Analyzing Information Transfer from Heterogeneous Sources via Precise High-dimensional Asymptotics

July 25, 2021

Abstract

We consider the problem of transfer learning—gaining knowledge from one task and applying it to a different but related target task. A fundamental question in transfer learning is when combining the data of both tasks works better than using only the target task's data (equivalently, when there is a "positive information transfer"). We study this question formally in a linear regression setting where a certain two-layer linear neural network estimator is used to combine both tasks' data. Under various settings involving covariate and model shifts, the estimator reproduces several interesting behaviors regarding positive (and negative) information transfer observed on real-world data.

We then show the precise asymptotic limit of risk of the estimator in terms of geometries of the data in a high-dimensional setting, where the sample sizes increase with the feature dimension proportionally at a fixed ratio. We also find that the asymptotics is accurate for finite dimensions and apply them to analyze the exact conditions for positive information transfer. This analysis leads to several novel insights regarding covariate and model shifts. For example, the risk curve is non-monotone under model shift, thus motivating a more efficient procedure for progressively adding data. The main ingredient of our analysis is showing the asymptotic limit of certain combinations of independent sample covariance matrices under covariate shift and different sample sizes, which may be of independent interest.

1 Introduction

Given data from two related tasks, how much does combining both data for learning a joint model help predict one of the tasks of interest? For example, suppose you have n_1 datapoints with p-dimensional covariates (sampled from a distribution D_1) and real-valued labels in the first task, and n_2 datapoints with p-dimensional covariates (sampled from a different distribution D_2) and real-valued labels in the second task. Does combining the $n_1 + n_2$ datapoints help learn a model whose performance is better than a model learned using the n_2 datapoints alone? Such questions are prevalent in transfer learning, where one often uses data from related tasks to expand the sample size of a task of interest. A frequent impediment to applying transfer learning is *negative information transfer* if combining both data performs worse than learning a single task clone (Pan and Yang, 2009). *Positive information transfer* refers to the desirable scenario where a task of interest benefits from another task's data.

The answer to the two questions above hinges on the tasks' heterogeneity. For example, negative information transfer can happen if the label distribution of the second task conditional on the covariates deviates significantly from that of the first task. Additionally, the "information transfer effect" also hinges on the covariates' distributions, D_1 and D_2. Following the existing literature (e.g., Kouw and Loog (2018)), we refer to settings where $D_1 \neq D_2$ as *covariate shift* and settings where label distributions differ (conditional on the covariates) as *model shift*. Both covariate and model shifts are prevalent in modern datasets (see e.g., Koh et al. (2021)).

This paper studies a linear regression setting involving two tasks with covariate and model shifts. We use a two-layer linear neural network to combine the data from both tasks and show that this network reproduces several interesting phenomena in transfer learning. We analyze the prediction risk of this network and its transfer effect, providing precise conditions based on geometries of D_1, D_2, and the sample sizes n_1, n_2. The analysis has also inspired several insights for understanding and mitigating covariate and model shifts.

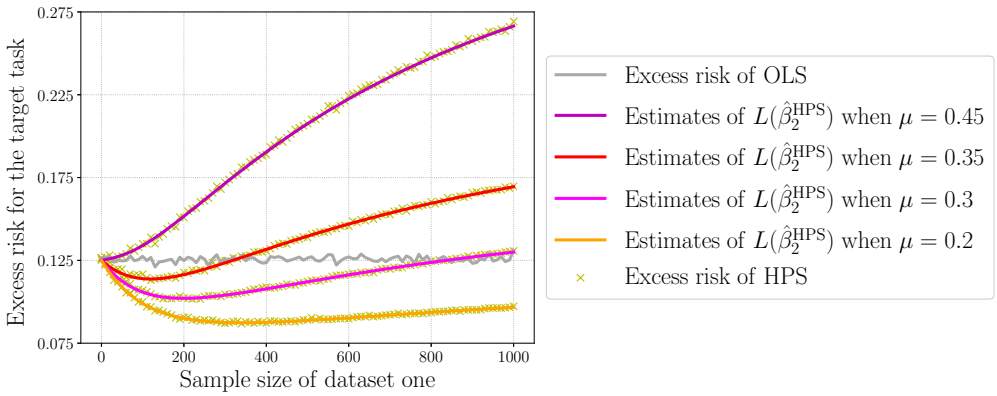


Figure 1: We illustrate that our setup can reproduce the phenomenon of *positive* (or *negative*) *information transfer*, which is prevalent in transfer learning. We vary the sample size of dataset one and the distance parameter μ such that $\|\beta^{(1)} - \beta^{(2)}\|^2 \approx 2\mu^2$. The region below the excess risk of the ordinary least squares (OLS) estimator corresponds to *positive information transfer*. This simulation uses $p = 100, n_2 = 300, \sigma = 1/2$. See also Figure 8b for a similar phenomenon observed in text classification tasks.

