HIGH-DIMENSIONAL ASYMPTOTICS OF TRANSFER LEARNING FOR LINEAR REGRESSION

By Fan Yang[1] Hongyang R. Zhang[2,†] Sen Wu[3,‡] Weijie J. Su [1,*] and Christopher Ré[3,§]

[1]*Department of Statistics, University of Pennsylvania, fyang75@wharton.upenn.edu; *suw@wharton.upenn.edu*

[2]*Khoury College of Computer Sciences, Northeastern University, †hrzhang@northeastern.edu*

[3]*Department of Computer Science, Stanford University, ‡senwu@stanford.edu; §chrismre@cs.stanford.edu*

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

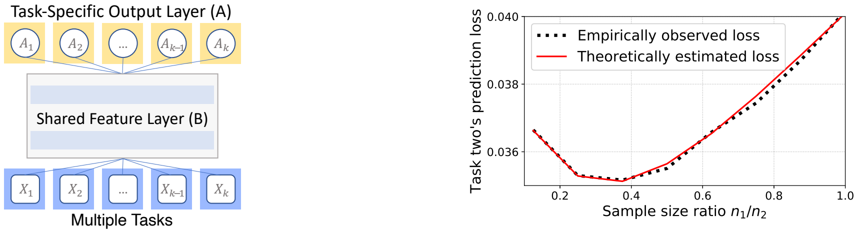
1. Introduction.

rewrite Hard parameter sharing (HPS) for multi-task learning is widely used in empirical research and goes back to the seminal work of [12]. Recent work has revived interests in this approach because it improves performance and reduces the cost of collecting labeled data [34]. It is generally applied by sharing the feature layers between all tasks while keeping an output layer for every task. Often, hard parameter sharing offers two critical advantages if successfully applied. First, it reduces model parameters since all tasks use the same feature space. Second, it reduces the amount of labeled data needed from each task by augmenting the entire training dataset.

Hard parameter sharing works as an inductive transfer mechanism and a regularizer that reduces overfitting, both of which have great intuitive appeal [34]. For example, by restricting the shared space's size, HPS encourages information sharing among multiple tasks [21]. Another source of inductive bias comes from the tasks and depends on datasets' properties such as sample sizes and task covariances [38]. However, how these dataset properties impact HPS has not been well established. Part of the challenge may be that HPS' generalization performance depends intricately on the sample size ratios and covariate shifts between tasks, and is not amenable to standard concentration results. Previous results based on Rademacher complexity or VC dimensions have considered cases where all tasks' sample sizes are equal to logarithmic factors of the feature dimension [8, 27], and when all tasks' sample sizes increase simultaneously [2, 26].

MSC2020 subject classifications: Primary 62J05, 60B20; secondary 62E20, 62H10.

Keywords and phrases: Hard parameter sharing, high-dimensional linear regression, random matrix theory, sample covariance matrices.



(a) A hard parameter sharing architecture

(b) Varying sample size ratio

Fig 1: Left: an illustrative picture of HPS. Right: an illustrative example of using HPS for two tasks X_1, Y_1 and X_2, Y_2 with sample size n_1, n_2, respectively. Increasing n_1/n_2 decreases task two's prediction loss initially but increase afterward. This phenomenon occurs due to different bias-variance tradeoffs as n_1/n_2 increases. Our result provides an estimated loss (solid line) that accurately matches the empirical loss (dotted line). See Section 5.1 for the precise setting.

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

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

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

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

Main results. Our first result (Theorem 4.2) applies to multi-label prediction settings where all datasets have the same features (and sample size), and we want to make several predictions for every input (cf. examples in [18]). We analyze the global minimizer of $f(A, B)$, and provide a sharp bias-variance decomposition of its (out-of-sample) prediction loss for any task. This setting is tractable even though in general, $f(A, B)$ is non-convex in A and B (e.g. matrix completion is a special case for suitably designed $X^{(i)}, Y^{(i)}$). Our result implies that when all tasks have the same features but different labels, for any task, HPS helps reduce the task's variance compared to single-task learning but increases bias.

Our second result (Theorem 3.4) applies to two tasks with arbitrarily different sample size ratios and covariate shifts. While we can no longer characterize $f(A, B)$'s global minimum because of non-convexity, we can still provide a sharp bias-variance tradeoff of any local minimizer's prediction loss for both tasks. Despite being a simple setting, we observe several non-trivial phenomena by varying sample size ratios and covariate shifts between the two tasks. See Figure 1b for an illustration of the former. Consequently, using our precise loss estimates, we observe several qualitative properties of HPS for varying dataset properties.

- *Sample efficiency (Example 4.3)*: One advantage of combining multiple datasets is that the requirement for labeled data reduces compared to single-task learning, a phenomenon that [40] has observed empirically. Our results further imply that HPS's sample efficiency depends on model-specific variances across tasks vs. the noise variance and is generally high when the latter is large.
- *Sample size ratio (Example 3.2)*: Increasing one task's sample size does not always reduce another task's loss. In a simplified setting, we find that the task loss either decreases first before increasing afterward or decreases monotonically depending on how fast the bias grows. These two trends result from different bias-variance tradeoffs. This result is surprising because previous generalization bounds in multi-task learning typically scale down as all tasks' sample sizes increase, thus do not apply for different sample size ratios.
- *Covariate shift (Example 3.5)*: In addition to sample sizes, variance also scales with two datasets' covariate shifts. For a large sample size ratio, HPS's variance is smallest when there is no covariate shift. Counterintuitively, for a small sample size ratio, having covariate shifts reduces variance through a complementary spectrum. We achieve this result through a novel characterization on the inverse of the sum of two sample covariance matrices with arbitrary covariate shifts. See our discussion of proof techniques below for details.

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

Proof techniques. There are two main ideas in our analysis. The proof of our first result uses a geometric intuition that hard parameter sharing finds a "rank-r" approximation of the datasets. We carefully keep track of the concentration error between the global minimizer of $f(A, B)$ and its population version (cf. equation (4.7)). The proof of our second result is significantly more involved because of different sample sizes and covariate shifts. We show that the inverse of the sum of two sample covariance matrices with arbitrary covariate shifts converges to a deterministic diagonal matrix asymptotically (cf. Theorem 3.4). We use recently developed techniques from random matrix theory to show a sharp convergence rate. One limitation of our analysis is that in Example 3.2, there is an error term that can result in vacuous bounds for very small n_1 (cf. equation (**??**)). We believe our result has provided significant initial insights, and it is an interesting question to tighten our result. See Section **??** for more discussions of the technical challenge.

1.2. *Related Work.* There is a large body of classical and recent works on multi-task learning. We focus our discussion on theoretical results and refer interested readers to several excellent surveys for general references [29, 41, 36]. The early work of [8, 10, 26] studied multi-task learning from a theoretical perspective, often using uniform convergence or

Rademacher complexity based techniques. An influential paper by [9] provides uniform convergence bounds that combine multiple datasets in certain settings. One limitation of uniform convergence based techniques is that the results often assume that all tasks have the same sample size, see e.g. [8, 27]. Moreover, these techniques do not apply to the high-dimensional setting because the results usually require a sample size of at least $p\log p$.

Our proof techniques use the so-called local law of random matrices [16], a recent development in the random matrix theory literature. In the single-task case, [11] first proved such a local law for sample covariance matrices with isotropic covariance. [20] later extended this result to arbitrary covariance matrices. These techniques provide almost sharp convergence rates to the asymptotic limit compared to other methods such as free probability [28]. To the best of our knowledge, we are not aware of any previous results in the multi-task case, even for two tasks (with arbitrary covariate shifts).

The problem we study here is also related to high-dimensional prediction in transfer learning [23, 7] and distributed learning [15]. For example, [23] provide minimax-optimal rates to predict a target regression task given multiple sparse regression tasks. One closely related work is [38], which studied hard parameter sharing for two linear regression tasks. However, their results only apply to sample size regimes at least logarithmic factors of dimension.

Organizations. The rest of this paper is organized as follows. In Section 4, we present the bias-variance decomposition for hard parameter sharing. In Section 3.2, we describe how varying sample sizes and covariate shifts impact hard parameter sharing using random matrix theory. In Sections 5.1, we validate our results in simulations. In Section **??**, we summarize our work and discuss future work. Section 5.2 describes our study on text classification tasks. Section C, F, and D present proofs of our results.

Notations. For an $n \times p$ matrix X, let $\lambda_{\min}(X)$ denote its smallest singular value and $\|X\|$ denote its largest singular value. Let $\lambda_1(X), \lambda_2(X), \cdots, \lambda_{p\wedge n}(X)$ denote the singular values of X in decreasing order. Let X^+ denote the Moore-Penrose psuedoinverse of X. We refer to random matrices of the form $\frac{X^\top X}{n}$ as sample covariance matrices. We say that an event Ξ holds with high probability if the probability that Ξ happens goes to 1 as p goes to infinity. We use the big-O notation $g(n) = \mathrm{O}(f(n))$ if there exists a constant C such that $g(n) \leqslant C\dot{f}(n)$ for large enough n. Moreover, we use the notation $g(n) \lesssim f(n)$ if $g(n) = \mathrm{O}(f(n))$, and the notation $g(n) \sim f(n)$ if $g(n) \lesssim f(n))$ and $f(n) \lesssim g(n)$. In this paper, we will often write an identity matrix $\mathrm{Id}_{n\times n}$ as 1 without causing any confusions.

2. Transfer learning for high-dimensional linear regression.

In this section, we define our model for the setting of transfer learning, and introduce several transfer learning estimators that will be considered in this paper.

2.1. *The model.*

Consider the transfer learning setting with two data sets. We denote their sample sizes by n_1 and n_2. For $i = 1, 2$, Let $X^{(i)} \in \mathbb{R}^{n_i \times p}$ denote dataset i's feature covariates. We assume that the label vector $Y^{(i)} \in \mathbb{R}^{n_i}$ for $X^{(i)}$ follows a linear model plus random noise:

$$Y^{(i)} = X^{(i)}\beta^{(i)} + \varepsilon^{(i)}, \quad i = 1, 2.$$

In this paper, we make the following assumptions on feature covariates $X^{(i)}$ and the noise vectors $\varepsilon^{(i)}$. We remark that all these assumptions are natural in high-dimensional statistics.

