} \tag{B.48}$$

where $(G^{(i)})^*$ denotes the complex conjugate transpose of $G^{(a)}$. Similarly, applying bounds (D.6) and (D.7) to $\mathcal{Z}_\mu$ and $\mathcal{Z}_\nu$ in equation (B.45) and using the estimate (B.39), we can obtain the same bound. Then we prove the off-diagonal estimate on Λ_o. For $i \in \mathcal{I}_1$ and $a \in \mathcal{I} \setminus \{i\}$, using equation (D.2), Lemma D.4 and equation (B.39), we can obtain that

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

We can get the same estimate for $\left|G^{-1}_{\mu\mu}G_{\mu b}\right|$, $\mu \in \mathcal{I}_1 \cup \mathcal{I}_2$ and $b \in \mathcal{I} \setminus \{\mu\}$, using a similar argument. Thus we obtain that $\Lambda_o \prec n^{-1/2}$, which concludes equation (B.47). □

Note that combining $\max_a |G_{aa}| = \mathrm{O}(1)$ by (B.39) with equation (B.47), we immediately conclude equation (B.38) for $a \neq b$.

Step 2: Self-consistent equations. In this step, we derive the approximate self-consistent equations satisfied by $m_1(z)$ and $m_2(z)$ defined in equation (B.2). More precisely, we will show that $(m_1(z), m_2(z))$ satisfies equation (B.32) for some small errors $|\mathcal{E}_1| + |\mathcal{E}_2| \prec n^{-1/2}$. Then in Step 3, we will apply Lemma B.8 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

$$\Xi(z) := \left\{|m_1(z) - M_1(z)| + |m_2(z) - M_2(z)| \leqslant (\log n)^{-1/2}\right\}. \tag{B.49}$$

Note that by equation (B.24), we have

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

for $z \in \mathbf{D}$. Together with equations (B.19) and the assumption that the singular values λ_i are bounded, we obtain the following estimates

$$|M_1| \sim |M_2| \sim 1, \quad |z + \lambda_i^2 r_1 M_1 + r_2 M_2| \sim 1, \quad \text{uniformly in } z \in \mathbf{D}. \tag{B.50}$$

Moreover, using equation (B.21) we get

$$|1 + \gamma_n M(z)| \sim |M_2^{-1}(z)| \sim 1, \quad |1 + \gamma_n M_0(z)| \sim |M_2^{-1}(z)| \sim 1, \tag{B.51}$$

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

$$M(z) := -\frac{1}{p}\sum_{i=1}^{p} \frac{1}{z + \lambda_i^2 r_1 M_1(z) + r_2 M_2(z)}, \quad M_0(z) := -\frac{1}{p}\sum_{i=1}^{p} \frac{\lambda_i^2}{z + \lambda_i^2 r_1 M_1(z) + r_2 M_2(z)}, \tag{B.52}$$

which are the asymptotic limit of $m(z)$ and $m_0(z)$, respectively. Plugging equation (B.50) into equation (3.11), we get that

$$|\mathfrak{G}_{aa}(z)| \sim 1 \ \text{ uniformly in } z \in \mathbf{D}, \ a \in \mathcal{I}. \tag{B.53}$$

Then we prove the following key lemma, which shows that $(m_1(z), m_2(z))$ satisfies equation (B.32) with some small errors $\mathcal{E}_1$ and $\mathcal{E}_2$.

Lemma B.14. *Under the assumptions of Proposition B.9, the following estimates hold uniformly in $z \in \mathbf{D}$:*

$$\mathbf{1}(\Xi)\left|\frac{1}{m_1} + 1 - \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{\lambda_i^2}{z + \lambda_i^2 r_1 m_1 + r_2 m_2}\right| \prec n^{-1/2}, \tag{B.54}$$

and

$$\mathbf{1}(\Xi)\left|\frac{1}{m_2} + 1 - \frac{\gamma_n}{p}\sum_{i=1}^{p} \frac{1}{z + \lambda_i^2 r_1 m_1 + r_2 m_2}\right| \prec n^{-1/2}. \tag{B.55}$$

Proof. By equations (D.1), (B.44) and (B.45), we obtain that

$$\frac{1}{G_{ii}} = -z - \frac{\lambda_i^2}{n}\sum_{\mu\in\mathcal{I}_1} G^{(i)}_{\mu\mu} - \frac{1}{n}\sum_{\mu\in\mathcal{I}_2} G^{(i)}_{\mu\mu} + \mathcal{Z}_i = -z - \lambda_i^2 r_1 m_1 - r_2 m_2 + \mathcal{E}_i, \quad \text{for } i \in \mathcal{I}_0, \tag{B.56}$$

$$\frac{1}{G_{\mu\mu}} = -1 - \frac{1}{n}\sum_{i\in\mathcal{I}_0} \lambda_i^2 G^{(\mu)}_{ii} + \mathcal{Z}_\mu = -1 - \gamma_n m_0 + \mathcal{E}_\mu, \quad \text{for } \mu \in \mathcal{I}_1, \tag{B.57}$$

$$\frac{1}{G_{\nu\nu}} = -1 - \frac{1}{n}\sum_{i\in\mathcal{I}_0} G^{(\nu)}_{ii} + \mathcal{Z}_\nu = -1 - \gamma_n m + \mathcal{E}_\nu, \quad \text{for } \nu \in \mathcal{I}_2, \tag{B.58}$$

where we denoted (recall equation (B.2) and Definition B.1)

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

and

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

Now using equations (D.3), (B.46) and (B.47), 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 (B.39) in the last step. Similarly, we can get 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 (B.47), we obtain the bound

$$\max_{i\in\mathcal{I}_0} |\mathcal{E}_i| + \max_{\mu\in\mathcal{I}_1\cup\mathcal{I}_2} |\mathcal{E}_\mu| \prec n^{-1/2}. \tag{B.59}$$

With equation (B.50) and the definition of the event Ξ, we get that

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

Combining it with equations (B.56) and (B.59), we obtain that

$$\mathbf{1}(\Xi)G_{ii} = \mathbf{1}(\Xi)\left[-\frac{1}{z+\lambda_i^2 r_1 m_1 + r_2 m_2} + \mathrm{O}_{\prec}\left(n^{-1/2}\right)\right]. \tag{B.60}$$

Plugging (B.60) into the definitions of m and m_0 in equation (B.2), we get

$$\mathbf{1}(\Xi)m = \mathbf{1}(\Xi)\left[-\frac{1}{p}\sum_{i\in\mathcal{I}_0}\frac{1}{z+\lambda_i^2 r_1 m_1 + r_2 m_2} + \mathrm{O}_{\prec}\left(n^{-1/2}\right)\right], \tag{B.61}$$

$$\mathbf{1}(\Xi)m_0 = \mathbf{1}(\Xi)\left[-\frac{1}{p}\sum_{i\in\mathcal{I}_0}\frac{\lambda_i^2}{z+\lambda_i^2 r_1 m_1 + r_2 m_2} + \mathrm{O}_{\prec}\left(n^{-1/2}\right)\right]. \tag{B.62}$$

As a byproduct, we obtain from these two estimates and equation (B.52) that

$$\mathbf{1}(\Xi)\left(|m(z)-M(z)| + |m_0(z)-M_0(z)|\right) \lesssim (\log n)^{-1/2}, \quad \text{with overwhelming probability on } \Xi. \tag{B.63}$$