The key ingredient of our analysis is showing the asymptotic limit of certain combinations of sample covariance matrices under distribution shift, using the *local laws* developed recently by Bloemendal et al. (2014) and Bao et al. (2017a). After formally defining the specific problem that we tackle, in Section 1.3, we provide a technical discussion of the related works in random matrix theory.

1.1 Setup and motivating examples

Let $i = 1$ or 2. Let $x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)}$ be the covariates. Let $y_1^{(i)}, y_2^{(i)}, \ldots, y_{n_i}^{(i)}$ be the corresponding labels. We assume a linear model specified by an unknown model vector $\beta^{(i)} \in \mathbb{R}^p$ as follows:

$$y_j^{(i)} = \langle x_j^{(i)}, \beta^{(i)} \rangle + \varepsilon_j^{(i)}, \text{ for any } j = 1, \ldots, n_i, \tag{1.1}$$

where $\varepsilon_j^{(i)} \in \mathbb{R}$ denotes a random noise variable with mean zero and variance σ^2. We focus on the so called *underparametrized* setting for the target task where $n_2 > p$ whereas for dataset one n_1 can be either smaller or greater than p (see Section 2.1 for the precise assumptions). For the ease of presentation, we refer to the first task as the source and the second task as the target.

We learn a two-layer linear neural network with parameters $A \in \mathbb{R}^{p \times r}$, $B_1 \in \mathbb{R}^r$, and $B_2 \in \mathbb{R}^r$ by minimizing the following optimization objective:

$$f(A, B_1, B_2) = \left\| X^{(1)} A B_1 - Y^{(1)} \right\|_2^2 + \left\| X^{(2)} A B_2 - Y^{(2)} \right\|_2^2. \tag{1.2}$$

The above objective uses A as the shared feature space for both tasks and B_1, B_2 as the prediction head for each task.[1] Such models are also known as *hard parameter sharing* (Caruana, 1997; Ruder, 2017). We focus on the case of $r = 1$—otherwise, the global minimizer of $f(A, B_1, B_2)$ reduces to learning each task alone, resulting in zero information transfer (cf. Proposition 2.2 in Section 2.2). For the case of $r = 1$, let $(\hat{A}, \hat{B}_1, \hat{B}_2)$ be $f(\cdot)$'s global minimizer, which is unique because $n_2 > p$. The hard parameter sharing (HPS) estimator for the target task is defined as $\hat{\beta}_2^{\mathrm{HPS}} := \hat{B}\hat{A}_2$. In order to evaluate the estimator, we will study the excess risk, which is the prediction risk at an unseen datapoint (of task two) minus σ^2, denoted as $L(\hat{\beta}_2^{\mathrm{HPS}})$.

As a motivating example of the above setup, we show an intriguing simulation where combining both datasets results in different transfer effects. Figure 1 illustrates a setting where X_1, X_2 are sampled from an isotropic Gaussian and β_1, β_2 are generated so that $\|\beta^{(1)} - \beta^{(2)}\|^2 \approx 2\mu^2$. We observe that for $\mu = 0.25$, the

[1] Popular variants of the objective function (1.2) also involve weight parameters for each task and ridge regularization. We focus on the unweighted and unregularized objective for simplicity. All of the asymptotic limits described below can be straightforwardly extended to weighted and regularized objectives.

transfer effect is always positive—$L(\hat{\beta}_2^{\mathrm{HPS}})$ is always smaller than the excess risk of OLS. For $\mu = 0.3$ or 0.35, the transfer effect is positive only for a restricted range of n_1. On the other hand, for $\mu = 0.45$, the transfer effect is always negative. Our result below will imply the exact conditions for positive information transfer, depending on μ and n_1.

1.2 Summary of results

We consider a high-dimensional setting on each dataset where the sample size n_i increases with p to infinity proportionally by a constant factor, which has recently received lots of interests (see e.g., Hastie et al. (2019) and Section 1.3 for more references). Let $\Sigma^{(i)} \in \mathbb{R}^{p\times p}$ be a deterministic positive semidefinite matrix. The covariates are assumed to have population covariance equal to $\Sigma^{(i)}$: $x_j^{(i)} = \left(\Sigma^{(i)}\right)^{1/2} z_j^{(i)}$, for any $j = 1, \ldots, n_i$, where $z_j^{(i)}$ consists of independent and identical entries sampled from a distribution that has zero mean, unit variance, and satisfies certain bounded moment condition.