We assume that the row vectors of $X^{(i)}$ are i.i.d. centered random vectors with population covariance matrix $\Sigma^{(i)}$. More precisely, let

$$X^{(i)} = Z^{(i)}(\Sigma^{(i)})^{1/2} \in \mathbb{R}^{n_i \times p}, \quad i = 1, 2, \tag{2.1}$$

where each $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are $p \times p$ deterministic positive definite symmetric matrices, and $Z^{(1)} = (z_{ij}^{(1)})$ and $Z^{(2)} = (z_{ij}^{(2)})$ are $n_1 \times p$ and $n_2 \times p$ random matrices with real i.i.d. entries satisfying

$$\mathbb{E} z_{ij}^{(1)} = \mathbb{E} z_{ij}^{(2)} = 0, \qquad \mathbb{E}|z_{ij}^{(1)}|^2 = \mathbb{E}|z_{ij}^{(2)}|^2 = 1, \tag{2.2}$$

Furthermore, we assume that the entries $z_{ij}^{(1)}$ and $z_{ij}^{(2)}$ have finite φ-th moment for some constant $\varphi > 4$:

$$\mathbb{E}|z_{ij}^{(1)}|^{\varphi} \leqslant \tau^{-1}, \quad \mathbb{E}|z_{ij}^{(2)}|^{\varphi} \leqslant \tau^{-1} \tag{2.3}$$

for a small constant $\tau > 0$. We assume that $\Sigma^{(i)}$ has eigendecomposition

$$\Sigma^{(i)} = O_i \Lambda_i O_i^{\top}, \quad \Lambda_1 = \text{diag}(\sigma_1^{(i)}, \ldots, \sigma_n^{(i)}), \quad i = 1, 2, \tag{2.4}$$

where O_i is the eigenmatrix and the eigenvalues satisfy that

$$\tau \leqslant \sigma_p^{(i)} \leqslant \cdots \leqslant \sigma_2^{(i)} \leqslant \sigma_1^{(i)} \leqslant \tau^{-1} \quad i = 1, 2. \tag{2.5}$$

We assume that $\varepsilon^{(1)} \in \mathbb{R}^n$ and $\varepsilon^{(2)} \in \mathbb{R}^n$ are two independent random noise vectors with i.i.d entries of mean zero, variance σ^2, and bounded moment up to any order: for any fixed $k \in \mathbb{N}$, there exists a constant $C_k > 0$ such that

$$\mathbb{E}|\varepsilon_i^{(1)}|^k \leqslant C_k, \quad \mathbb{E}|\varepsilon_i^{(2)}|^k \leqslant C_k. \tag{2.6}$$

Finally, we assume that $\beta^{(1)}$ and $\beta^{(2)}$ are two arbitrary deterministic or random vectors that are independent of $X^{(1)}$, $X^{(2)}$, $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$.

In this paper, we consider the high-dimensional setting, where the sample ratios satisfy that

$$1 + \tau \leqslant \rho_1 := \frac{n_1}{p} \leqslant p^{\tau^{-1}}, \quad 1 + \tau \leqslant \rho_2 := \frac{n_2}{p} \leqslant p^{\tau^{-1}}, \quad \tau \leqslant \frac{\rho_1}{\rho_2} \leqslant \tau^{-1}, \tag{2.7}$$

for a small constant $\tau > 0$. If $\rho_i > p^{\tau^{-1}}$, $i = 1, 2$, we are basically in the low-dimensional region, where the law of large numbers and central limit theorem already give good enough results without using the theory developed in this paper. The lower bounds $\rho_1 > 1 + \tau$ and $\rho_2 > 1 + \tau$ are to ensure that the sample covariance matrices $(X^{(1)})^{\top} X^{(1)}$ and $(X^{(2)})^{\top} X^{(2)}$ are non-singular with high probability, so that the ordinary least squares (OLS) estimator is well-defined for the linear regression problem on each task.

We summarize our basic assumptions here for future reference.

ASSUMPTION 2.1. Let τ be a small constant.

(i) $X^{(1)}$ and $X^{(2)}$ take the form (2.1), where $Z^{(1)}$ and $Z^{(2)}$ are respectively $n_1 \times p$ and $n_2 \times p$ random matrices with i.i.d. entries satisfying (2.2) and (2.3), $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are deterministic positive definite symmetric matrices satisfying (2.4) and (2.5).

(ii) $\varepsilon^{(1)} \in \mathbb{R}^n$ and $\varepsilon^{(2)} \in \mathbb{R}^n$ are random vectors independent from $X^{(1)}$ and $X^{(2)}$, and with i.i.d entries of mean zero, variance σ^2, and bounded moments as in (2.6).

(iii) $\beta^{(1)}$ and $\beta^{(2)}$ are independent of $X^{(1)}$, $X^{(2)}$, $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$.

(iv) ρ_1 and ρ_2 satisfy (2.7).

2.2. *Transfer learning estimators.* [FY: Introduce some popular transfer learning estimators: total loss estimator, and model averaging estimator.]

We study the standard hard parameter sharing architecture: a shared feature representation layer $B \in \mathbb{R}^p$ for all datasets and a separate output layer $A_i \in \mathbb{R}$ for every dataset i. Then we study the following minimization problem:

$$f(A,B) = \|X^{(1)}BA_1 - Y^{(1)}\|^2 + \|X^{(2)}BA_2 - Y^{(2)}\|^2, \tag{2.8}$$

where we abbreviate $A = [A_1, A_2]$. Let $(\hat{A}, \hat{B})$ be the minimizer of $f(A,B)$. We define the hard parameter sharing (HPS) estimator for task i as

$$\hat{\beta}_i^{\text{HPS}} = \hat{B}\hat{A}_i, \quad i = 1, 2. \tag{2.9}$$

For the optimization objective $f(A,B)$ in (2.8), using the local optimality condition $\frac{\partial f}{\partial B} = 0$, we can solve that

$$\hat{B} = A_2^{-1}\hat{\Sigma}(a)^{-1}\left[a(X^{(1)})^\top Y^{(1)} + (X^{(2)})^\top Y^{(2)}\right], \tag{2.10}$$

where we denote $a := A_1/A_2$ and $\hat{\Sigma}(a) := a^2(X^{(1)})^\top X^{(1)} + (X^{(2)})^\top X^{(2)}$. Applying $\hat{B}$ to equation (2.8), we obtain an objective that only depends on a as follows

$$\begin{aligned} g(a) := \bigg\| & X^{(1)}\hat{\Sigma}(a)^{-1}(X^{(2)})^\top X^{(2)}(a\beta^{(2)} - \beta^{(1)}) \\ & + \left(a^2 X^{(1)}\hat{\Sigma}(a)^{-1}(X^{(1)})^\top - \mathrm{Id}_{n_1\times n_1}\right)\varepsilon^{(1)} + aX^{(1)}\hat{\Sigma}(a)^{-1}(X^{(2)})^\top\varepsilon^{(2)}\bigg\|^2 \\ + \bigg\| & X^{(2)}\hat{\Sigma}(a)^{-1}(X^{(1)})^\top X^{(1)}(a\beta^{(1)} - a^2\beta^{(2)}) \\ & + \left(X^{(2)}\hat{\Sigma}(a)^{-1}(X^{(2)})^\top - \mathrm{Id}_{n_2\times n_2}\right)\varepsilon^{(2)} + aX^{(2)}\hat{\Sigma}(a)^{-1}(X^{(1)})^\top\varepsilon^{(1)}\bigg\|^2. \end{aligned} \tag{2.11}$$

Let $\hat{a}$ be the minimizer of $g(a)$. Throughout this paper, we regard $Y^{(1)}$ as the source data, and $Y^{(2)}$ as the target data. Then the HPS estimator (2.9) for the target task 2 is

$$\hat{\beta}_2^{\text{HPS}}(\hat{a}) = \hat{\Sigma}(\hat{a})^{-1}\left[\hat{a}(X^{(1)})^\top Y^{(1)} + (X^{(2)})^\top Y^{(2)}\right]. \tag{2.12}$$

In this paper, we study the out-of-sample predication loss (test error) of $\hat{\beta}_2^{\text{HPS}}(\hat{a})$. Consider a test data point (x,y) generated from the same model as task 2: $y = x^\top\beta^{(2)} + \varepsilon$, where $x \in \mathbb{R}^p$ and $\varepsilon \in \mathbb{R}$ are independent of $X^{(1)}$, $X^{(2)}$, $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$, and only x is observable. We want to use $x^\top\hat{\beta}_2^{\text{HPS}}(\hat{a})$ to predict y, and we measure the predication loss using the mean squared error

$$\mathop{\mathbb{E}}_{x}\left[\left\|y - x^\top\hat{\beta}_2^{\text{HPS}}(\hat{a})\right\|^2\right] = \left\|(\Sigma^{(2)})^{1/2}\left(\hat{\beta}_2^{\text{HPS}}(\hat{a}) - \beta^{(2)}\right)\right\|^2 + \sigma^2.$$

Since σ^2 is a constant that does not depend on the model, we ignore it and define the predication loss as

$$L(\hat{\beta}_2^{\text{HPS}}(\hat{a})) := \left\|(\Sigma^{(2)})^{1/2}\left(\hat{\beta}_2^{\text{HPS}}(\hat{a}) - \beta^{(2)}\right)\right\|^2. \tag{2.13}$$

We will compare it to the out-of-sample predication loss $L(\hat{\beta}_2^{\text{OLS}})$ of the single-task OLS estimator

$$\hat{\beta}^{\text{OLS}} = [(X^{(2)})^\top X^{(2)}]^{-1}(X^{(2)})^\top Y^{(2)}$$

as a baseline.