Together with equation (B.51), we get that

$$|1+\gamma_n m_0(z)| \sim 1, \quad |1+\gamma_n m(z)| \sim 1, \quad \text{with overwhelming probability on } \Xi. \tag{B.64}$$

Now combining equations (B.57), (B.58), (B.59) and (B.64), we obtain that for $\mu \in \mathcal{I}_1$ and $\nu \in \mathcal{I}_2$,

$$\mathbf{1}(\Xi)\left(G_{\mu\mu} + \frac{1}{1+\gamma_n m_0}\right) = \mathrm{O}_{\prec}\left(n^{-1/2}\right), \quad \mathbf{1}(\Xi)\left(G_{\nu\nu} + \frac{1}{1+\gamma_n m}\right) = \mathrm{O}_{\prec}\left(n^{-1/2}\right). \tag{B.65}$$

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

$$\mathbf{1}(\Xi)\left(m_1 + \frac{1}{1+\gamma_n m_0}\right) = \mathrm{O}_{\prec}\left(n^{-1/2}\right), \quad \mathbf{1}(\Xi)\left(m_2 + \frac{1}{1+\gamma_n m}\right) = \mathrm{O}_{\prec}\left(n^{-1/2}\right), \tag{B.66}$$

which further implies

$$\mathbf{1}(\Xi)\left(\frac{1}{m_1} + 1 + \gamma_n m_0\right) \prec n^{-1/2}, \quad \mathbf{1}(\Xi)\left(\frac{1}{m_2} + 1 + \gamma_n m\right) \prec n^{-1/2}. \tag{B.67}$$

Finally, plugging equations (B.61) and (B.62) into equation (B.67), we conclude equations (B.54) and (B.55). □

Step 3: Ξ holds with overwhelming probability. In this step, we show that the event $\Xi(z)$ in (B.49) actually holds with overwhelming probability for all $z \in \mathbf{D}$. Once we have proved this fact, applying Lemma B.8 to equations (B.54) and (B.55) 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

$$|m_1(0) - M_1(0)| + |m_2(0) - M_2(0)| \prec n^{-1/2}. \tag{B.68}$$

In fact, notice that by equations (B.24) and (B.40) 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 (B.68) holds, we can obtain that

$$\sup_{z\in\mathbf{D}}\left(|m_1(z) - M_1(z)| + |m_2(z) - M_2(z)|\right) \lesssim (\log n)^{-1} \quad \text{with overwhelming probability,} \tag{B.69}$$

and

$$\sup_{z\in\mathbf{D}}\left(|m_1(z) - M_1(0)| + |m_2(z) - M_2(0)|\right) \lesssim (\log n)^{-1} \quad \text{with overwhelming probability.} \tag{B.70}$$

The equation (B.69) shows that Ξ holds with overwhelming probability, while the equation (B.70) verifies the condition (B.31) of Lemma B.8. Now applying Lemma B.8 to equations (B.54) and (B.55), we obtain that

$$|m_1(z) - M_1(z)| + |m_2(z) - M_2(z)| \prec n^{-1/2} \tag{B.71}$$

uniformly for all $z \in \mathbf{D}$. Together with equations (B.65) and (B.66), equation (B.71) implies that

$$\max_{\mu\in\mathcal{I}_1} |G_{\mu\mu}(z) - M_1(z)| + \max_{\nu\in\mathcal{I}_2} |G_{\nu\nu}(z) - M_2(z)| \prec n^{-1/2}. \tag{B.72}$$

Then plugging estimate (B.71) into equation (B.60) and recalling (B.20), we obtain that

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

Together with equation (B.72), it gives the diagonal estimate

$$\max_{a\in\mathcal{I}} |G_{aa}(z) - \Pi_{aa}(z)| \prec n^{-1/2}. \tag{B.73}$$

Combining equation (B.73) with the off-diagonal estimate on Λ_o in equation (B.47), we conclude the proof of Lemma B.11. □

Finally, we give the proof of equation (B.68).

Proof of equation (B.68). By equation (B.42), we get

$$m(0) = \frac{1}{p}\sum_{i\in\mathcal{I}_0} G_{ii}(0) = \frac{1}{p}\sum_{i\in\mathcal{I}_0}\sum_{k=1}^{p} \frac{|\xi_k(i)|^2}{\lambda_k} \geqslant \lambda_1^{-1} \gtrsim 1,$$

where we used Fact D.3 in the last step. ≪HZ notes: to check≫ Similarly, we can also get that $m_0(0)$ is positive and has size $m_0(0) \sim 1$. Hence we have the estimates

$$1 + \gamma_n m(0) \sim 1, \quad 1 + \gamma_n m_0(0) \sim 1.$$

Combining these estimates with equations (B.57), (B.58) and (B.59), we obtain that equation (B.66) 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 (B.56) and (B.59), we obtain that equations (B.61) and (B.62) also hold at $z = 0$ even without the indicator function $\mathbf{1}(\Xi)$. Finally, plugging equations (B.61) and (B.62) into equation (B.67), we conclude that equations (B.54) and (B.55) hold at $z = 0$, that is,

$$\begin{aligned}
&\left|\frac{1}{m_1(0)} + 1 - \frac{1}{n}\sum_{i=1}^{p} \frac{\lambda_i^2}{\lambda_i^2 r_1 m_1(0) + r_2 m_2(0)}\right| \prec n^{-1/2}, \\
&\left|\frac{1}{m_2(0)} + 1 - \frac{1}{n}\sum_{i=1}^{p} \frac{1}{\lambda_i^2 r_1 m_1(0) + r_2 m_2(0)}\right| \prec n^{-1/2}.
\end{aligned} \tag{B.74}$$

Denoting $y_1 = -m_1(0)$ and $y_2 = -m_2(0)$, by equation (B.67) 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 there exists a constant $c > 0$ such that

$$c \leqslant y_1 \leqslant 1, \quad c \leqslant y_2 \leqslant 1, \quad \text{with overwhelming probability.} \tag{B.75}$$

Also one can verify from equation (B.74) that $(r_1 y_1, r_2 y_2)$ satisfies approximately the same system of equations as equation (3.5):

$$r_1 y_1 + r_2 y_2 = 1 - \gamma_n + \mathrm{O}_{\prec}(n^{-1/2}), \quad r_1^{-1} f(r_1 y_1) = 1 + \mathrm{O}_{\prec}(n^{-1/2}), \tag{B.76}$$

where recall that the function f was defined in equation (B.18). The first equation of (B.76) and equation (B.75) together imply that $y_1 \in [0, r_1^{-1}(1-\gamma_n)]$ with overwhelming probability. For the second equation of (B.76), 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 (B.76), 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 (B.68). □

C Proof of Corollary 3.3

We follow a similar logic from Theorem 2.1. We characterize the global minimizer of $f(A, B)$ in the random-effects model. Based on the characterization, we reduce the prediction loss of hard parameter sharing to the asymptotic limits provided in Theorem 3.1. Then, we prove Corollary 3.3 based on the bias and variance limits. We set up several notations. In the two-task case, the optimization objective $f(A, B)$ is equal to

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

where $B \in \mathbb{R}^p$ and $A = [A_1, A_2] \in \mathbb{R}^2$ because the width of B is one. We assume that A_1 and A_2 are both nonzero. Otherwise, we add a tiny amount of perturbation δ to them and the result remains the same. Using the local optimality condition $\frac{\partial f}{\partial B} = 0$, we have that $\hat{B}$ satisfies the following

$$\hat{B} := (A_1^2 (X^{(1)})^\top X^{(1)} + A_2^2 (X^{(2)})^\top X^{(2)})^{-1} (A_1 (X^{(1)})^\top Y^{(1)} + A_2 (X^{(2)})^\top Y^{(2)}). \tag{C.2}$$