Main results. Thus, the OLS estimator exists and it is well-known that the limit of its excess risk when p approaches infinity is equal to $\frac{\sigma^2 p}{n_2 - p}$ (see e.g., Bai and Silverstein (2010)). By contrast, showing the asymptotic limit of $L(\hat{\beta}_2^{\mathrm{HPS}})$ in the high-dimensional setting requires dealing with the inverse of the sum of two sample covariance matrices under distribution shift. Our main result addresses both challenges by ≪Todo notes: illustrate≫.

- Our first result applies to settings where $\Sigma^{(1)}$ and $\Sigma^{(2)}$ can be arbitrarily different but $\beta^{(1)} = \beta^{(2)}$. In Theorem 3.1, we characterize the asymptotic limit of $L(\hat{\beta}_2^{\mathrm{HPS}})$ as a function of the singular values of the matrix $\left(\Sigma^{(1)}\right)^{1/2}\left(\Sigma^{(2)}\right)^{-1/2}$ and the sample sizes n_1, n_2.
- Our second result applies to settings where $\beta^{(1)}$ and $\beta^{(2)}$ can be arbitrarily different but $\Sigma^{(1)} = \Sigma^{(2)}$. In Theorem 3.5, we provide the asymptotic limit of $L(\hat{\beta}_2^{\mathrm{HPS}})$ as a function of the ℓ_2 distance between $\beta^{(1)}$ and $\beta^{(2)}$, and the sample sizes n_1, n_2.
- Our third result applies to settings where both the covariance matrices and the model vectors are different between the two tasks. In Theorem **??**, we show that when $\Sigma^{(2)}$ is isotropic, the asymptotic limit of $L(\hat{\beta}_2^{\mathrm{HPS}})$ is characterized by the singular values of $\Sigma^{(1)}$, the ℓ_2 distance between $\beta^{(1)}$ and $\beta^{(2)}$), and the sample sizes n_1, n_2. Next, in Theorem 3.7, we consider the most general setting concerning an arbitrary $\Sigma^{(2)}$. We identify an estimate of $L(\hat{\beta}_2^{\mathrm{HPS}})$ that becomes more and more accurate as n_1/n_2 increases.

Extension to multiple sources. We also extend our setup to the case of having multiple data sources. We consider a setting where all datasets have the same features but different labels, which is also known as multi-label prediction (Hsu et al., 2009). We show the asymptotic limit of $L(\hat{\beta}_i^{\mathrm{HPS}})$ under model shift (cf. Theorem 4.1) and use the result to illustrate how to select the hidden width r of B.

Theoretical insights and empirical implications. Next, we demonstrate the power of these precise asymptotics by explaining several empirical phenomena in transfer learning. We consider a random-effect model where $\beta^{(i)}$ is equal to a shared model vector plus per-task variances (cf. Dobriban and Wager (2018)). Under this model, we show the following *theoretical insights*.

i) *Whether or not covariate shift helps depends on sample sizes:* We consider the covariate shift setting and show that (cf. Proposition 3.3) when $n_1 < n_2$, transferring from a data source with $\Sigma^{(1)} \neq \Sigma^{(2)}$ can achieve a smaller $L(\hat{\beta}_2^{\mathrm{HPS}})$ compared to transferring from a data source with $\Sigma^{(1)} = \Sigma^{(2)}$. When $n_1 > n_2$, transferring from from a data source with $\Sigma^{(1)} \neq \Sigma^{(2)}$ always results in higher $L(\hat{\beta}_2^{\mathrm{HPS}})$ than transferring from a data source with $\Sigma^{(1)} = \Sigma^{(2)}$.

ii) *Identical covariance provides approximately optimal transfer under imbalanced sizes:* Additionally, in Proposition 3.4, we show that when n_1 is greater than n_2 times a certain large constant factor, transferring from a data source with $\Sigma^{(1)} = \Sigma^{(2)}$ achieves an approximately minimum excess risk value, within a certain family of covariance matrices of $\Sigma^{(1)}$.

iii) *Three regimes of information transfer under model shift*: Next, we consider the model shift setting and characterize the precise range of n_1 under which there is a positive transfer. In Proposition 3.6, we show that when $\mu^2 \leqslant \frac{\sigma^2 p}{2(n_2-p)}$ (recall that $\|\beta^{(1)} - \beta^{(2)}\|^2 \approx 2\mu^2$) is small, the transfer effect is positive for any n_1. When $\frac{\sigma^2 p}{2(n_2-p)} < \mu^2 < \frac{\sigma^2 n_2}{2(n_2-p)}$, there is a transition from positive to negative transfer, as n_1 increases. When $\frac{\sigma^2 n_2}{2(n_2-p)} \leqslant \mu^2$, the transfer effect is negative for any n_1. Figure 1 suggests that the asymptotic limits are also accurate in finite dimensions. Additionally, these three regimes of information transfer can be observed in settings of covariate and model shifts (cf. Figure 4b).