Plugging (2.12) into (2.13), we get

$$L(\hat{\beta}_2^{\text{HPS}}(\hat{a})) = \Big\| (\Sigma^{(2)})^{1/2}\hat{\Sigma}(\hat{a})^{-1}(X^{(1)})^\top X^{(1)}(\hat{a}\beta^{(1)} - \hat{a}^2\beta^{(2)}) \\ + (\Sigma^{(2)})^{1/2}\hat{\Sigma}(\hat{a})^{-1}\left[(X^{(2)})^\top \varepsilon^{(2)} + \hat{a}(X^{(1)})^\top \varepsilon^{(1)}\right]\Big\|^2. \tag{2.14}$$

Note that if we replace $\hat{a}$ with a fixed number a in (2.14), then taking the expectation over $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$, we can get a clean bias-variance decomposition of the predication loss:

$$\mathop{\mathbb{E}}_{\varepsilon^{(1)},\varepsilon^{(2)}}\left[L(\hat{\beta}_2^{\text{HPS}}(a))\right] = L_{\text{bias}}(a) + L_{\text{Var}}(a), \tag{2.15}$$

where

$$L_{\text{bias}}(a) := \left\|(\Sigma^{(2)})^{1/2}\hat{\Sigma}(a)^{-1}(X^{(1)})^\top X^{(1)}\left(a\beta^{(1)} - a^2\beta^{(2)}\right)\right\|^2 \tag{2.16}$$

is called the bias term, which depends on the model bias between task 1 and task 2, and

$$L_{\text{Var}}(a) := \sigma^2 \operatorname{Tr}\left[\Sigma^{(2)}\hat{\Sigma}(a)^{-1}\right] \tag{2.17}$$

is called the variance term, which depends on the noise variance. Using concentration of the noise vectors $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$, we can show that $L(\hat{\beta}_2^{\text{HPS}}(a))$ is close to (2.15) up to a small error as in the next lemma. In this paper, we say that an event Ξ holds *with high probability* (w.h.p.) if $\mathbb{P}(\Xi) \to 1$ as $p \to \infty$.

LEMMA 2.2. *Under Assumption 2.1, for any small constant $c > 0$ and large constant $C > 0$, there exists a high probability event Ξ, on which the following estimates hold uniformly in $a \in \mathbb{R}$:*

$$L(\hat{\beta}_2^{\text{HPS}}(a)) = \left[1 + \mathrm{O}(p^{-1/2+c})\right] \cdot \left[L_{\textit{bias}}(a) + L_{\text{Var}}(a)\right] \\ + \mathrm{O}\left[p^{-C}\left(\|\beta^{(1)}\|^2 + \|\beta^{(2)}\|^2\right)\right], \tag{2.18}$$

and

$$L(\hat{\beta}_2^{\text{OLS}}) = \left[1 + \mathrm{O}(p^{-1/2+c})\right] \cdot \sigma^2 \operatorname{Tr}\left[\Sigma^{(2)}[(X^{(2)})^\top X^{(2)}]^{-1}\right]. \tag{2.19}$$

Since (2.18) holds uniformly for all $a \in \mathbb{R}$, we can also apply it to $\hat{\beta}_2^{\text{HPS}}(\hat{a})$, where $\hat{a}$ is a random variable that may depend on $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$. In practice, the parameter a is up to one's choice and may not be the global minimizer, so we have stated the result for an completely arbitrary a. (In fact, the optimization objective $g(a)$ is generally non-convex, so in some cases one can only find a local minimizer.) For a fixed $a \in \mathbb{R}$, the proof of (2.18) is based on the sharp concentration bounds in Lemma A.5 of the supplement [39]. To extend uniformly to all $a \in \mathbb{R}$, we will use a standard ε-net argument, which leads to a small error $p^{-C}\left(\|\beta^{(1)}\|^2 + \|\beta^{(2)}\|^2\right)$. Note that this error is negligible unless $\beta^{(1)}$ and $a\beta^{(2)}$ cancel each other almost exactly, and the noise variance σ is very small. The proof of Lemma 2.2 will be given in Appendix B of the supplement [39].

For the single-task predication loss (2.19), we can calculate its asymptotic limit exactly using the following classical result in multivariate statistics.

LEMMA 2.3 (Theorem 2.4 of [11] and Theorem 3.14 of [13]). *Under Assumption 2.1, we have that*

$$\operatorname{Tr}\left[\Sigma^{(2)}\frac{1}{(X^{(2)})^\top X^{(2)}}\right] = \operatorname{Tr}\left[\frac{1}{(Z^{(2)})^\top Z^{(2)}}\right] = \frac{p}{n_2 - p} \cdot \left[1 + \mathrm{O}\left(\frac{p^c}{\sqrt{np}}\right)\right]$$

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

If the entries of $Z^{(2)}$ are i.i.d. Gaussian, then this result follows from the classical result for the mean of inverse Wishart distribution [1]. For a general non-Gaussian $Z^{(2)}$, this result can be obtained using the well-known Stieltjes transform method (cf. Lemma 3.11 of [4]). Here we have presented the results from [11, 13], which give an almost sharp convergence rate. With Lemma 2.3, we obtain that w.h.p.,

$$L(\hat{\beta}_2^{\mathrm{OLS}}) = \left[1 + \mathrm{O}(p^{-1/2+c})\right] \cdot \frac{p\sigma^2}{n_2 - p}. \tag{2.20}$$

On the other hand, much less is known about the predication loss of the HPS estimator. In particular, its exact dependence on the model parameters, including the model bias, noise variance, sample sizes, and population covariance matrices, is not well-understood so far. [FY: What is known so far about this topic?] In the rest of this paper, we give a rigorous analysis of the bias term (2.16) and variance term (2.17) in three different settings of increasing complexity: (i) same covariates $X^{(1)} = X^{(2)}$; (ii) independent $X^{(1)}$ and $X^{(2)}$ with different sample sizes and same population covariance matrices $\Sigma^{(1)} = \Sigma^{(2)}$; (iii) independent $X^{(1)}$ and $X^{(2)}$ with different sample sizes and different population covariance matrices. In each case, we provide the exact asymptotic bias and variance limits, together with almost sharp convergence rates. In particular, we will use our results to analyze the effect of the following factors: *bias-variance tradeoff*, *sample sizes*, and *covariate shift*.

discuss about the bias-variance trade-off

3. Predication loss for HPS estimator.

3.1. *Varying sample sizes.* In this section, we consider the setting where $X^{(1)}$ and $X^{(2)}$ are independent and have the same population covariance matrices $\Sigma^{(1)} = \Sigma^{(2)}$. However, the two tasks can have different sample sizes $n_1 \neq n_2$. In this case, we can obtain the exact asymptotic limits of the bias term (2.16) and the variance term (2.17). We will use these results to illustrate the effect of the sample sizes on the predication loss of the HPS estimator.

Theorem 3.1. *Under Assumption 2.1 holds, suppose that $\Sigma^{(1)} = \Sigma^{(2)}$ and the entries of $Z^{(1)}$ and $Z^{(2)}$ are i.i.d. Gaussian random variables. Then for any small constant $c > 0$ and large constant $C > 0$, there exists a high probability event Ξ, on which the following estimates hold uniformly in all $a \in \mathbb{R}$:*

$$L_{\mathrm{Var}}(a) = \left[\mathcal{L}_1(a) + \mathrm{O}\left(\frac{p^c}{n_1}\right)\right] \cdot \sigma^2, \tag{3.1}$$

and

$$\begin{aligned} L_{bias}(a) = {} & \left[\mathcal{L}_2(a) + \mathrm{O}\left(p^{-1/2+c}\right)\right] \|\beta^{(1)} - a\beta^{(2)}\|^2 \\ & + \mathrm{O}\left[p^{-C}\left(\|\beta^{(1)}\|^2 + \|\beta^{(2)}\|^2\right)\right]. \end{aligned} \tag{3.2}$$

Here we define the functions

$$\mathcal{L}_1(a) := \frac{2p}{n_2(1-\xi_2) + a^2 n_1(1-\xi_1) + \sqrt{[n_2(1-\xi_2) + a^2 n_1(1-\xi_1)]^2 + 4a^2 n_2 n_1(\xi_1 + \xi_2 - \xi_1\xi_2)}},$$

$$\mathcal{L}_2(a) := \frac{1}{a^2} \cdot \frac{1 - 2\frac{\mathcal{L}_1(a)}{\xi_2[1+\mathcal{L}_1(a)]} + \kappa(a)}{1 - \xi_2\kappa(a)},$$

where we abbreviate $\xi_1 := p/n_1$, $\xi_2 := p/n_2$ and

$$\kappa(a) := \frac{\mathcal{L}_1(a)^2}{\xi_2^2[1+\mathcal{L}_1(a)]^2}\left[1 - \frac{a^4\mathcal{L}_1(a)^2}{\xi_1[1+a^2\mathcal{L}_1(a)]^2}\right]^{-1}.$$

The proof of Theorem 3.1 will be given in Appendix D in the supplement [39]. For the variance estimate in (3.1), it is not necessary to assume the Gaussian distributions of the entries of $Z^{(1)}$ and $Z^{(2)}$. In fact, (3.1) is a special case of Theorem 3.4 below for the more general case with possibly different $\Sigma^{(1)}$ and $\Sigma^{(2)}$. On the other hand, the Gaussian assumption is needed in our current proof of the bias limit (3.2). In the setting of Theorem 3.1, we can write

$$L_{\text{bias}}(a) = \mathbf{v}_a^\top (Z^{(1)})^\top Z^{(1)} \frac{1}{\left[a^2 (Z^{(1)})^\top Z^{(1)} + (Z^{(2)})^\top Z^{(2)}\right]^2} (Z^{(1)})^\top Z^{(1)} \mathbf{v}_a,$$

where we denote $\mathbf{v}(a) := (\Sigma^{(1)})^{1/2}\left(a\beta^{(1)} - a^2\beta^{(2)}\right)$. Using the rotational invariance of $(Z^{(1)})^\top Z^{(1)}$ and $(Z^{(2)})^\top Z^{(2)}$, we have that

$$L_{\text{bias}}(a) \approx \| \mathbf{v}_a \|^2 \frac{1}{p} \operatorname{Tr}\left[[(Z^{(1)})^\top Z^{(1)}]^2 \frac{1}{\left[a^2 (Z^{(1)})^\top Z^{(1)} + (Z^{(2)})^\top Z^{(2)}\right]^2} \right], \tag{3.3}$$

up to a small error. Notice that we can write (3.3) into a simpler form

$$L_{\text{bias}}(a) \approx \| \mathbf{v}_a \|^2 \left.\frac{\mathrm{d}}{\mathrm{d}x}\right|_{x=0} \frac{1}{p} \operatorname{Tr}\left[\frac{1}{a^2 (Z^{(1)})^\top Z^{(1)} + x[(Z^{(1)})^\top Z^{(1)}]^2 + (Z^{(2)})^\top Z^{(2)}} \right].$$