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

$$\begin{aligned} g(x) := & \left\| X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} (x\beta_2 - \beta_1) + \left(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1} \right) \varepsilon^{(1)} \right. \\ & \left. + x X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top \varepsilon^{(2)} \right\|^2 \\ & + \left\| X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} (x\beta_1 - x^2 \beta_2) + \left(X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top - \mathrm{Id}_{n_2} \right) \varepsilon^{(2)} \right. \\ & \left. + x X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top \varepsilon^{(1)} \right\|^2. \end{aligned} \tag{C.3}$$

The conditional expectation of $g(x)$ over ε_1 and ε_2 is

$$\begin{aligned} \mathop{\mathbb{E}}_{\varepsilon^{(1)}, \varepsilon^{(2)}} [g(x) \mid X_1, X_2, \beta_1, \beta_2] = & \left\| X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} (\beta_1 - x\beta_2) \right\|^2 \\ & + x^2 \left\| X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} (\beta_1 - x\beta_2) \right\|^2 \\ & + \mathop{\mathbb{E}}_{\varepsilon^{(1)}} \left[(\varepsilon^{(1)})^\top \left(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1} \right)^2 \varepsilon^{(1)} \right] \\ & + \mathop{\mathbb{E}}_{\varepsilon^{(2)}} \left[(\varepsilon^{(2)})^\top \left(X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top - \mathrm{Id}_{n_2} \right)^2 \varepsilon^{(2)} \right] \\ & + x^2 \mathop{\mathbb{E}}_{\varepsilon^{(2)}} \left[(\varepsilon^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top \varepsilon^{(2)} \right] \\ & + x^2 \mathop{\mathbb{E}}_{\varepsilon^{(1)}} \left[(\varepsilon^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top \varepsilon^{(1)} \right]. \end{aligned} \tag{C.4}$$

Using the following identity

$$(X^{(1)})^\top X^{(1)}\hat{\Sigma}(x)^{-1}(X^{(2)})^\top X^{(2)} = \left(x^2[(X^{(2)})^\top X^{(2)}]^{-1} + [(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1}$$
$$= (X^{(2)})^\top X^{(2)}\hat{\Sigma}(x)^{-1}(X^{(1)})^\top X^{(1)},$$

we can simplify that

$$\begin{aligned}
&\left\|X^{(1)}\hat{\Sigma}(x)^{-1}(X^{(2)})^\top X^{(2)}(\beta_1 - x\beta_2)\right\|^2 + x^2\left\|X^{(2)}\hat{\Sigma}(x)^{-1}(X^{(1)})^\top X^{(1)}(\beta_1 - x\beta_2)\right\|^2 \\
=&(\beta_1 - x\beta_2)^\top (X^{(2)})^\top X^{(2)}\hat{\Sigma}(x)^{-1}(X^{(1)})^\top X^{(1)}\hat{\Sigma}(x)^{-1}(X^{(2)})^\top X^{(2)}(\beta_1 - x\beta_2) \\
&+ (\beta_1 - x\beta_2)^\top x^2(X^{(1)})^\top X^{(1)}\hat{\Sigma}(x)^{-1}(X^{(2)})^\top X^{(2)}\hat{\Sigma}(x)^{-1}(X^{(1)})^\top X^{(1)}(\beta_1 - x\beta_2) \\
=&(\beta_1 - x\beta_2)^\top (X^{(1)})^\top X^{(1)}\hat{\Sigma}(x)^{-1}(X^{(2)})^\top X^{(2)}(\beta_1 - x\beta_2).
\end{aligned}$$

Using this identity, we can simplify equation (C.4) to

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\varepsilon^{(1)},\varepsilon^{(2)}}[g(x) \mid X_1, X_2, \beta_1, \beta_2] \\
=&(\beta_1 - x\beta_2)^\top (X^{(1)})^\top X^{(1)}\hat{\Sigma}(x)^{-1}(X^{(2)})^\top X^{(2)}(\beta_1 - x\beta_2) \\
&+ \sigma^2 \operatorname{Tr}\left[\left(x^2 X^{(1)}\hat{\Sigma}(x)^{-1}(X^{(1)})^\top - \mathrm{Id}_{n_1}\right)^2 + \left(X^{(2)}\hat{\Sigma}(x)^{-1}(X^{(2)})^\top - \mathrm{Id}_{n_2}\right)^2\right] \\
&+ 2x^2\sigma^2 \operatorname{Tr}\left[\hat{\Sigma}(x)^{-1}(X^{(1)})^\top X^{(1)}\hat{\Sigma}(x)^{-1}(X^{(2)})^\top X^{(2)}\right] \\
=&(\beta_1 - x\beta_2)^\top (X^{(1)})^\top X^{(1)}\hat{\Sigma}(x)^{-1}(X^{(2)})^\top X^{(2)}(\beta_1 - x\beta_2) + \sigma^2(n_1 + n_2 - p).
\end{aligned}$$

For the random-effects model, recall that the entries of $\beta_1, \beta_2 \in \mathbb{R}^p$ are i.i.d. Gaussian random variables with mean zero and variance $\kappa^2/p, d^2/p$, respectively. Hence,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\varepsilon^{(1)},\varepsilon^{(2)},\beta_1,\beta_2}[g(x) \mid X_1, X_2] \\
&= \frac{(x-1)^2\kappa^2 + (x^2+1)d^2/2}{p} \operatorname{Tr}\left[(X^{(1)})^\top X^{(1)}\hat{\Sigma}(x)^{-1}(X^{(2)})^\top X^{(2)}\right] + \sigma^2(n_1 + n_2 - p),
\end{aligned} \tag{C.5}$$

Part 1: characterizing the global minimum of $f(A, B)$. Let $\hat{x}$ denote the 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 $f(A, B)$, combined with the local optimality condition for $\hat{B}$.

First, we show that $g(x)$ and $\mathbb{E}\left[g(x) \mid X^{(1)}, X^{(2)}\right]$ are close using standard concentration bounds.

Claim C.1. *In the setting of Corollary 3.3, we have that with high probability*

$$\left|g(x) - \mathop{\mathbb{E}}_{\varepsilon^{(1)},\varepsilon^{(2)},\beta_1,\beta_2}[g(x) \mid X_1, X_2]\right| \leqslant p^{1/2+c}\left(\sigma^2 + \kappa^2 + d^2\right).$$

Proof. By Fact D.3(ii), we have that with high probability,

$$\|X^{(1)}\| \leqslant \sqrt{(\sqrt{n_1} + \sqrt{p})^2 + n_1 \cdot p^{-c_\varphi}} \lesssim \sqrt{p}, \quad \|X^{(2)}\| \leqslant \sqrt{(\sqrt{n_2} + \sqrt{p})^2 + n_2 \cdot p^{-c_\varphi}} \lesssim \sqrt{p}, \tag{C.6}$$

and

$$\|\hat{\Sigma}(x)^{-1}\| \leqslant \frac{1}{x^2[(\sqrt{n_1} - \sqrt{p})^2 - n_1 \cdot p^{-c_\varphi}] + [(\sqrt{n_2} - \sqrt{p})^2 - n_2 \cdot p^{-c_\varphi}]} \lesssim \frac{1}{(x^2+1)p}, \tag{C.7}$$

where we used that $1 + \tau \leqslant \rho_1 = n_1/p \leqslant \tau^{-1}$ and $1 + \tau \leqslant \rho_2 = n_2/p \leqslant \tau^{-1}$ for a small constant $\tau > 0$.