We also complement our theoretical analysis with empirical evaluations. As mentioned in Section 1.1, our motivation for studying the HPS estimator stems from its superior empirical performance for transfer learning. We affirm that HPS indeed outperforms several commonly used estimators including weighted averages of the OLS estimators of each task (cf. Bastani (2020)) and the ridge estimator of the target task in our setting. See Figure 6 for the result.

We then conduct a case study on six text classification tasks using HPS neural network models. We demonstrate that our theoretical insights above imply interesting new methods for mitigating covariate shift and model shift, respectively.

First, our insight ii) above suggests the importance of mitigating covariate shift under imbalanced dataset sizes. To this end, we consider a certain covariance alignment procedure proposed in Wu et al. (2020) for mitigating covariate shift. We show that such an alignment procedure provides more significant improvement when n_1/n_2 increases.

Second, our insight iii) above suggests subsampling the source dataset under model shift. An efficient instantiation of this above would be to increase the sample size of the source dataset gradually until performance drops. For example, in the setting of Figure 1, this procedure will stop right at the minimum of the local basin. On both two-task and multi-task case studies using the six text classification datasets, such a training procedure can reduce the computational cost by 65% compared to a stochastic gradient descent procedure without sacrificing accuracy.

1.3 Related work

Sample covariance matrices under distribution shift. In the *underparametrized* setting ($n_2 > p$) which this work focuses on, when the covariates are sampled from Gaussian distributions, the precise asymptotic limit can be derived directly from properties of the Wishart distribution (see also Lemma 2.5 and the historical notes in Section 2.3). In the high-dimensional setting, the eigenvalues of Wishart matrices satisfy the well-known Marchenko–Pastur (MP) law (Marčenko and Pastur, 1967), whose Stieltjes transform characterizes the variance limit. Furthermore, it is well-known that the MP law holds universally regardless of the underlying data distribution of the covariates (**?**). Bloemendal et al. (2014) obtained a sharp convergence rate of the empirical spectral distribution (ESD) of the MP law for sample covariance matrices with isotropic population covariance matrices. Knowles and Yin (2016); Ding and Yang (2018) later extended this result to sample covariance matrices with arbitrary population covariances. These results are proved by establishing certain optimal convergence estimates of the Stieltjes transforms of sample covariance matrices, which are referred to as *local laws* in the random matrix theory literature. We refer interested readers to Erdos and Yau (2017) and the references therein for a detailed review of related concepts. Our technical contribution is to extend these techniques to the two-task setting and prove a local law for the sum of two sample covariance matrices with arbitrary covariate shifts. Using the local law, we can then derive the variance limit with a precise dependence on the singular values of the covariate shift matrix $\left(\Sigma^{(1)}\right)^{1/2}\left(\Sigma^{(2)}\right)^{-1/2}$.

The asymptotic limit of the variance of $L(\hat{\beta}_2^{\mathrm{HPS}})$ (cf. equation (2.9)) under covariate shift may also be derived using free addition in free probability theory (Nica and Speicher, 2006). However, this approach is not completely justified when the covariates are sampled from non-Gaussian distributions with non-diagonal covariate shift matrices. Furthermore, our technique provides almost sharp convergence rates to the asymptotic limit, while it is unclear how to obtain such rates using free probability techniques, to the best of our knowledge.

The asymptotic limit of the bias of $L(\hat{\beta}_2^{\mathrm{HPS}})$ (cf. equation (2.8)) involves "asymmetric" additions of two sample covariance matrices, whose analysis is technically involved. Our result on the bias limit is inspired by

the recent work of Bao et al. (2017a,b). Additionally, we show a sharp convergence rate to the bias limit using free probability techniques. Our work provides the first result in the model shift setting assuming that $\Sigma^{(1)} = \Sigma^{(2)}$ or that $\Sigma^{(2)}$ is isotropic, and the covariates are sampled from Gaussian distributions. Showing the asymptotic bias limit under the most general covariate shift and model shift is an interesting open problem for future work. See Section 6 for further discussion on the technical challenges.