It is well-known that the empirical eigenvalue distributions (ESD) of $(Z^{(1)})^\top Z^{(1)}$ and $(Z^{(2)})^\top Z^{(2)}$ satisfy the famous Marchenko-Pastur (MP) law asymptotically [25]. From the MP law of $(Z^{(1)})^\top Z^{(1)}$, we can also derive the asymptotic ESD of $a^2(Z^{(1)})^\top Z^{(1)} + x[(Z^{(1)})^\top Z^{(1)}]^2$ for any fixed $a \in \mathbb{R}$ and $x > 0$. Due to the Gaussian assumption, $a^2(Z^{(1)})^\top Z^{(1)} + x[(Z^{(1)})^\top Z^{(1)}]^2$ and $(Z^{(2)})^\top Z^{(2)}$ are asymptotically freely independent from each other. Hence the asymptotic ESD of $a^2(Z^{(1)})^\top Z^{(1)} + x[(Z^{(1)})^\top Z^{(1)}]^2 + (Z^{(2)})^\top Z^{(2)}$ is given by the free additive convolution (or free addition) of the asymptotic ESD of $a^2(Z^{(1)})^\top Z^{(1)} + x[(Z^{(1)})^\top Z^{(1)}]^2$ and the MP law of $(Z^{(2)})^\top Z^{(2)}$. In particular, a sharp convergence estimate has been proved in [5, 6] for the free addition of two probability measure. We use that result to obtain a convergence estimate on

$$\frac{1}{p} \operatorname{Tr}\left[\frac{1}{a^2 (Z^{(1)})^\top Z^{(1)} + x[(Z^{(1)})^\top Z^{(1)}]^2 + (Z^{(2)})^\top Z^{(2)}} \right].$$

Taking derivative with respect to x at $x = 0$ gives the exact asymptotic limit of $L_{\text{bias}}(a)$.

We believe that the above argument can be extended to the case without Gaussian assumption. For example, instead of using the results in [5, 6] on free addition, we can use the sharp local laws on polynomials of random matrices in [3]. However, to apply the result in [3], we need to check some hard technical conditions for our setting. Hence we do not pursue this direction in this paper.

Now we use the following example to illustrate the effect of varying sample sizes on the predication loss of the HPS estimator.

EXAMPLE 3.2 (Sample sizes). We again consider the random effect model in Example 4.3 with $t = 2$. First, we show that if $\|\beta_0\|^2 \gg d^2$, then the global minimizer $\hat{a}$ of the function $g(a)$ in (2.11) is close to 1. The proof of Proposition 3.3 will be given in Appendix E.

PROPOSITION 3.3. *Suppose Assumption 2.1 and the setting in Example 4.3 hold. Suppose that*

$$\|\beta_0\|^2 \geqslant p^{c_0} d^2 + p^{-1/2+c_0}\sigma^2. \tag{3.4}$$

for a constant $c_0 > 0$. Then we have that for any small constant $c > 0$ and large constant $C > 0$,

$$\hat{a} = 1 + \mathrm{O}\left(\frac{d^2}{\|\beta_0\|^2} + p^{-1/4+c} \frac{d+\sigma}{\|\beta_0\|} + p^{-C} \right) \quad \textit{w.h.p.} \tag{3.5}$$

Combining Lemma 2.2, Theorem 3.1 and Proposition 3.3, and using that $\|\beta^{(1)}-\beta^{(2)}\|^2=(1+\mathrm{o}(1))d^2$ with high probability, the predication loss $L(\hat\beta_2^{\mathrm{HPS}}(\hat a))$ is approximately equal to

$$\ell(n_1,n_2)=\sigma^2\mathcal{L}_1(1)+2d^2\mathcal{L}_2(1)=\frac{p\sigma^2}{n_1+n_2-p}+2d^2\frac{n_1^2(n_1+n_2-p)+pn_1n_2}{(n_1+n_2)^2(n_1+n_2-p)},$$

with high probability, where in the second step we obtain $\mathcal{L}_1(1)$ and $\mathcal{L}_2(1)$ through direct calculations. Now we compare $\ell(n_1,n_2)$ with (2.20). First, we notice that the variance term is smaller than $L(\hat\beta_2^{\mathrm{OLS}})$, while the bias term is always positive. Moreover, calculating the derivative with respect to n_1, it is not hard to see that the variance term always decreases as n_1 increases, while the bias term always increases with n_1.

For a fixed n_2, we study the intricate bias-variance tradeoff with respect to n_1 using the following function of $\rho:=\rho_1+\rho_2$,

$$h(\rho):=\frac{p}{n_2-p}\cdot\frac{\sigma^2}{2d^2}-\frac{\ell(\rho_1p,\rho_2p)}{2d^2}=\frac{\rho-\rho_2}{(\rho-1)(\rho_2-1)}\cdot\frac{\sigma^2}{2d^2}-\frac{(\rho-\rho_2)^2(\rho-1)+(\rho-\rho_2)\rho_2}{\rho^2(\rho-1)}.$$

This function characterizes $L(\hat\beta_2^{\mathrm{OLS}})-L(\hat\beta_2^{\mathrm{HPS}}(\hat a))$, which gives the quantitive information transfer from the source task to the target task. First, we have positive (resp. negative) transfer if and only if $h(\rho)>0$ (resp. $h(\rho)<0$). Moreover, the sign of $h(\rho)$ is determined by the sign of the second order polynomial

$$\left(\frac{1}{\rho_2-1}\cdot\frac{\sigma^2}{2d^2}-1\right)\rho^2+(\rho_2+1)\rho-2\rho_2. \tag{3.6}$$

Then we observe the following dichotomy. [FY: add several plots to illustrate this dichotomy]

(i) If $\frac{1}{\rho_2-1}\cdot\frac{\sigma^2}{2d^2}-1>0$, then the polynomial (3.6) is positive for all $\rho\in(\rho_2+1,\infty)$, so that the transfer is always positive.

(ii) If $\frac{1}{\rho_2-1}\cdot\frac{\sigma^2}{2d^2}-1<0$, then we have the following cases.

- If $(\rho_2+1)^2<8\rho_2\left(1-\frac{1}{\rho_2-1}\cdot\frac{\sigma^2}{2d^2}\right)$, the polynomial (3.6) is negative for all ρ, so that the transfer is always negative.
- If $(1+\rho_2)^2<8\rho_2\left(1-\frac{1}{\rho_2-1}\cdot\frac{\sigma^2}{2d^2}\right)$, the polynomial (3.6) has two positive roots, where one of them is always smaller than ρ_2+1. Hence if the larger root, say ρ_c, is smaller than ρ_2+1, then the transfer is always negative for $\rho\in(\rho_2+1,\infty)$; otherwise, there is a transition from positive transfer to negative transfer as ρ crosses ρ_c.

Second, we calculate the derivative of $h(\rho)$, and find that its sign is determined by the sign of the third order polynomial

$$\frac{1}{\rho_2}\cdot\frac{\sigma^2}{2d^2}\rho^3-2(\rho-\rho_2)(\rho-1)(\rho-2)-(\rho_2-1)\rho. \tag{3.7}$$

Then we observe the following dichotomy. [FY: add several plots to illustrate this dichotomy]

(i) If $\frac{1}{\rho_2}\cdot\frac{\sigma^2}{2d^2}>2$, the polynomial (3.7) is always positive for all $\rho\in(\rho_2+1,\infty)$, so that the information transfer always increases as n_1 increases.

(ii) If $\frac{1}{\rho_2}\cdot\frac{\sigma^2}{2d^2}<2$, the polynomial (3.7) is positive initially around $\rho=\rho_2+1$, and then becomes negative when ρ crosses its unique root, say ρ_c, in (ρ_2+1,∞). Hence the information transfer achieves the global maximum at $\rho=\rho_c$.

[FY: add applications, simulations, and algorithm consequences of the results in this section]

3.2. *Covariate shift.* Finally, in this section, we consider the most general setting, where the feature covariates $X^{(1)}$ and $X^{(2)}$ of the two tasks have both different sample sizes and different population covariance matrices. In particular, the fact that population covariance matrices differ across tasks is often called "covariate shift", which is characterized by the matrix $(\Sigma^{(1)})^{1/2}(\Sigma^{(2)})^{-1/2}$. In this section, we describe the exact asymptotic variance limit in the high-dimensional setting, while the bias limit is much more complicated, and we can only give a rough estimate on it.

Compared to the results in Section 3.1, the spectrum of $\hat{\Sigma}(a)^{-1}$ now not only depends on the sample sizes of both tasks, but also depends on the "misalignment" between $\Sigma^{(1)}$ and $\Sigma^{(2)}$. To capture this misalignment quantitatively, we introduce the covariate shift matrix (rescaled by a)

$$M(a) := a(\Sigma^{(1)})^{1/2}(\Sigma^{(2)})^{-1/2}.$$

Let $\lambda_1(a), \lambda_2(a), \dots, \lambda_p(a)$ denote the singular values of M in descending order. By (2.5), these singular values satisfy that

$$|a|\tau \leqslant \lambda_p(a) \leqslant \cdots \leqslant \lambda_2(a) \leqslant \lambda_1(a) \leqslant |a|\tau^{-1}. \tag{3.8}$$

The main result of this section is the following theorem on the variance limit, which characterizes the exact dependence of $L_{\mathrm{Var}}(a)$ on the singular values of M.