Now we expand the first term on the right-hand side of (C.3):

$$\begin{aligned}
&\Big\|X^{(1)}\hat{\Sigma}(x)^{-1}(X^{(2)})^\top X^{(2)}(x\beta_2 - \beta_1) + \left(x^2 X^{(1)}\hat{\Sigma}(x)^{-1}(X^{(1)})^\top - \mathrm{Id}_{n_1}\right)\varepsilon^{(1)} \\
&+ xX^{(1)}\hat{\Sigma}(x)^{-1}(X^{(2)})^\top \varepsilon^{(2)}\Big\|^2 = h_1(x) + h_2(x) + h_3(x) + 2h_4(x) + 2h_5(x) + 2h_6(x),
\end{aligned} \tag{C.8}$$

where

$$
\begin{aligned}
h_1(x) &:= (\beta_1 - x\beta_2)^\top (X^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} (\beta_1 - x\beta_2), \\
h_2(x) &:= (\varepsilon^{(1)})^\top \left(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1} \right)^2 \varepsilon^{(1)}, \\
h_3(x) &:= x^2 (\varepsilon^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top \varepsilon^{(2)}, \\
h_4(x) &:= (\varepsilon^{(1)})^\top \left(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1} \right) X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} (x\beta_2 - \beta_1), \\
h_5(x) &:= x (\varepsilon^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} (x\beta_2 - \beta_1), \\
h_6(x) &:= x (\varepsilon^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top \left(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1} \right) \varepsilon^{(1)}.
\end{aligned}
$$

Next we estimate each term using Lemma D.4 in Appendix D.

For $h_1(x)$, using Lemma D.4 in Appendix D 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 $p^{-1}[(x-1)^2\kappa^2 + (x^2+1)d^2/2]$, we obtain the following estimate with high probability for any small constant $c > 0$:

$$
\begin{aligned}
&\left| h_1(x) - \mathop{\mathbb{E}}_{\beta_1, \beta_2} [h_1(x) \mid X_1, X_2] \right| \\
&\leqslant p^c \cdot p^{-1}[(x-1)^2\kappa^2 + (x^2+1)d^2/2] \cdot \left\| (X^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} \right\|_F \\
&\leqslant p^c \cdot p^{-1}[(x-1)^2\kappa^2 + (x^2+1)d^2/2] \cdot p^{1/2} \left\| (X^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} \right\| \\
&\lesssim p^{1/2+c} \cdot \frac{(x-1)^2\kappa^2 + (x^2+1)d^2/2}{(x^2+1)^2} \lesssim p^{1/2+c}(\kappa^2 + d^2).
\end{aligned} \tag{C.9}
$$

Here in the third step we used (C.6) and (C.7) to bound the operator norm:

$$
\begin{aligned}
\left\| (X^{(2)})^\top X^{(2)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} \right\| &\leqslant \left\| (X^{(2)})^\top X^{(2)} \right\|^2 \cdot \left\| (X^{(1)})^\top X^{(1)} \right\| \cdot \left\| \hat{\Sigma}(x)^{-1} \right\|^2 \\
&\lesssim \frac{p}{(x^2+1)^2}.
\end{aligned}
$$

Similarly, using Lemma D.4, the fact that 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, and bounds (C.6)-(C.7), we can obtain that with high probability,

$$
\left| h_2(x) - \mathop{\mathbb{E}}_{\varepsilon^{(1)}} [h_2(x) \mid X_1, X_2] \right| \lesssim p^{1/2+c}\sigma^2, \quad \left| h_3(x) - \mathop{\mathbb{E}}_{\varepsilon^{(2)}} [h_3(x) \mid X_1, X_2] \right| \lesssim p^{1/2+c}\sigma^2. \tag{C.10}
$$

For $h_4(x)$, using Lemma D.4 in Appendix D, we obtain the following estimate with high probability for any small constant $c > 0$:

$$
\begin{aligned}
&|h_4(x)| \\
&\leqslant p^c \cdot \sigma \sqrt{p^{-1}[(x-1)^2\kappa^2 + (x^2+1)d^2/2]} \cdot \left\| \left(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1} \right) X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} \right\|_F \\
&\leqslant p^c \cdot \sigma \sqrt{p^{-1}[(x-1)^2\kappa^2 + (x^2+1)d^2/2]} \cdot p^{1/2} \left\| \left(x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top - \mathrm{Id}_{n_1} \right) X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} \right\| \\
&\lesssim p^c \cdot \sigma \sqrt{(x-1)^2\kappa^2 + (x^2+1)d^2/2} \cdot \left\| \mathrm{Id}_{n_1} - x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top \right\| \left\| X^{(1)} \right\| \left\| \hat{\Sigma}(x)^{-1} \right\| \left\| (X^{(2)})^\top X^{(2)} \right\| \\
&\lesssim p^{1/2+c} \frac{\sigma \sqrt{(x-1)^2\kappa^2 + (x^2+1)d^2/2}}{x^2+1} \lesssim p^{1/2+c}(\sigma^2 + \kappa^2 + d^2),
\end{aligned} \tag{C.11}
$$

where in the fourth step we used $\left\| \mathrm{Id}_{n_1} - x^2 X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(1)})^\top \right\| \leqslant 1$ and (C.6)-(C.7), and in the last step we used AM-GM inequality. With same argument, we can show that

$$
|h_5(x)| \leqslant p^{1/2+c}(\sigma^2 + \kappa^2 + d^2), \quad |h_6(x)| \leqslant p^{1/2+c}\sigma^2, \tag{C.12}
$$

with high probability for any small constant $c > 0$.

Finally, we can obtain similar estimates for the second term on the right-hand side of (C.3) as in (C.9), (C.10), (C.11) and (C.12). Hence the proof is complete. □

Next, we show that $\hat{x}$ is close to 1.

Claim C.2. *In the setting of Corollary 3.3, we have that with high probability,*

$$|\hat{x} - 1| \leqslant \frac{2d^2}{\kappa^2} + p^{-1/4+c}. \tag{C.13}$$

Proof. Corresponding to equation (C.5), we define the function

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

Using Fact D.3(ii), we can obtain that with high probability,

$$\begin{aligned} p(x^2+1)^{-1} &\lesssim \left(x^2\left[(\sqrt{n_2}-\sqrt{p})^2 - n_2 p^{-c_\varphi}\right]^{-1} + \left[(\sqrt{n_1}-\sqrt{p})^2 - n_1 p^{-c_\varphi}\right]^{-1}\right)^{-1} \\ &\preceq (X^{(1)})^\top X^{(1)} \hat{\Sigma}(x)^{-1} (X^{(2)})^\top X^{(2)} = \left(x^2[(X^{(2)})^\top X^{(2)}]^{-1} + [(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1} \\ &\preceq \left(x^2\left[(\sqrt{n_2}+\sqrt{p})^2 + n_2 p^{-c_\varphi}\right]^{-1} + \left[(\sqrt{n_1}+\sqrt{p})^2 + n_1 p^{-c_\varphi}\right]^{-1}\right)^{-1} \lesssim p(x^2+1)^{-1}, \end{aligned} \tag{C.14}$$

where we also used that $1 + \tau \leqslant n_1/p \leqslant \tau^{-1}$ and $1 + \tau \leqslant n_2/p \leqslant \tau^{-1}$ for a small constant $\tau > 0$. Then we get that

$$h(1) = d^2 \cdot p^{-1} \operatorname{Tr}\left[(X^{(1)})^\top X^{(1)} \hat{\Sigma}(1)^{-1} (X^{(2)})^\top X^{(2)}\right] \lesssim pd^2. \tag{C.15}$$