Transfer learning theory. Earlier works such as Baxter (2000); Ben-David and Schuller (2003); Maurer (2006) use uniform convergence to obtain generalization bounds on transfer learning (also called domain adaptation). The influential paper by Ben-David et al. (2010) proposes a rigorous model for domain adaptation in classification (see also Crammer et al. (2008) for an earlier result). Their main result provides a generalization bound for combining multiple data sources using an objective that averages over the risk of every datapoint. This bound provides a principled way to reweight each data source by minimizing the bound. From a technical perspective, uniform convergence-based techniques provide upper bounds on the excess risks. However, analyzing the empirical phenomena of information transfer such as the ones in Figure 1 requires tight lower bounds on the excess risks. The very recent work of Kalan et al. (2020) introduced a minimax framework to quantify the fundamental limit of transfer learning. Their result quantities the minimax risk using the distance between the source and target models. Our work provides precise estimates that accurately match the empirical risk of commonly used estimators compared to their work. These precise estimates are necessary for analyzing these estimators' transfer effect.

Finally, we discuss several concurrent works that study transfer learning in linear regression settings. The work of Li et al. (2020) considers the problem of how to select useful data sources for transfer learning under model shift. Their work provides an adaptive procedure that applies even when the set of beneficial data sources is unknown. Our work differs from theirs in two aspects. First, we considered a high-dimensional setting where the sample size increases with dimension proportionally by a constant factor. We use this setting to reproduce several interesting phenomena in transfer learning. Second, our estimates apply even under covariate shift, whereas the result of Li et al. (2020) requires that there is no covariate shift between the data sources and the target task. Another work by Lei et al. (2021) considers a linear regression setting under distribution shift, including covariate shift and model shift. The difference is that our setting deals with a labeled target dataset. In contrast, Lei et al. (2021) deal with an unlabeled target dataset (i.e., unsupervised domain adaptation).

Random matrix theory in machine learning. This work joins an emerging line of recent work that has found random matrix theory useful for studying modern machine learning techniques. A prominent example is looking at interpolators in over-parametrized linear and logistic regression (Bartlett et al., 2020; Hastie et al., 2019; Montanari et al., 2019; Liang et al., 2020; Liang and Sur, 2020). Interpolators can reproduce similar phenomena observed under neural networks because both models learn by "interpolating" the data using many parameters (Mei and Montanari, 2019). The precise rates achieved via random matrix theory can also explain the double-descent phenomenon (Belkin et al., 2019). Whereas these works deal with covariates drawn from a single distribution, our work instead tackles distribution shift given covariates sampled from two distributions.

1.4 Organizations

The rest of this paper is organized as follows. In Section 2, we formally define the data model and describe the underlying assumptions. We show a bias-variance decomposition of the HPS estimator, and connect the bias and variance equations to random matrix theory. In Section 3, we show precise estimates for the excess risk of the HPS estimator under covariate and model shifts. In Section 5, we evaluate the performance of HPS estimators against several commonly used transfer learning estimators, and present two implications for migating covariate and model shift in neural networks. In Section 4, we extend our setup to transferring from multiple data sources under model shift. Section 6 concludes the paper and discusses potential future directions. Section **??**, **??**, and **??** present proofs of our results.

2 Preliminaries

We begin by formally defining the data model involving large dimensional random matrices together with the underling assumptions. We define the covariate shift and model shift settings. Then, we describe the HPS estimators for these settings using two-layer linear neural networks. Finally, we present a bias-variance decomposition of the HPS estimators. We explain why this decomposition is crucial for analyzing information transfer and the need for new techniques in the analysis. Our setup in this section concerns the case of a single data source. Later on in Section 4, we will extend the setup to the case of multiple data sources.

2.1 Data model

Recall we have two regression datasets (or tasks) with sample sizes denoted as n_1 and n_2, respectively. Suppose that $i = 1$ or $i = 2$. Let $X^{(i)} \in \mathbb{R}^{n_i \times p}$ be the matrix notation corresponding to dataset i's covariates. Let $Y^{(i)} \in \mathbb{R}^{n_i}$ and $\varepsilon \in \mathbb{R}^{n_i}$ be the vector notation corresponding to its labels and additive noise. Equation (1.1) thus implies that $Y^{(i)} = X^{(i)}\beta^{(i)} + \varepsilon^{(i)}$. Both $\beta^{(1)}$ and $\beta^{(2)}$ are two arbitrary deterministic or random vectors that are independent of the feature covariates and the noise vector. Throughout the paper, we make the following assumptions on $X^{(i)}$ and $\varepsilon^{(i)}$, which are standard in the random matrix theory literature (see e.g., Tulino and Verdú (2004); Bai and Silverstein (2006)).