THEOREM 3.4. *Under Assumption 2.1, for any small constant $c > 0$, there exists a high probability event Ξ, on which the following estimate holds for the variance term $L_{\mathrm{Var}}(a)$ in* (2.17)*:*

$$\left| L_{\mathrm{Var}}(a) - \frac{\sigma^2}{n_1+n_2} \operatorname{Tr}\left[\frac{1}{a_1 M(a)^\top M(a) + a_2}\right]\right| \leqslant \frac{(n_1+n_2)^{2/\varphi+c}}{p^{1/2}(n_1+n_2)^{1/2}} \cdot \frac{p\sigma^2}{n_1+n_2}, \tag{3.9}$$

uniformly in all $a \in \mathbb{R}$. Here (a_1, a_2) is the solution of the following self-consistent equations

$$a_1 + a_2 = 1 - \frac{p}{n_1+n_2}, \quad a_1 + \frac{1}{n_1+n_2}\left(\sum_{i=1}^p \frac{\lambda_i^2 a_1}{\lambda_i^2 a_1 + a_2}\right) = \frac{n_1}{n_1+n_2}. \tag{3.10}$$

With (3.8), it is easy to see that

$$\frac{\sigma^2}{n_1+n_2} \operatorname{Tr}\left[\frac{1}{a_1 M(a)^\top M(a) + a_2}\right] \sim \frac{p\sigma^2}{n_1+n_2}.$$

Hence the right-hand side of (3.9) is much smaller than this main term by a factor of $p^{-1/2}(n_1 + n_2)^{-1/2+2/\varphi+c}$. Lemma 2.3 can be also regarded as a special case of Theorem 3.4. To see this, when $M(a) = 0$, we solve equation (3.10) to obtain that

$$a_1 = \frac{n_1}{n_1+n_2}, \quad a_2 = \frac{n_2-p}{n_1+n_2},$$

and plug them into equation (3.9). The proof of Theorem 3.4 will be given in Appendix F in the supplement [39].

EXAMPLE 3.5 (Covariate shift). To illustrate the effect of covariate shift, we consider a similar setting as in Example 3.2, such that (3.4) holds. In Proposition 3.3, we have seen that the global minimizer $\hat{a}$ is close to 1 up to a small error. Hence we take $a = 1$ in (3.9), and study the asymptotic limit

$$\ell_{\mathrm{Var}}(M) := \frac{\sigma^2}{n_1+n_2} \operatorname{Tr}\left[\frac{1}{a_1(M)\cdot M^\top M + a_2(M)}\right],$$

where $M = (\Sigma^{(1)})^{1/2}(\Sigma^{(2)})^{-1/2}$, and $a_1(M)$ and $a_2(M)$ are defined through (3.10). We compare different choices of M that are scaled to have the same determinant. More precisely, for a fixed $\mu > 0$ we define

$$\mathcal{S}_\mu := \left\{ M \,\middle|\, \prod_{i=1}^{p} \lambda_i = \mu^p, \tau \leqslant \lambda_p \leqslant \lambda_1 \leqslant \tau^{-1} \right\}.$$

Our first observation is that if $n_1 \gg n_2$, i.e. there is enough source data compared to the target data, then $\ell_{\mathrm{Var}}(M)$ is minimized approximately when M is a scalar matrix. From equation (3.10), we obtain the following estimates on $a_1(M)$ and $a_2(M)$ for any $M \in \mathcal{S}_\mu$:

$$a_1(M) = \frac{n_1 + \mathrm{O}(p)}{n_1 + n_2}, \quad a_2(M) = \frac{n_2 + \mathrm{O}(p)}{n_1 + n_2}. \tag{3.11}$$

Inserting (3.11) into $\ell_{\mathrm{Var}}(M)$, we obtain that

$$\ell_{\mathrm{Var}}(M) = \left[1 + \mathrm{O}\left(\frac{n_2}{n_1}\right)\right] \cdot \frac{\sigma^2}{n_1} \mathrm{Tr}\left(\frac{1}{M^\top M}\right).$$

By AM-GM inequality, we observe that

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

is minimized when $\lambda_1 = \cdots \lambda_p = \mu$ under the restriction $\prod_{i=1}^p \lambda_i = \mu^p$. Hence in this case, $M = \mu\, \mathrm{Id}_{p\times p}$ is approximately the optimal choice.

However, we now use an example to show that the above observation fails when n_1 is comparable to n_2. We take $\mu = 1$, and consider the subset of covariate shift matrices satisfying $M = M^{-1}$. In other words, half of the singular values of M satisfy that $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{p/2} \geqslant 1$, while the other half eigenvalues are $\lambda_{p/2}^{-1} \geqslant \cdots \geqslant \lambda_2^{-1} \geqslant \lambda_1^{-1}$. When $\lambda_1/\lambda_{p/2} = 1$, we have $M = \mathrm{Id}_{p\times p}$ and there is no covariate shift. We claim the following dichotomy.

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

We explain why the above dichotomy happens. We can write that variance limit as

$$\ell_{\mathrm{Var}}(M) = \frac{\sigma^2}{n_1 + n_2} \sum_{i=1}^{p/2} \left(\frac{1}{\lambda_i^2 a_1 + a_2} + \frac{1}{\lambda_i^{-2} a_1 + a_2} \right).$$

When $M = \mathrm{Id}_{p\times p}$, by the first equation of (3.10), we have

$$\ell_{\mathrm{Var}}(\mathrm{Id}_{p\times p}) = \frac{\sigma^2}{n_1 + n_2} \sum_{i=1}^{p/2} \frac{2}{1-\gamma},$$

where we abbreviate $\gamma := p/(n_1 + n_2)$. Then using $a_1 + a_2 = 1 - \gamma$, through a direct calculation we find that

$$\ell_{\mathrm{Var}}(M) - \ell_{\mathrm{Var}}(\mathrm{Id}_{p\times p}) = \frac{\sigma^2}{n_1 + n_2 - p} \sum_{i=1}^{p/2} \frac{(\lambda_i^2 - 1)^2 a_1(M)\,[a_1(M) - a_2(M)]}{[a_1(M) + \lambda_i^2 a_2(M)][\lambda_i^2 a_1(M) + a_2(M)]}.$$

We claim that $a_1(M) > a_2(M)$ if and only if $n_1 > n_2$, which then explains the dichonomy. In fact, if $a_1 > a_2$, then the first equation in (3.10) gives that $a_1 > (1-\gamma)/2$, and the second equation in (3.10) gives that

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

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

For the bias limit, we have the following proposition.

PROPOSITION 3.6. *Under Assumption 2.1, for any small constant $c > 0$ and large constant $C > 0$, there exists a high probability event Ξ, on which the following estimate holds for the bais term $L_{bias}(a)$ in (2.16):*

$$\left| L_{bias}(a) - (\beta^{(1)} - a\beta^{(2)})^\top (\Sigma^{(1)})^{1/2} \Pi(a) (\Sigma^{(1)})^{1/2} (\beta^{(1)} - a\beta^{(2)}) \right|$$

$$\leqslant \left[\left(1 + \sqrt{\frac{p}{n_1}}\right)^4 - 1 + n_1^{-1/2+2/\varphi+c} \right] \frac{n_1^2 \lambda_1^2 \left\| (\Sigma^{(1)})^{1/2} \left(\beta^{(1)} - a\beta^{(2)}\right) \right\|^2}{[(\sqrt{n_1} - \sqrt{p})^2 \lambda_p^2 + (\sqrt{n_2} - \sqrt{p})^2]^2}$$

$$+ p^{-C} \left[\|\beta^{(1)}\|^2 + \|\beta^{(2)}\|^2 \right], \tag{3.12}$$

uniformly in all $a \in \mathbb{R}$. Here $\Pi(a)$ is a $p \times p$ matrix defined as

$$\Pi(a) := \frac{n_1^2}{(n_1+n_2)^2} M(a) \frac{a_3 M(a)^\top M(a) + (a_4+1)}{[a_1 M(a)^\top M(a) + a_2]^2} M(a)^\top,$$

λ_1 and λ_p are respectively the largest and smallest singular values of $M(a)$, and (a_3, a_4) is the solution of the following self-consistent equations

$$\begin{aligned} & a_3 + a_4 = \frac{1}{n_1+n_2} \sum_{i=1}^p \frac{1}{\lambda_i^2 a_1 + a_2}, \\ & a_3 + \frac{1}{n_1+n_2} \sum_{i=1}^p \frac{\lambda_i^2 (a_2 a_3 - a_1 a_4)}{(\lambda_i^2 a_1 + a_2)^2} = \frac{1}{n_1+n_2} \sum_{i=1}^p \frac{\lambda_i^2 a_1}{(\lambda_i^2 a_1 + a_2)^2}, \end{aligned} \tag{3.13}$$

where recall that (a_1, a_2) is the solution of (3.10).

Note that the first error term on the right-hand side of (3.12) is typically smaller than the main term $(\beta^{(1)} - a\beta^{(2)})^\top (\Sigma^{(1)})^{1/2} \Pi(a) (\Sigma^{(1)})^{1/2} (\beta^{(1)} - a\beta^{(2)})$ by a factor of $\mathrm{O}(\sqrt{p/n_1} + n_1^{-1/2+2/\varphi+c})$. Hence (3.12) only gives an exact asymptotic limit in the regime $n_1 \gg p$. Moreover, by equations (3.10) and (3.13) we have

$$a_1 = \frac{n_1}{n_1+n_2} + \mathrm{O}\left(\frac{p}{n_1+n_2}\right), \quad a_3 = \frac{n_3}{n_1+n_2} + \mathrm{O}\left(\frac{p}{n_1+n_2}\right),$$

and

$$a_3 = \mathrm{O}\left(\frac{p}{n_1+n_2}\right), \quad a_4 = \mathrm{O}\left(\frac{p}{n_1+n_2}\right).$$

Using these estimates, it is easy to check that

$$(\beta^{(1)} - a\beta^{(2)})^\top (\Sigma^{(1)})^{1/2} \Pi(a) (\Sigma^{(1)})^{1/2} (\beta^{(1)} - a\beta^{(2)})$$

$$= \left\| (\Sigma^{(2)})^{1/2} \frac{1}{a^2 \Sigma^{(1)} + \Sigma^{(2)}} a \Sigma^{(1)} \left(\beta^{(1)} - a\beta^{(2)}\right) \right\|^2 + \mathrm{O}\left(\frac{p \left\|\beta^{(1)} - a\beta^{(2)}\right\|^2}{n_1+n_2} \right). \tag{3.14}$$

Hence (3.12) is consistent with the result obtained by replacing $(X^{(1)})^\top X^{(1)}$ and $(X^{(2)})^\top X^{(2)}$ with $n_1 \Sigma^{(1)}$ and $n_2 \Sigma^{(2)}$, respectively, in $L_{\text{bias}}(a)$ by the law of large numbers in the regime $n_1 \gg p$. However, simulations show that our estimate (3.12) is more precise than the first term on the right-hand side of (3.14).