On the other hand, if $|x-1| \geqslant \delta$ for a small constant $\delta > 0$, then we have

$$h(x) \gtrsim (1+x^2)\kappa^2 \cdot p\left(x^2+1\right)^{-1} = p\kappa^2, \tag{C.16}$$

where in the first step we used that $|x-1|^2 \gtrsim 1 + x^2$ for $|x-1| \geqslant \delta$ and (C.14). Hence using Claim C.1 and equations (C.15)-(C.16), we obtain that with high probability,

$$\begin{aligned} g(x) &\geqslant h(x) + \sigma^2(n_1+n_2-p) - p^{-1/2+e_1} \cdot p\left(\sigma^2+\kappa^2+d^2\right) \\ &\geqslant h(1) + \sigma^2(n_1+n_2-p) + p^{-1/2+e_1} \cdot p\left(\sigma^2+\kappa^2+d^2\right) \geqslant g(1), \end{aligned} \tag{C.17}$$

under conditions $\sigma^2 \lesssim \kappa^2$ and $d^2 \leqslant p^{-c_\varphi}\kappa^2$. This gives that $|\hat{x} - 1| \leqslant \delta$ for any small constant $\delta > 0$.

To obtain the better bound (C.13), we study the function $h(x)$ more closely and find its minimizer, denoted by $x^\star$. First it is easy to observe that $h(x) < h(-x)$ for $x > 0$. Then we consider the case $x \geqslant 1$. We write

$$h(x) := [(1-x^{-1})^2\kappa^2 + (1+x^{-2})d^2/2] \cdot p^{-1} \operatorname{Tr}\left[\left([(X^{(2)})^\top X^{(2)}]^{-1} + x^{-2}[(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1}\right].$$

Notice that

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

is an increasing functions in x. Hence taking derivative of $h(x)$ with respect to x, we obtain that

$$h'(x) \geqslant \left[2(1-x^{-1})\frac{\kappa^2}{x^2} - \frac{d^2}{x^3}\right] \cdot p^{-1} \operatorname{Tr}\left[\left([(X^{(2)})^\top X^{(2)}]^{-1} + x^{-2}[(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1}\right] \geqslant 0, \tag{C.18}$$

as long as $x \geqslant 1 + d^2/(2\kappa^2)$. Next we consider the case $1 - \delta \leqslant x \leqslant 1$. Notice that

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

is a decreasing functions in x. Hence taking derivative of $h(x)$, we obtain that

$$h'(x) \leqslant [-2(1-x)\kappa^2 + xd^2] \cdot p^{-1} \operatorname{Tr}\left[\left(x^2[(X^{(2)})^\top X^{(2)}]^{-1} + [(X^{(1)})^\top X^{(1)}]^{-1}\right)^{-1}\right] \leqslant 0, \tag{C.19}$$

as long as $x \leqslant 1 - d^2/(2\kappa^2)$. In sum, we see that the minimizer $x^\star$ must satisfy

$$1 - d^2/(2\kappa^2) \leqslant x^\star \leqslant 1 + d^2/(2\kappa^2).$$

Now we are ready to prove equation (C.13). First, using equations (C.14) and (C.18), we obtain that for $1 + d^2/\kappa^2 \leqslant x \leqslant 1 + \delta$,

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

Thus we get that if $1 + 2d^2/\kappa^2 + p^{-1/4+c} \leqslant x \leqslant 1 + \delta$,

$$h(x) - h(1) \geqslant h(x) - h(1 + d^2/\kappa^2) \geqslant \int_{1+d^2/\kappa^2}^{x} h'(x)\mathrm{d}x \gtrsim p\kappa^2 \cdot |x-1|^2 \gtrsim p\frac{d^4}{\kappa^2} + p^{-1/2+2c} \cdot p\kappa^2.$$

On the other hand, by Claim C.1 we have that

$$g(x) = h(x) + \sigma^2(n_1 + n_2 - p) + \mathrm{O}\left(p^{1/2+c}\left(\sigma^2 + \kappa^2 + d^2\right)\right)$$

with high probability. Combining the above two estimates, we obtain that under the conditions $\sigma^2 \lesssim \kappa^2$ and $d^2 \leqslant p^{-c_\varphi}\kappa^2$,

$$g(x) - g(1) = h(x) - h(1) + \mathrm{O}\left(p^{1/2+c}\left(\sigma^2 + \kappa^2 + d^2\right)\right) > 0$$

with high probability, i.e. x cannot be the minimizer of g. Second, using equations (C.14) and (C.19), we obtain that for $1 - \delta \leqslant x \leqslant 1 - d^2/\kappa^2$,

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

Then with a similar argument as above, we can get that if $1 - 2d^2/\kappa^2 - p^{-1/4+c} \leqslant x \leqslant 1 - \delta$, then

$$h(x) - h(1) \gtrsim p\frac{d^4}{\kappa^2} + p^{1/2+2c}\kappa^2,$$

which again implies that $g(x) > g(1)$, i.e. x cannot be the minimizer of g. In sum, we obtain that the minimizer $\hat{x}$ must satisfy equation (C.13). □

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 $\hat{B}\hat{A}_2$. The predication loss of hard parameter sharing is as follows

$$\begin{aligned} L(\hat{\beta}_2^{\mathrm{HPS}}) &= \left\|\Sigma_2^{1/2}\left(\hat{B}\hat{A}_2 - \beta_2\right)\right\| \\ &= \left\|(\Sigma^{(2)})^{1/2}\hat{\Sigma}(\hat{x})^{-1}\left[(X^{(1)})^\top X^{(1)}(\hat{x}\beta_1 - \hat{x}^2\beta_2) + (X^{(2)})^\top \varepsilon^{(2)} + \hat{x}(X^{(1)})^\top \varepsilon^{(1)}\right]\right\|^2. \end{aligned} \tag{C.20}$$

Using Lemma C.2 and the concentration estimates in Lemma D.4, we can simplify $L(\hat{\beta}_2^{\mathrm{HPS}})$ as in the following claim.