First, recall that the row vectors of $X^{(i)}$ are i.i.d. centered random vectors with population covariance matrix $\Sigma^{(i)} \in \mathbb{R}^{p\times p}$:

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

where $Z^{(i)} = (z^{(i)}_{j,k})$ is a $n_i \times p$ random matrix with real-valued independent entries having zero mean and unit variance. Let $\tau > 0$ be a small constant. We assume that for some constant $\varphi > 4$, the φ-th moment of each entry $z^{(i)}_{j,k}$ is bounded from above by $1/\tau$:

$$\mathbb{E}\left[|z^{(i)}_{j,k}|^{\varphi}\right] \leqslant \frac{1}{\tau}. \tag{2.2}$$

Additionally, We assume that the eigenvalues of $\Sigma^{(i)}$ are bounded between τ and $1/\tau$.

Second, we assume that $\varepsilon^{(i)} \in \mathbb{R}^{n_i}$ is a random noise vector with i.i.d entries having 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

$$\mathbb{E}\left[|\varepsilon^{(i)}_j|^k\right] \leqslant C_k. \tag{2.3}$$

Finally, we assume a high-dimensional setting when the sample sizes are proportional to the dimension up to a constant factor. Denote by $\rho_i = n_i/p$, we require that

$$\tau \leqslant \rho_1 \leqslant p^{\tau^{-1}} \text{ and } 1 + \tau \leqslant \rho_2 \leqslant p^{\tau^{-1}}. \tag{2.4}$$

The first inequality ($1+\tau \leqslant \rho_i$) ensures that the sample covariance matrices $(X^{(1)})^\top X^{(1)}$ and $(X^{(2)})^\top X^{(2)}$ are non-singular with high probability. The second inequality ($\rho_i \leqslant p^{\tau^{-1}}$) is without loss of generality. Otherwise, standard concentration results such as the central limit theorem already imply accurate estimates of the linear model under the setting defined by equation (2.4).

To summarize, the underlying assumptions in our data model are as follows.

Assumption 2.1 (Data model). Let τ be a small constant. Let $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are deterministic positive definite symmetric matrices whose eigenvalues are bounded between τ and $1/\tau$.

(i) $X^{(1)}$ and $X^{(2)}$ take the form of equation (2.1), where $Z^{(1)}$ and $Z^{(2)}$ are both random matrices with i.i.d. entries having zero mean, unit variance, and bounded moments as in equation (2.2).

(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 equation (2.3).

(iii) ρ_1 and ρ_2 satisfy equation (2.4). Furthermore, $\tau \leqslant \frac{\rho_1}{\rho_2} \leqslant \tau^{-1}$.

We will consider a random-effect model as our running example, which has been studied in several recent works concerning high-dimensional (single-task) ridge regression (Dobriban and Wager, 2018) and distributed regression (Dobriban and Sheng, 2020). Suppose each $\beta^{(i)}$ consists of two random components, one that is shared among the two tasks and one that is task-specific. More precisely, let

$$\beta^{(i)} = \beta_0 + \widetilde{\beta}^{(i)}, \quad i = 1, 2, \tag{2.5}$$

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 μ^2/p.

2.2 Risks

We measure the performance of $\hat{\beta}_i^{\mathrm{HPS}}$ over samples (x, y) generated from the i-th data model: $y = x^\top \beta^{(i)} + \varepsilon$. The expected prediction risk (under mean squared loss) is given by

$$\mathop{\mathbb{E}}_{x,\varepsilon}\left[\left\|y - x^\top \hat{\beta}_i^{\mathrm{HPS}}\right\|^2\right] = \left\|\left(\Sigma^{(i)}\right)^{1/2}\left(\hat{\beta}_i^{\mathrm{HPS}} - \beta^{(i)}\right)\right\|^2 + \sigma^2.$$

The excess risk is the difference between the above risk and the expected risk of the population risk optimizer (which is equal to σ^2):

$$L(\hat{\beta}_i^{\mathrm{HPS}}) := \left\|\left(\Sigma^{(i)}\right)^{1/2}\left(\hat{\beta}_i^{\mathrm{HPS}} - \beta^{(i)}\right)\right\|^2. \tag{2.6}$$

We focus on the setting where the hidden dimension $r = 1$. Otherwise, the excess risk of the HPS estimator and that of the ordinary least squares (OLS) estimator are equal to each other (Wu et al., 2020).