REMARK 3.7. The main error in Proposition 3.6 comes from approximating $(Z^{(1)})^\top Z^{(1)}$ by $n_1 \operatorname{Id}_{n_1 \times n_2}$ using Corollary A.7 in the supplement [39]. In order to improve this estimate and obtain an exact asymptotic result, one needs to study the singular value distribution of the random matrix $\mathcal{X} + a^2$ for any fixed $a \in \mathbb{R}$, where $\mathcal{X} := [(X^{(1)})^\top X^{(1)}]^{-1}(X^{(2)})^\top X^{(2)}$. We remark that the eigenvalues of $\mathcal{X}$ have been studied in the name of Fisher matrices [42]. However, since $\mathcal{X}$ is not symmetric, its singular values are different from its eigenvalues. To the best of our knowledge, the asymptotic singular value behavior of $\mathcal{X}$ is still an open problem in random matrix theory, and the study of the singular values of $\mathcal{X} + a^2$ will be even harder. We leave this problem to future study.

We also remark for the general case with covariate shift, the method in Section 3.1 for dealing with the bias term also fails, because we cannot reduce the problem into the study of the addition of two random matrices that are asymptotically freely independent.

[FY: add applications, simulations, and algorithm consequences of the results in this section]

4. Extension to multi-task learning. In this section, we consider the setting where the two tasks have the same covariates $X^{(1)} = X^{(2)}$. We define $A^\star \in \mathbb{R}^2$ as the normalized eigenvector corresponding to the larger eigenvalue of the 2×2 matrix ${B^\star}^\top \Sigma B^\star$, where $B^\star := [\beta^{(1)}, \beta^{(2)}] \in \mathbb{R}^{p\times 2}$ is the matrix formed by the linear model parameters of the two tasks. Without loss of generality, we assume that the two eigenvalues of ${B^\star}^\top \Sigma B^\star$ are not degenerate, so that $A^\star$ is well-defined. Otherwise, Theorem 4.1 will give a null result.

THEOREM 4.1. *Under Assumption 2.1, suppose that $X^{(1)} = X^{(2)}$ and $n_1 = n_2$. Let $c > 0$ be an arbitrary small constant. Then we have that with high probability,*

$$\left\| u_{\hat a} u_{\hat a}^\top - A^\star {A^\star}^\top \right\|_F \leqslant \left[\frac{n^{-1/2+2/\varphi+c} \| {B^\star}^\top \Sigma B^\star \| + n^{-1/2+c}\sigma^2}{\lambda_1 - \lambda_2} \right]^{1/2}, \tag{4.1}$$

where $u_{\hat a}$ is a unit vector defined as $u_{\hat a} := \frac{1}{\hat a^2+1}\begin{pmatrix}\hat a\\ 1\end{pmatrix}$, and λ_1 and λ_2 are respectively the larger and smaller eigenvalues of ${B^\star}^\top \Sigma B^\star$. Moreover, the prediction loss of the HPS estimator satisfies that with high probability,

$$\begin{aligned} &\left| L(\hat\beta_2^{\mathrm{HPS}}(\hat a)) - \left\| (\Sigma^{(2)})^{1/2} \left(A^\star(2)(B^\star A^\star) - \beta^{(2)} \right) \right\|^2 - |A^\star(2)|^2 \frac{p\sigma^2}{n-p} \right| \\ &\leqslant \left[\frac{n^{-1/2+2/\varphi+c} \| {B^\star}^\top \Sigma B^\star \| + n^{-1/2+c}\sigma^2}{\lambda_1 - \lambda_2} \right]^{1/2} \left(\|\Sigma^{1/2} B^\star\|^2 + \sigma^2 \right), \end{aligned} \tag{4.2}$$

where $A^\star(2)$ denotes the second entry of $A^\star$.

Recall that φ is a constant larger than 4, hence $n^{-1/2+2/\varphi+c}$ is a negligible factor asymptotically as long as c is smaller than $1/2 - 2/\varphi$. Moreover, the factor $(\lambda_1 - \lambda_2)^{-1}$ is natural, because the two eigenspaces of ${B^\star}^\top \Sigma B^\star$ are not stable when λ_1 and λ_2 are close to each other. The estimate (4.1) shows that the minimizer $\hat a$ is approximately equal to $A^\star(1)/A^\star(2)$, while (4.2) gives the exact asymptotic limit of $L(\hat\beta_2^{\mathrm{HPS}}(\hat a))$, together with an explicit convergence rate that we believe to be sharp.

It is not hard to extend the above result to the cases with more than two tasks. We make this extension for the following reasons. First, it provides a clearer geometric intuition than the two-task setting as we will discuss below. Second, the corresponding multi-task setting is prevalent in applications of multi-task learning to image classification, where there are multiple prediction labels/tasks for every image [33, 17]. Finally, it provides useful insights into a more general theory of multi-task learning, which we will explore in greater details in future works.

Multi-task setting. Suppose we have t datasets whose sample sizes are all equal to n and whose feature covariates are all equal to $X \in \mathbb{R}^{n\times p}$. The label vector of the i-th task follows a linear model

$$Y^{(i)} = X\beta^{(i)} + \varepsilon^{(i)}, \quad i = 1, 2, \cdots, t.$$

We assume that $X = Z\Sigma^{1/2}$ is a random matrix satisfying the same assumption as $X^{(2)}$ in Assumption 2.1, that is, Z is an $n \times p$ random matrix with i.i.d. entries satisfying (2.2) and (2.3), Σ is a deterministic positive definite symmetric matrix satisfying (2.4) and (2.5), and $\rho := n/p$ satisfies (2.7). $\varepsilon^{(i)} \in \mathbb{R}^n$, $i = 1, 2, \cdots, t$, are independent random vectors with i.i.d entries of mean zero, variance σ^2, and bounded moments as in (2.6). Finally, X, $\varepsilon^{(i)}$ and $\beta^{(i)}$, $i = 1, 2, \cdots, t$, are all independent from each other.

To define the HPS parameter for the above setting, we study the following minimization problem

$$f(A, B) = \sum_{j=1}^{t} \left\| XBA_j - Y^{(j)} \right\|^2, \tag{4.3}$$

where $B \in \mathbb{R}^{p\times r}$ is a rank-r shared feature representation layer, and $A := [A_1, A_2, \ldots, A_t] \in \mathbb{R}^{r\times t}$ with $A_i \in \mathbb{R}^r$ being a separate output layer for task i. We will focus on cases with $r < t$, because otherwise the global minimum of $f(A, B)$ reduces to single-task learning (cf. Proposition 1 of [38]).

For the optimization objective in (4.3), using the local optimality condition over B, that is, $\frac{\partial f}{\partial B} = 0$, we obtain $\hat{B}$ as a function of A:

$$\hat{B}(A) = (X^\top X)^{-1} X^\top Y A^\top (AA^\top)^+, \tag{4.4}$$

where $Y := [Y^{(1)}, Y^{(2)}, \ldots, Y^{(t)}]$ and $(AA^\top)^+$ denotes the pseudoinverse of $AA^\top$. Plugging $\hat{B}(A)$ into equation (4.3), we obtain the following objective that only depends on A (in matrix notation):

$$g(A) = \left\| X(X^\top X)^{-1} X^\top Y A^\top (AA^\top)^+ A - Y \right\|_F^2. \tag{4.5}$$

Let $\hat{A}$ be the global minimizer of $g(A)$. Then $(\hat{A}, \hat{B}(\hat{A}))$ is the global minimizer of $f(A, B)$. We define the HPS estimator for task i as $\hat{\beta}_i^{\text{HPS}} := \hat{B}(\hat{A})\hat{A}_i$, and its (out-of-sample) prediction loss as

$$L_i(\hat{\beta}_i^{\text{HPS}}) = \left\| \Sigma^{1/2} \left(\hat{\beta}_i^{\text{HPS}} - \beta^{(i)} \right) \right\|^2. \tag{4.6}$$

Our main result of this section, Theorem 4.2, shows that hard parameter sharing essentially approximates all tasks through a rank-r subspace. To formalize this geometric intuition, as in the two-task case, we introduce the matrix $B^\star := [\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(t)}] \in \mathbb{R}^{p\times t}$ formed by the linear model parameters of all the t tasks. Let $A^\star A^{\star\top}$ denote the best rank-r subspace approximation of $B^{\star\top}\Sigma B^\star$ (which is the covariance of the task labels):

$$A^\star := \mathop{\arg\min}_{U\in\mathbb{R}^{t\times r}: U^\top U = \mathrm{Id}_{r\times r}} \langle UU^\top, B^{\star\top}\Sigma B^\star \rangle, \tag{4.7}$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product between two matrices. To ensure that $A^\star$ is unique, we assume that the r-th largest eigenvalue of $B^{\star\top}\Sigma B^\star$ is strictly larger than the $(r+1)$-th largest eigenvalue. Otherwise, Theorem 4.2 will give a null result. Let $a_i^\star := A^\star A^{\star\top} e_i \in \mathbb{R}^r$ denote the i-th column of $A^\star A^{\star\top}$, where e_i is the standard basis unit vector along i-th direction. Then we have the following result.