Claim C.3. *Recall that $\hat{\Sigma}(1)$ is equal to $\hat{\Sigma}$ (cf. Section 3). In the setting of Proposition 3.3, we have that with high probability*

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

Proof. Our proof is divided into two steps. First, with the concentration estimates in Lemma D.4, we can simplify $L(\hat{\beta}_2^{\mathrm{HPS}})$ as follows. With high probability, we have that for any small constant $c > 0$,

$$\left| L(\hat{\beta}_2^{\mathrm{HPS}}) - \mathcal{L}(\hat{x}) \right| \leqslant p^{-1/2+c} \left(\sigma^2 + \kappa^2 + d^2 \right), \tag{C.21}$$

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

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

We are ready to finish the proof of Claim C.3. We defer the proof of equation (C.21) until the end. Next, we can further simplify $\mathcal{L}(\hat{x})$ using Claim C.2 and $\Sigma^{(1)} = \Sigma^{(2)} = \mathrm{Id}_p$. More precisely, we claim that with high probability

$$\begin{aligned} &\left| \mathcal{L}(\hat{x}) - d^2 \cdot p^{-1} \operatorname{Tr}\left[\hat{\Sigma}^{-2} \left((X^{(1)})^\top X^{(1)} \right)^2 \right] - \sigma^2 \operatorname{Tr}\left[\hat{\Sigma}^{-1} \right] \right| \\ &\lesssim \frac{d^4 + \sigma^2 d^2}{\kappa^2} + p^{-1/2+2c} \kappa^2 + p^{-1/4+c} (\sigma^2 + d^2). \end{aligned} \tag{C.22}$$

Combining (C.21) and (C.22), we obtain Claim C.3.

It remains to prove equation (C.22). Using Lemma C.2 and equations (C.6)-(C.7), we obtain that with high probability,

$$\|\hat{\Sigma}^{-1} - \hat{\Sigma}(\hat{x})^{-1}\| \leqslant |\hat{x}^2 - 1| \|\hat{\Sigma}^{-1}\| \|(X^{(1)})^\top X^{(1)}\| \|\hat{\Sigma}(\hat{x})^{-1}\| \lesssim p^{-1} \left(\frac{d^2}{\kappa^2} + p^{-1/4+c} \right). \tag{C.23}$$

Using similar arguments, we get that with high probability,

$$\left\| \left(\hat{\Sigma}^{-2} - \hat{\Sigma}(\hat{x})^{-2} \right) \left((X^{(1)})^\top X^{(1)} \right)^2 \right\| \lesssim \frac{d^2}{\kappa^2} + p^{-1/4+c}. \tag{C.24}$$

Moreover, using equations (C.6)-(C.7) we can bound that with high probability,

$$\operatorname{Tr}\left[\hat{\Sigma}^{-2} \left((X^{(1)})^\top X^{(1)} \right)^2 \right] \leqslant p \left\| \hat{\Sigma}^{-2} \left((X^{(1)})^\top X^{(1)} \right)^2 \right\| \lesssim p. \tag{C.25}$$

Now we bound the left hand side of equation (C.22) as

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

Applying the estimates (C.13), (C.23), (C.24) and (C.25) to the three terms on the right-hand side, we can conclude (C.22).

Now we give the proof of the estimate (C.21). The proof of this claim is very similar to Claim C.1, where the only difference is that $\hat{x}$ now may depend on $\varepsilon^{(1)}$, $\varepsilon^{(2)}$, β_1 and β_2. In this case, we can still use Lemma D.4 to conclude the proof because $\hat{x}$ can be pulled out as a coefficient. Here we only give a proof sketch and omit the details. Recall that β_0 is the shared component of β_1 and β_2 with i.i.d. Gaussian entries of mean zero and variance $p^{-1}\kappa^2$. Moreover, we denote the task-specific components by $\widetilde{\beta}_1$ and $\widetilde{\beta}_2$, whose entries are i.i.d. Gaussian random variables of mean zero and variance $p^{-1}d^2/2$. Then we write $L(\hat{\beta}_2^{\mathrm{HPS}})$ in (C.20) as:

$$\begin{aligned} L(\hat{\beta}_2^{\mathrm{HPS}}) =& \left\| (\Sigma^{(2)})^{1/2} \hat{\Sigma}(\hat{x})^{-1} \left[(X^{(1)})^\top X^{(1)} (\hat{x} - \hat{x}^2) \beta_0 + (X^{(1)})^\top X^{(1)} \hat{x} \widetilde{\beta}_1 - (X^{(1)})^\top X^{(1)} \hat{x}^2 \widetilde{\beta}_2 \right] \right. \\ &\left. + (\Sigma^{(2)})^{1/2} \hat{\Sigma}(\hat{x})^{-1} \left[(X^{(2)})^\top \varepsilon^{(2)} + \hat{x} (X^{(1)})^\top \varepsilon^{(1)} \right] \right\|^2. \end{aligned} \tag{C.26}$$

As in (C.8), we can expand this expression into the sum of 15 terms, and bound each term as in (C.9)-(C.12). For example, for the main 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 D.4 and equation (C.6)-(C.7), we can obtain the following estimate with high probability:

$$\begin{aligned}
&\left|\widetilde{\beta}_1^\top (X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(1)})^\top X^{(1)}\widetilde{\beta}_1 - \frac{d^2}{2}\cdot p^{-1}\operatorname{Tr}\left[(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(1)})^\top X^{(1)}\right]\right| \\
&\leqslant p^{-1+c}d^2\cdot\left\|(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(1)})^\top X^{(1)}\right\|_F \\
&\leqslant p^{-1/2+c}d^2\cdot\left\|(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(1)})^\top X^{(1)}\right\| \lesssim p^{-1/2+c}d^2.
\end{aligned}$$

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 D.4 and (C.6)-(C.7), we can obtain the following estimate with high probability for any small constant $c>0$:

$$\begin{aligned}
\left|\widetilde{\beta}_1^\top (X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(2)})^\top\varepsilon^{(2)}\right| &\leqslant p^c\cdot\sigma\sqrt{p^{-1}d^2}\cdot\|(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(2)})^\top\|_F \\
&\lesssim p^c\sigma d\cdot\|(X^{(1)})^\top X^{(1)}\hat{\Sigma}(\hat{x})^{-1}\Sigma^{(2)}\hat{\Sigma}(\hat{x})^{-1}(X^{(2)})^\top\| \\
&\lesssim p^{-1/2+c}\sigma d\leqslant p^{-1/2+c}(\sigma^2+d^2).
\end{aligned}$$

The rest of the terms in the expansion of (C.26) can be bounded in the same way, and we omit the details. □

Part 3: applying the bias and 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 $d^2\cdot p^{-1}\operatorname{Tr}\left[\hat{\Sigma}^{-2}\left((X^{(1)})^\top X^{(1)}\right)^2\right]$ using Theorem 3.1.

Proof of Corollary 3.3. Using (3.4), we obtain that

$$\operatorname{Tr}[\hat{\Sigma}^{-1}] = \operatorname{Tr}\left[\left((X^{(1)})^\top X^{(1)} + (X^{(2)})^\top X^{(2)}\right)^{-1}\right] = \operatorname{Tr}\left[\frac{(a_1\operatorname{Id}_p + a_2\operatorname{Id}_p)^{-1}}{n_1+n_2}\right] + \mathrm{O}(p^{-c_\varphi}) \tag{C.27}$$

with high probability. Solving equation (3.5) with $\lambda_i\equiv 1$, $1\leqslant i\leqslant p$, we get that

$$a_1 = \frac{\rho_1(\rho_1+\rho_2-1)}{(\rho_1+\rho_2)^2},\quad a_2 = \frac{\rho_2(\rho_1+\rho_2-1)}{(\rho_1+\rho_2)^2}. \tag{C.28}$$

Applying the above to equation (C.27), we obtain that

$$\operatorname{Tr}\left[\hat{\Sigma}^{-1}\right] = \frac{p}{n_1+n_2}\cdot\frac{\rho_1+\rho_2}{\rho_1+\rho_2-1} + \mathrm{O}(p^{-c_\varphi}) = \frac{1}{\rho_1+\rho_2-1} + \mathrm{O}(p^{-c_\varphi}) \tag{C.29}$$

with high probability.