Proposition 2.2. *Suppose that Assumption 2.1 holds. Let $\hat{\beta}_i^{\mathrm{OLS}} = \left((X^{(i)})^\top X^{(i)}\right)^{-1}(X^{(i)})^\top Y^{(i)}$ denote the OLS estimator, for either $i = 1$ or $i = 2$. Then, for any $r \geqslant 2$, we have that $L(\hat{\beta}_i^{\mathrm{HPS}}) = L(\hat{\beta}_i^{\mathrm{OLS}})$.*

Proof. Since both n_1 and n_2 are at least $(1+\tau)p$, the inverse of both ${X^{(1)}}^\top X^{(1)}$ and ${X^{(2)}}^\top X^{(2)}$ exists. Thus, both $\hat{\beta}_1^{\mathrm{OLS}}$ and $\hat{\beta}_2^{\mathrm{OLS}}$ are well-defined. The global minimum of each summand of equation (1.2) is achieved if and only if $\hat{\beta}_1^{\mathrm{OLS}}$ and $\hat{\beta}_2^{\mathrm{OLS}}$ are both in the column subspace of A. In this case, $\hat{\beta}_1^{\mathrm{HPS}} = AB_1$ must be equal to $\hat{\beta}_1^{\mathrm{OLS}}$ (same for $\hat{\beta}_2^{\mathrm{HPS}}$). □

When $r = 1$, both B_1 and B_2 reduce to scalars. In particular, the case when B_1 and B_2 are both equal to 1 reduces to a pooling estimator, which treats both datasets as if they are drawn independently from the same distribution. Such pooling estimators have been shown to be effective on e-commence and healthcare datasets (Bastani, 2020). Additionally, they have been used as a preprocessing step of minimax optimal estimators, under certain conditions of the data (Li et al., 2020) (see also Section 1.3 for further discussion of both works).

2.3 Bias and variance

Next, we provide the bias-variance decomposition of the excess risk under this setting. Without loss of generality, we focus on task two's excess risk $L(\hat{\beta}_2^{\mathrm{HPS}})$. As a remark, for general r, problem (1.2) is non-convex since $f(\cdot)$ is a degree-four polynomial. Nevertheless, the setting when $r = 1$ remains tractable since we can write a closed-form solution of A given any (scalar) B_1 and B_2. Since B_1 and B_2 are both scalars, let $a = B_1/B_2$. Let $\hat{\Sigma}(a) := a^2(X^{(1)})^\top X^{(1)} + (X^{(2)})^\top X^{(2)}$ be a weighted combination of the sample covariance matrices. We present a bias-variance decomposition of $L(\hat{\beta}_2^{\mathrm{HPS}})$ for the setting of $r = 1$ as follows. We say that an event Ξ holds *with high probability* (w.h.p.) if $\mathbb{P}(\Xi) \to 1$ as $p \to \infty$.

Proposition 2.3 (Bias-variance decomposition)**.** *Under Assumption 2.1, for any small constant $c > 0$ and large constant $C > 0$, there exists an event Ξ, on which the following estimates hold uniformly in $a \in \mathbb{R}$ and $\lambda \in [0, 1]$ w.h.p.:*

$$L(\hat{\beta}_2^{\mathrm{HPS}}) = \left(1 + \mathrm{O}(p^{-1/2+c})\right) \cdot \left(L_{\mathrm{Bias}}(a) + L_{\mathrm{Var}}(a)\right) + \mathrm{O}\left(p^{-C}\left(\|\beta^{(1)}\|^2 + \|\beta^{(2)}\|^2\right)\right), \tag{2.7}$$

where the bias and variance equations are defined as follows:

$$L_{\text{Bias}}(a) := \left\| (\Sigma^{(2)})^{1/2} \hat{\Sigma}(a)^{-1} (X^{(1)})^\top X^{(1)} \left(a\beta^{(1)} - a^2 \beta^{(2)} \right) \right\|^2, \tag{2.8}$$

$$L_{\text{Var}}(a) := \sigma^2 \operatorname{Tr}\left[\Sigma^{(2)} \hat{\Sigma}(a)^{-1} \right]. \tag{2.9}$$