THEOREM 4.2. *Suppose the above multi-task setting holds. Let $c > 0$ be an arbitrary small constant. Then for any task $i = 1, \cdots, t$, we have that high probability,*

$$\left\| \hat{A}^\top (\hat{A}\hat{A}^\top)^+ \hat{A} - A^\star A^{\star\top} \right\|_F \leqslant \left[\frac{n^{-1/2+2/\varphi+c} \| B^{\star\top} \Sigma B^\star \| + n^{-1/2+c}\sigma^2}{\lambda_r - \lambda_{r+1}} \right]^{1/2}, \tag{4.8}$$

where λ_r and λ_{r+1} are respectively the r-th and $(r+1)$-th largest eigenvalues of $B^{\star\top}\Sigma B^\star$. Moreover, the prediction loss of the HPS estimator $\hat{\beta}_i^{\text{HPS}}$ satisfies that

$$\begin{aligned} & \left| L_i(\hat{\beta}_i^{\text{HPS}}) - L_i(B^\star a_i^\star) - \frac{p\sigma^2}{n-p} \|a_i^\star\|^2 \right| \\ & \leqslant \left[\frac{n^{-1/2+2/\varphi+c} \| B^{\star\top} \Sigma B^\star \| + n^{-1/2+c}\sigma^2}{\lambda_r - \lambda_{r+1}} \right]^{1/2} \left(\|B^{\star\top}\Sigma B^\star\| + \sigma^2 \right). \end{aligned} \tag{4.9}$$

Finally, we have a better bound for the averaged prediction loss:

$$\begin{aligned} & \left| \frac{1}{t} \sum_{i=1}^t L_i(\hat{\beta}_i^{\text{HPS}}) - \frac{1}{t} \left\| \Sigma^{1/2} B^\star (A^\star A^{\star\top} - \mathrm{Id}_{t\times t}) \right\|_F^2 - \frac{p\sigma^2}{n-p} \cdot \frac{r}{t} \right| \\ & \leqslant n^{-1/2+2/\varphi+c} \|B^{\star\top}\Sigma B^\star\| + n^{-1/2+c}\sigma^2, \end{aligned} \tag{4.10}$$

with high probability.

The bound (4.8) verifies our intuition that hard parameter sharing approximates the matrix $B^{\star\top}\Sigma B^\star$ through a best rank-r subspace. The estimate (4.9) shows that the prediction loss of $\hat{\beta}_i^{\text{HPS}}$ decomposes into a bias term $L_i(B^\star a_i^\star)$ that measures the prediction loss of $B^\star a_i^\star$, plus a variance term that scales with $\|a_i^\star\|^2$. Since $\|a_i^\star\|^2 \leqslant 1$, compared with the single-task predication loss (2.20), the variance term always decreases in the HPS for multi-task setting. On the other hand, the bias term always increases, because the bias in single-task linear regression is zero. Hence whether the HPS estimator is better than the OLS estimator depends on an intricate *bias-variance tradeoff*. We can observe a similar bias-variance tradeoff for the averaged predication loss in (4.10) using the fact that $r < t$. Note that the estimate (4.10) can be applied even when the best rank-r subspace approximation of $B^{\star\top}\Sigma B^\star$ is not unique. For all the estimates in Theorem 4.2, we believe that their convergence rates are asymptotically tight when n goes to infinity.

The proof of Theorem 4.2 will be given in Appendix C of the supplement [39]. Moreover, Theorem 4.2 implies Theorem 4.1 as a special case with $t = 2$ and $r = 1$.

To illustrate the bias-variance tradeoff quantitively, we consider a random-effect model, which has been studied for single-task linear regression and ridge regression (see e.g. [14, 15]).

EXAMPLE 4.3 (Random-effect model). Suppose every $\beta^{(i)}$ consists of two random components, one that is shared among all tasks and one that is task-specific. More precisely,

$$\beta^{(i)} = \beta_0 + \widetilde{\beta}^{(i)}, \quad i = 1, 2, \cdots, t,$$

where β_0 denotes the shared component, and $\widetilde{\beta}^{(i)}$ denotes the i-th task-specific component whose entries are i.i.d. Gaussian random variables of mean zero and variance $p^{-1}d^2$.

In the random-effect model described above, using the concentration of Gaussian random vectors (e.g. Lemma A.5 in the supplement [39]), the (i,j)-th entry of $B^{\star\top}\Sigma B^\star$ is equal to

$$\beta_i^\top \Sigma \beta_j = \beta_0^\top \Sigma \beta_0 + \delta_{ij} \frac{d^2}{p} \operatorname{Tr} \Sigma + \mathrm{O}\left(p^{-1/2+c} \|\beta_0\|^2 + p^{-1/2+c} d^2 \right), \tag{4.11}$$

with high probability for any constant $c > 0$. We omit the details to show the error bound using Lemma A.5. With (4.11), it is easy to calculate that with high probability, the eigenvalues of $B^{\star\top}\Sigma B^{\star}$ are given by

$$\lambda_1 = (1 + \mathrm{O}(p^{-1/2+c}))\left(t\beta_0^\top \Sigma \beta_0 + \frac{d^2}{p}\operatorname{Tr}\Sigma\right),$$

and

$$\lambda_i = (1 + \mathrm{O}(p^{-1/2+c}))\frac{d^2}{p}\operatorname{Tr}\Sigma, \quad i = 2, \cdots, t.$$

Thus for the best rank-r approximation $A^\star A^{\star\top}$ of $B^{\star\top}\Sigma B^\star$, we have

$$\left\|\Sigma^{1/2}B^\star(A^\star A^{\star\top} - \mathrm{Id}_{t\times t})\right\|_F^2 = [1 + \mathrm{O}(p^{-1/2+c})] \cdot (t-r)\frac{d^2}{p}\operatorname{Tr}\Sigma$$

with high probability. Then using (4.10), we obtain that

$$\frac{1}{t}\sum_{i=1}^t L_i(\hat\beta_i^{\mathrm{HPS}}) = \left(1 - \frac{r}{t}\right)\frac{d^2}{p}\operatorname{Tr}\Sigma + \frac{r}{t}\cdot\frac{p\sigma^2}{n-p} + \mathrm{o}(\|\beta_0\|^2 + d^2 + \sigma^2), \quad \text{w.h.p.}$$

If the error is sufficiently small, then comparing the above equation with (2.20), we have the following observations.

(i) **Positive vs. negative transfer.** The averaged HPS prediction loss is smaller than the single-task OLS prediction loss if and only if $\frac{d^2}{p}\operatorname{Tr}\Sigma < \frac{p\sigma^2}{n-p}$, that is, the "task-specific variance" is smaller than the "noise variance" up to some constant factor.

(ii) **The optimal rank r.** If $\frac{d^2}{p}\operatorname{Tr}\Sigma < \frac{\sigma^2 p}{n-p}$, then the smallest averaged HPS prediction loss is achieved when $r = 1$. Hence increasing the width r of the shared feature representation layer does not help.

(iii) **Sample efficiency.** Suppose $\frac{d^2}{p}\operatorname{Tr}\Sigma < \frac{\sigma^2 p}{n-p}$ and we choose the optimal rank $r = 1$. Following [FY: add citation], we define the data efficiency ratio of HPS as the proportion of labelled data needed to achieve comparable performance to single-task linear regression. More precisely, for some $x \in (0,1)$, if we only use xn many data, then the averaged HPS predication loss is

$$\frac{1}{t}\sum_{i=1}^t L_i(\hat\beta_i^{\mathrm{HPS}}, x) = \left(1 - \frac{1}{t}\right)\frac{d^2}{p}\operatorname{Tr}\Sigma + \frac{1}{t}\cdot\frac{p\sigma^2}{xn-p} + \mathrm{o}(d^2 + \sigma^2), \quad \text{with high probability.}$$

Comparing it to the single-task OLS predication loss $t^{-1}\sum_{i=1}^t L_i(\hat\beta_i^{\mathrm{OLS}})$, we find that HPS requires at most (recall that $\rho = n/p$)

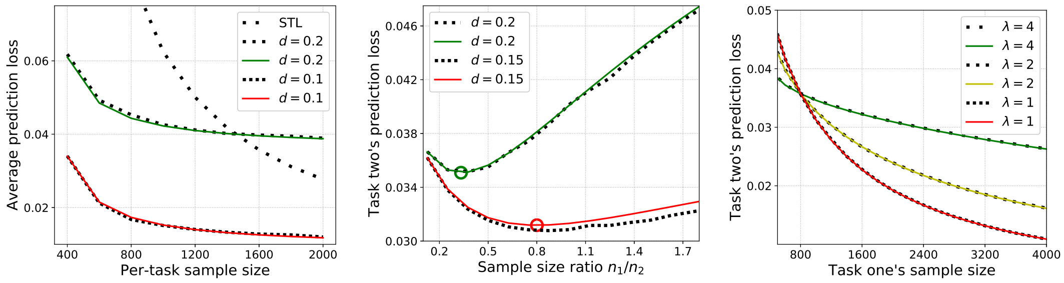
$$x = \frac{1}{\rho} + \frac{1 - \rho^{-1}}{t - (t-1)(\rho - 1)\frac{d^2}{\sigma^2}\cdot\frac{1}{p}\operatorname{Tr}\Sigma}$$

proportion of the samples to achieve the same performance. This is the data efficiency ratio for the above random-effect model.

[FY: add applications, simulations, and algorithm consequences of the results in this section]

5. Experiments.

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



(a) Example 4.3

(b) Example 3.2

(c) Example 3.5

Fig 2: Our estimated losses (solid line) match the empirical losses (dotted line) accurately under various settings in dimension $p = 200$. **Left.** Validating Example 4.3 for ten tasks: the noise variance σ^2 is $1/4$. **Middle.** Validating Example 3.2 for two tasks: we discover an interesting phenomena by fixing task two's sample size and increasing task one's sample size. Moreover, our result accurately predicts the critical point (marked in circle) of the loss curve. **Right.** We show how different levels of covariate shift affect hard parameter sharing when there is no bias. Having covariate shift increases task two's prediction loss when task two's sample size is smaller than task one. Otherwise, having covariate shift (surprisingly) decreases task two's prediction loss.