On the other hand, we have that with high probability,

$$\begin{aligned}
\frac{(\sqrt{n_1}-\sqrt{p})^4\cdot(1-p^{-c_\varphi})}{p}\sum_{i=1}^p\left(\hat{\Sigma}^{-2}\right)_{ii} &\leqslant p^{-1}\operatorname{Tr}\left[\hat{\Sigma}^{-2}\left((X^{(1)})^\top X^{(1)}\right)^2\right] \\
&\leqslant \frac{(\sqrt{n_1}+\sqrt{p})^4\cdot(1+p^{-c_\varphi})}{p}\sum_{i=1}^p\left(\hat{\Sigma}^{-2}\right)_{ii}.
\end{aligned} \tag{C.30}$$

To obtain this inequality, we used

$$(\sqrt{n_1}-\sqrt{p})^4\cdot(1-p^{-c_\varphi}) \preceq \left((X^{(1)})^\top X^{(1)}\right)^2 \preceq (\sqrt{n_1}+\sqrt{p})^4\cdot(1+p^{-c_\varphi}) \quad \text{with high probability,}$$

by Fact D.3(ii), and the fact that for the product of two PSD matrices, its trace is always nonnegative. Using (3.6) with $\Sigma^{(1)}=\Sigma^{(2)}=\Lambda=V=\operatorname{Id}_p$ and w being the i-th coordinate vector, we can calculate that

$$\left(\hat{\Sigma}^{-2}\right)_{ii} = \frac{1}{(n_1+n_2)^2}\cdot\frac{a_3+a_4+1}{(a_1+a_2)^2} + \mathrm{O}(p^{-c_\varphi}) \tag{C.31}$$

with high probability. Solving equation (3.7) with a_1, a_2 in (C.28) and $\lambda_i \equiv 1$, $1 \leqslant i \leqslant p$, we can obtain that

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

Inserting them into (C.31), we obtain that

$$\left(\hat{\Sigma}^{-2}\right)_{ii} = \frac{1}{(n_1+n_2)^2} \cdot \frac{(\rho_1+\rho_2)^3}{(\rho_1+\rho_2-1)^3} + \mathrm{O}(p^{-c_\varphi})$$

with high probability. In fact, we need this estimate to hold simultaneously for all $1 \leqslant i \leqslant p$ with high probability. For this purpose, we shall use (B.16) to get that

$$\left| \left(\hat{\Sigma}^{-2}\right)_{ii} - \frac{1}{(n_1+n_2)^2} \cdot \frac{(\rho_1+\rho_2)^3}{(\rho_1+\rho_2-1)^3} \right| \prec p^{-\frac{\varphi-4}{2\varphi}}$$

on a high probability event that does not depend on i. Then using Fact B.3 (i), we obtain that

$$\left| \sum_{i=1}^{p} \left(\hat{\Sigma}^{-2}\right)_{ii} - \frac{p}{(n_1+n_2)^2} \cdot \frac{(\rho_1+\rho_2)^3}{(\rho_1+\rho_2-1)^3} \right| \prec p^{1-\frac{\varphi-4}{2\varphi}}$$

with high probability, which by Definition B.2 gives that

$$\left| p^{-1} \sum_{i=1}^{p} \left(\Sigma(1)^{-2}\right)_{ii} - \frac{1}{(n_1+n_2)^2} \cdot \frac{(\rho_1+\rho_2)^3}{(\rho_1+\rho_2-1)^3} \right| \leqslant p^{-c_\varphi} \tag{C.32}$$

with high probability. Inserting (C.32) into (C.30), we get that

$$\begin{aligned}
&\frac{(\sqrt{n_1}-\sqrt{p})^4 \cdot (1-p^{-c_\varphi}) - n_1^2 \cdot (1+\mathrm{O}(p^{-c_\varphi}))}{(n_1+n_2)^2} \cdot \frac{(\rho_1+\rho_2)^3}{(\rho_1+\rho_2-1)^3} \\
\leqslant & p^{-1} \operatorname{Tr}\left[\hat{\Sigma}^{-2}\left((X^{(1)})^\top X^{(1)}\right)^2\right] - \frac{n_1^2}{(n_1+n_2)^2} \cdot \frac{(\rho_1+\rho_2)^3}{(\rho_1+\rho_2-1)^3} \\
\leqslant & \frac{(\sqrt{n_1}+\sqrt{p})^4 \cdot (1+p^{-c_\varphi}) - n_1^2 \cdot (1+\mathrm{O}(p^{-c_\varphi}))}{(n_1+n_2)^2} \cdot \frac{(\rho_1+\rho_2)^3}{(\rho_1+\rho_2-1)^3}
\end{aligned}$$

with high probability.

Combining equation (C.29) and the above with Claim C.3, we get that

$$\begin{aligned}
\left| \mathcal{L}(\hat{x}) - \frac{\sigma^2}{\rho_1+\rho_2-1} - d^2 \cdot \frac{\rho_1^2(\rho_1+\rho_2)}{(\rho_1+\rho_2-1)^3} \right| \leqslant & \left[\left(1+\frac{1}{\sqrt{\rho_1}}\right)^4 - 1 \right] d^2 \cdot \frac{\rho_1^2(\rho_1+\rho_2)}{(\rho_1+\rho_2-1)^3} \\
& + C\left(p^{-c_\varphi}(\sigma^2+d^2) + \frac{d^4+\sigma^2 d^2}{\kappa^2} + p^{-1/2+2c}\kappa^2 + p^{-1/4+c}(\sigma^2+d^2) \right).
\end{aligned}$$

Then using Claim C.21 and the conditions $\sigma^2 \lesssim \kappa^2$ and $d^2 \leqslant p^{-c_\varphi}\kappa^2$, we conclude the proof. □

D Tools

D.1 Algebraic Inequalities

Fact D.1. The minimum singular value of a matrix $X^\top Y X$ is at least the minimum singular value of Y times the minimum singular value of $X^\top X$.

The following simple resolvent identities are important tools for our proof. Recall that the resolvent minors have been defined in Definition B.1.

Lemma D.2. *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

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

(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

$$G_{ia} = -G_{ii}\left(AG^{(i)}\right)_{ia}, \quad G_{\mu b} = -G_{\mu\mu}\left(A^\top G^{(\mu)}\right)_{\mu b}. \tag{D.2}$$

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

$$G^{(a)}_{a_1 a_2} = G_{a_1 a_2} - \frac{G_{a_1 a} G_{a a_2}}{G_{aa}}. \tag{D.3}$$

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

D.2 Random Variables with Bounded Moments

Fact D.3. In the setting of Theorem 2.1, with high probability over the randomness of X, we have that

(i) When n/p converges to $\rho > 1$, the sample covariance matrix $\frac{X^\top X}{n}$ is full rank.

(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).

The following lemma gives (almost) sharp concentration bounds for linear and quadratic forms of bounded supported random variables. Here we recall that the stochastic domination "$\prec$" was defined in Definition B.2.

Lemma D.4 (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 variances at most 1, and x_i and y_j satisfy the bounded support condition (B.5) for a deterministic parameter q. Then we have the following bounds:*

$$\Big|\sum_{i=1}^n A_i x_i\Big| \prec q \max_i |A_i| + \Big(\sum_i |A_i|^2\Big)^{1/2}, \quad \Big|\sum_{i,j=1}^n x_i B_{ij} y_j\Big| \prec q^2 B_d + q n^{1/2} B_o + \Big(\sum_{i\neq j} |B_{ij}|^2\Big)^{1/2}, \tag{D.4}$$

$$\Big|\sum_{i=1}^n (|x_i|^2 - \mathbb{E}|x_i|^2) B_{ii}\Big| \prec q n^{1/2} B_d, \quad \Big|\sum_{1\leqslant i\neq j\leqslant n} \bar{x}_i B_{ij} x_j\Big| \prec q n^{1/2} B_o + \Big(\sum_{i\neq j} |B_{ij}|^2\Big)^{1/2}, \tag{D.5}$$

where $B_d := \max_i |B_{ii}|$ 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 stronger bounds

$$\Big|\sum_i A_i x_i\Big| \prec \Big(\sum_i |A_i|^2\Big)^{1/2}, \quad \Big|\sum_{i,j} x_i B_{ij} y_j\Big| \prec \Big(\sum_{i\neq j} |B_{ij}|^2\Big)^{1/2}, \tag{D.6}$$

$$\Big|\sum_i (|x_i|^2 - \mathbb{E}|x_i|^2) B_{ii}\Big| \prec \Big(\sum_i |B_{ii}|^2\Big)^{1/2}, \quad \Big|\sum_{i\neq j} \bar{x}_i B_{ij} x_j\Big| \prec \Big(\sum_{i\neq j} |B_{ij}|^2\Big)^{1/2}. \tag{D.7}$$

As a corollary, we can obtain the following concentration estimates for random variables with bounded φ-th moment.

Corollary D.5. *Let $Z \in \mathbb{R}^{n\times p}$ be a random matrix as in Assumption 2.2. Then for any deterministic vector $v \in \mathbb{R}^p$, we have that*

$$\big|\|Zv\|^2 - n\|v\|^2\big| \leqslant n^{1-c_\varphi} \|v\|^2 \tag{D.8}$$

with high probability.

Proof. We introduce the truncated matrices $\widetilde{Z}$ with entries $\widetilde{Z}_{ij} := \mathbf{1}\left(|Z_{ij}| \leqslant q\right) \cdot Z_{ij}$ for $q = n^{1/2-\frac{\varphi-4}{2\varphi}} \log n$. By equation (B.6), we have that

$$\mathbb{P}(\widetilde{Z} = Z) = 1 - \mathrm{O}((\log n)^{-\varphi}). \tag{D.9}$$

Furthermore, by equation (B.12) we have that

$$|\mathbb{E}\widetilde{Z}_{ij}| = \mathrm{O}(n^{-3/2}), \quad \mathbb{E}|\widetilde{Z}_{ij}|^2 = 1 + \mathrm{O}(n^{-1}). \tag{D.10}$$

Then we can centralize and rescale $\widetilde{Z}$ as $\widehat{Z} := \frac{\widetilde{Z} - \mathbb{E}\widetilde{Z}}{(\mathbb{E}|\widetilde{Z}_{11}|^2)^{1/2}}$. It suffices to show that

$$\left| \|\widehat{Z}v\|^2 - n\|v\|^2 \right| \prec n^{1/2} q \|v\|^2. \tag{D.11}$$

In fact, provided that (D.11) holds, using (D.10) it is easy to get that

$$\left| \|\widetilde{Z}v\|^2 - n\|v\|^2 \right| \prec n^{1/2} q \|v\|^2,$$

which implies (D.8) for any fixed value c_φ within $(0, \frac{\varphi-4}{2\varphi})$.

For (D.11), we first observe that for $1 \leqslant i \leqslant n$, $(\widehat{Z}v)_i = \sum_{1 \leqslant j \leqslant p} \widehat{Z}_{ij} v_j$ are i.i.d random variables of mean zero and variance $\|v\|^2$. Moreover, using (D.4) we get that

$$|(\widehat{Z}v)_i| \prec q \max_{1 \leqslant i \leqslant p} |v_i| + \|v\|.$$

Together with the trivial $\max_{1 \leqslant i \leqslant p} |v_i| \leqslant \|v\|$, this implies that the random variables $\frac{(\widehat{Z}v)_i}{\|v\|}$, $1 \leqslant i \leqslant p$, are i.i.d random variables of mean zero, variance 1 and bounded support q. Then applying (D.5), we get that

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

This concludes (D.11). □

Next we can obtain the following concentration estimates for random noise in the setting of Section 2.

Corollary D.6. *Suppose the entries of $\mathcal{E} = [\varepsilon^{(1)}, \varepsilon^{(2)}, \cdots, \varepsilon^{(t)}] \in \mathbb{R}^{n \times t}$ are i.i.d. random variables of mean zero, variance σ^2 and bounded moments up to any order. Then for any deterministic vector $v \in \mathbb{R}^n$ and deterministic $n \times n$ matrix $B \in \mathbb{R}^{n \times n}$, we have that with high probability,*

$$|v^\top \varepsilon^{(i)}| \leqslant \sigma n^c \|v\|, \tag{D.12}$$

and

$$\left| (\varepsilon^{(i)})^\top B \varepsilon^{(j)} - \delta_{ij} \cdot \sigma^2 \operatorname{Tr}(B) \right| \leqslant n^c \|B\|_F, \tag{D.13}$$

for any small constant $c > 0$. As a consequence, equations (D.12) *and* (D.13) *imply that*

$$\|v^\top \mathcal{E}\| \leqslant \sigma n^c \|v\|, \quad \text{and} \quad \left\| \mathcal{E}^\top B \mathcal{E} - \sigma^2 \operatorname{Tr}(B) \cdot \mathrm{Id}_{t \times t} \right\|_F \leqslant n^c \|B\|_F, \tag{D.14}$$

with high probability for any small constant $c > 0$.

Proof. Note that $\varepsilon_j^{(i)}/\sigma$, $1 \leqslant i \leqslant t, 1 \leqslant j \leqslant n$, are i.i.d. random variable of mean zero, variable one and bounded moments up to any order. Then using the first estimate in equation (D.6), we can obtain that $|v^\top \varepsilon^{(i)}| = \left| \sum_{j=1}^n v_j \varepsilon_j^{(i)} \right| \prec \sigma \|v\|$, which concludes (D.12) together with the definition of stochastic domination in Definition B.2. Similarly, using the second estimate in equation (D.6), we can obtain that for $i \neq j$,

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

Using the two estimates in equation (D.7), we obtain that

$$\begin{aligned}\Big|(\varepsilon^{(i)})^\top B\varepsilon^{(j)} - \delta_{ij}\cdot\sigma^2\operatorname{Tr}(B)\Big| &\leqslant \Big|\sum_{k=1}\Big(|\varepsilon_k^{(i)}|^2 - \mathbb{E}|\varepsilon_k^{(i)}|^2\Big)B_{ii}\Big| + \Big|\sum_{k\neq l}\varepsilon_k^{(i)}\varepsilon_l^{(i)}B_{kl}\Big| \\ &\prec \sigma^2\Big(\sum_k |B_{kk}|^2\Big) + \sigma^2\Big(\sum_{k\neq l}|B_{kl}|^2\Big)^{1/2} \lesssim \sigma^2\|B\|_F.\end{aligned}$$

The above two estimates conclude (D.13). Finally, the estimates in equation (D.14) are easy consequences of equations (D.12) and (D.13) since t is a fixed integer that does not change with p. □