5.1.0.1. Sample efficiency.. First, we validate the result of Example 4.3. Figure 2a shows the average prediction loss over ten tasks as we increase the number of samples per-task from 400 to 2000. In all the parameter settings, our results estimate the empirical losses accurately. We also observe a trend that the average prediction loss increases as we increase distance d from 0.1 to 0.2. Our work explains the differences between these two settings since $d^2 = 0.1^2$ is always smaller than $\frac{\sigma^2 p}{n-p}$, but $d^2 = 0.2^2$ is not. Indeed, we observe a crossover point between hard parameter sharing and STL. Finally, for $d = 0.2$, looking horizontally, we find that HPS requires fewer samples per-task than STL to achieve the same loss level.

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

5.1.0.3. Covariate shift.. Finally, we validate the result of Example 3.5. Figure 2c shows task two's prediction loss as we increase task one's sample size. Recall that λ measures the severity of covariate shifts—a larger λ means a larger covariate shift. We indeed observe the dichotomy in Example 3.5 at $n_1 = 800$. The sample size n_2 is 800 and the noise variance σ^2 is $1/4$.

5.2. *Further Studies on Text Classification Tasks.* Our results and simulations are all in the high-dimensional linear regression setting. How well do they extend to other scenarios? In this section, we conduct further studies on six text classification datasets. Our datasets include a movie review sentiment dataset (MR) [31], a sentence subjectivity dataset (SUBJ) [30], a customer reviews dataset (CR) [19], a question type dataset (TREC) [24], an opinion polarity dataset (MPQA) [37], and the Stanford sentiment treebank (SST) dataset [35]. Our model consists of a word embedding layer with GloVe embeddings [32] followed by a long-short term memory (LSTM) or a multi-layer perception (MLP) layer [22].[1]

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

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

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

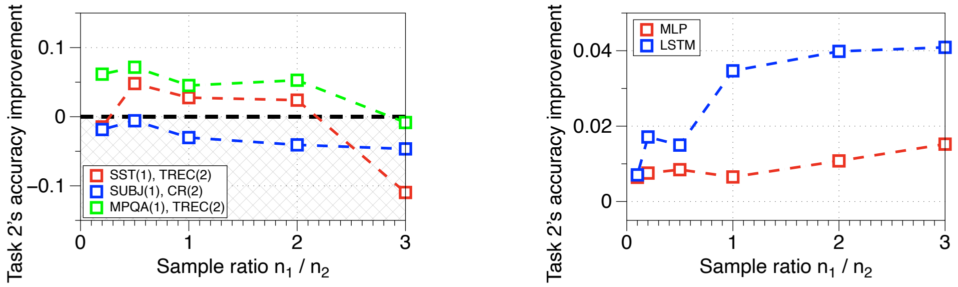
If we apply this procedure to the settings of Figure 3a and 2b, it will terminate once reaching the optimal sample ratio. The advantage of this procedure is that it reduces the computational cost compared to standard round-robin training schedules. For example, if the procedure terminates at 30% of all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

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

5.2.0.2. Covariate shift.. Recall from Example 3.5 that having covariate shifts worsens the variance (hence the loss) of hard parameter sharing when the sample ratio increases. This highlights the need for correcting covariate shifts when the sample size ratio rises. To this end, we study a covariance alignment procedure proposed in [38], 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 [38] for more details about the procedure and the implementation.

We conduct multi-task training on all 15 task pairs from the six datasets. In Figure 3b, we measure the performance gains from performing covariance alignment vs. HPS. To get a

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



(a) HPS vs. STL

(b) HPS vs. covariance alignment

Fig 3: Comparing hard parameter sharing (HPS) to single-task learning (STL) and a covariance alignment approach proposed by [38]: In Figure 3a, we observe that for multiple task pairs, increasing task one's sample size improves task two's prediction accuracy initially, but hurts eventually – a phenomenon similar to Figure 2b. In Figure 3b, we observe that as task one's sample size increases, covariance alignment improves more over HPS.

robust comparison, we average the improvements over the 15 task pairs. The result shows that as the sample size ratio increases, performing covariance alignment provides more significant gains over HPS. We fix task two's sample size at $1,000$, and increase task one's sample size from $1,000$ to $3,000$.

SUPPLEMENTARY MATERIAL

Supplement to "Hard Parameter Sharing in High-dimensional Linear Regression". In [39], we provide the proofs of the technical results in Sections 2-3.2, including Lemma 2.2, Theorem 4.2, Theorem 3.1, Proposition 3.3, Theorem 3.4 and Proposition 3.6.
().

REFERENCES

[1] ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley New York.

[2] ANDO, R. K. and ZHANG, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research* **6** 1817–1853.

[3] ERDŐS, L., KRÜGER, T. and NEMISH, Y. (2020). Local laws for polynomials of Wigner matrices. *Journal of Functional Analysis* **278** 108507.

[4] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral analysis of large dimensional random matrices*, 2nd ed. *Springer Series in Statistics*. Springer, New York.

[5] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2017). Local Law of Addition of Random Matrices on Optimal Scale. *Communications in Mathematical Physics* **349** 947–990.

[6] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2017). Convergence rate for spectral distribution of addition of random matrices. *Advances in Mathematics* **319** 251 - 291.

[7] BASTANI, H. (2020). Predicting with proxies: Transfer learning in high dimension. *Management Science*.

[8] BAXTER, J. (2000). A model of inductive bias learning. *Journal of artificial intelligence research* **12** 149–198.

[9] BEN-DAVID, S., BLITZER, J., CRAMMER, K., KULESZA, A., PEREIRA, F. and VAUGHAN, J. W. (2010). A theory of learning from different domains. *Machine learning* **79** 151–175.

[10] BEN-DAVID, S. and SCHULLER, R. (2003). Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines* 567–580. Springer.

[11] BLOEMENDAL, A., ERDŐS, L., KNOWLES, A., YAU, H. T. and YIN, J. (2014). Isotropic Local Laws for Sample Covariance and Generalized Wigner Matrices. *Electron. J. Probab.* **19** 1-53.

[12] CARUANA, R. (1997). Multitask learning. *Machine learning* **28** 41–75.

[13] DING, X. and YANG, F. (2018). A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *Ann. Appl. Probab.* **28** 1679–1738.

[14] DOBRIBAN, E. and SHENG, Y. (2020). WONDER: Weighted One-shot Distributed Ridge Regression in High Dimensions. *Journal of Machine Learning Research* **21** 1–52.

[15] DOBRIBAN, E., WAGER, S. et al. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics* **46** 247–279.

[16] ERDOS, L. and YAU, H.-T. (2017). A dynamical approach to random matrix theory. *Courant Lecture Notes in Mathematics* **28**.

[17] EYUBOGLU, S., ANGUS, G., PATEL, B. N., PAREEK, A., DAVIDZON, G., JAREDDUNNMON and LUNGREN, M. P. (2020). Multi-task weak supervision enables automatedabnormality localization in whole-body FDG-PET/CT.

[18] HSU, D. J., KAKADE, S. M., LANGFORD, J. and ZHANG, T. (2009). Multi-label prediction via compressed sensing. In *Advances in neural information processing systems* 772–780.

[19] HU, M. and LIU, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* 168–177. ACM.

[20] KNOWLES, A. and YIN, J. (2016). Anisotropic local laws for random matrices. *Probability Theory and Related Fields* 1–96.

[21] KUMAR, A. and DAUMÉ III, H. (2012). Learning task grouping and overlap in multi-task learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*.

[22] LEI, T., ZHANG, Y., WANG, S. I., DAI, H. and ARTZI, Y. (2018). Simple recurrent units for highly parallelizable recurrence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* 4470–4481.

[23] LI, S., CAI, T. T. and LI, H. (2020). Transfer Learning for High-dimensional Linear Regression: Prediction, Estimation, and Minimax Optimality. *arXiv preprint arXiv:2006.10593*.

[24] LI, X. and ROTH, D. (2002). Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1* 1–7. Association for Computational Linguistics.

[25] MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik* **1** 457.

[26] MAURER, A. (2006). Bounds for linear multi-task learning. *Journal of Machine Learning Research* **7** 117–139.

[27] MAURER, A., PONTIL, M. and ROMERA-PAREDES, B. (2016). The benefit of multitask representation learning. *The Journal of Machine Learning Research* **17** 2853–2884.

[28] NICA, A. and SPEICHER, R. (2006). *Lectures on the combinatorics of free probability* **13**. Cambridge University Press.

[29] PAN, S. J. and YANG, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22** 1345–1359.

[30] PANG, B. and LEE, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics* 271. Association for Computational Linguistics.

[31] PANG, B. and LEE, L. (2005). 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* 115–124. Association for Computational Linguistics.

[32] PENNINGTON, J., SOCHER, R. and MANNING, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* 1532–1543.

[33] RAJPURKAR, P., IRVIN, J., ZHU, K., YANG, B., MEHTA, H., DUAN, T., DING, D., BAGUL, A., LANGLOTZ, C., SHPANSKAYA, K. et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

[34] RUDER, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

[35] SOCHER, R., PERELYGIN, A., WU, J., CHUANG, J., MANNING, C. D., NG, A. and POTTS, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* 1631–1642.

[36] VANDENHENDE, S., GEORGOULIS, S., PROESMANS, M., DAI, D. and VAN GOOL, L. (2020). Revisiting Multi-Task Learning in the Deep Learning Era. *arXiv preprint arXiv:2004.13379*.

[37] WIEBE, J., WILSON, T. and CARDIE, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation* **39** 165–210.

[38] WU, S., ZHANG, H. R. and RÉ, C. (2020). Understanding and Improving Information Transfer in Multi-Task Learning. In *International Conference on Learning Representations*.

[39] YANG, F., ZHANG, H. R., WU, S., SU, W. J. and RÉ, C. (2020). Supplement to “Sharp bias-variance tradeoffs of hard parameter sharing in high-dimensional linear regression".

[40] ZAMIR, A. R., SAX, A., SHEN, W., GUIBAS, L. J., MALIK, J. and SAVARESE, S. (2018). Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3712–3722.

[41] ZHANG, Y. and YANG, Q. (2017). A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

[42] ZHENG, S., BAI, Z. and YAO, J. (2017). CLT for eigenvalue statistics of large-dimensional general Fisher matrices with applications. *Bernoulli* **23** 1130–1178.