Equation (2.7) shows that the excess risk of the HPS estimator admits a clean bias-variance decomposition plus lower-order terms that decrease to zero with the dimension p. Since equation (2.7) holds uniformly for all $a \in \mathbb{R}$, we can apply the result to $\hat{a} = \hat{B}_1/\hat{B}_2$, for any $\hat{B}_1$ and $\hat{B}_2$ that may depend on $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ (after solving problem (1.2)). We briefly describe why Proposition 2.3 holds. First, suppose that a is an arbitrary value that is independent of the random noise vectors $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$. By taking the expectation over both of them, we get the following result:

$$\mathop{\mathbb{E}}_{\varepsilon^{(1)}, \varepsilon^{(2)}} \left[L(\hat{\beta}_2^{\text{HPS}}) \right] = L_{\text{Bias}}(\hat{a}) + L_{\text{Var}}(\hat{a}). \tag{2.10}$$

Secondly, 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 (2.10) up to a small error as in the next lemma. For a fixed $a \in \mathbb{R}$, the proof of (2.7) is based on the sharp concentration bounds in Lemma **??** of Section **??**. 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 Proposition 2.3 can be found in Section **??**.

Equation (2.10) suggests that the bias and variance of HPS both depend on the scalar $\hat{a}$. We provide some intuition about this scalar in the setting of random-effect models. We show that under natural assumptions, the global minimizer $\hat{a} = \hat{B}_1/\hat{B}_2$ is approximately equal to 1. The proof of Proposition 2.4 can be found in Section **??**.

Proposition 2.4 (Random-effect model)**.** *Suppose Assumption 2.1 and the random-effect model under equation (2.5) holds. Suppose that for a constant $c_0 > 0$,*

$$\|\beta_0\|^2 \geqslant p^{c_0}\mu^2 + p^{-1/2+c_0}\sigma^2. \tag{2.11}$$

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

$$\hat{a} = 1 + \mathrm{O}\left(\frac{\mu^2}{\|\beta_0\|^2} + p^{-\frac{1}{4}+c} \frac{\mu + \sigma}{\|\beta_0\|} + p^{-C} \right) \quad w.h.p. \tag{2.12}$$

≪Todo notes: change d to μ because a reviewer complained≫

We now illustrate that the bias-variance decomposition of Proposition 2.3 characterizes whether or not the information transfer is positive. First, we compare the bias. Since $n_1 \geqslant (1+\tau)p$ by Assumption 2.1, the bias of $\hat{\beta}_2^{\text{OLS}}$ is zero. Thus, the bias of $\hat{\beta}_2^{\text{HPS}}$ is always higher than that of $\hat{\beta}_2^{\text{OLS}}$. Second, we compare the variance. By a similar argument as in Proposition 2.3, the variance of $\hat{\beta}_2^{\text{OLS}}$ is equal to $\sigma^2 \operatorname{Tr}\left[\Sigma^{(2)}((X^{(2)})^\top X^{(2)})^{-1}\right]$. Then, a straightforward algebraic calculation (e.g. using the Woodbury matrix identity) shows that the variance of $\hat{\beta}_2^{\text{HPS}}$ is always higher than that of $\hat{\beta}_2^{\text{OLS}}$:

$$\operatorname{Tr}\left[\Sigma^{(2)}\left(\hat{\Sigma}(a)\right)^{-1}\right] \leqslant \operatorname{Tr}\left[\Sigma^{(2)}\left((X^{(2)})^\top X^{(2)}\right)^{-1}\right].$$

Combining both comparisons, we conclude that the transfer effect of HPS is mixed: the bias always increases while the variance always decreases. Thus, whether or not the resulting information transfer is *positive* depends on which effect is more significant.

Building on the above connection, we apply recent developments from random matrix theory to analyze information transfer. As a warmup, we describe the following classical result in multivariate statistics that characterizes the excess risk (or equivalently, the variance) of the OLS estimator.

Lemma 2.5 (Variance estimates of OLS estimators, cf. Theorem 2.4 of Bloemendal et al. (2014) and Theorem 3.14 of Ding and Yang (2018)). *Under Assumption 2.1, we have that for any deterministic matrix* $\Sigma \in \mathbb{R}^{p\times p}$,

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

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

≪Todo notes: the scaling $\sqrt{n_2}$ above doesn't seem correct≫

If the entries of $Z^{(2)}$ are i.i.d. Gaussian, then this result follows from the classical result for the inverse Wishart distribution (Anderson, 2003). For a general non-Gaussian $Z^{(2)}$, this result can be obtained using the well-known Stieltjes transform method (cf. Lemma 3.11 of Bai and Silverstein (2010)). Here we have presented the results from Bloemendal et al. (2014) and Ding and Yang (2018), who give an almost sharp convergence rate.

3 Precise asymptotics under distribution shift

We show precise estimates of the bias and variance of HPS and almost sharp convergence rates. In Section 3.1, we characterize the variance of HPS under covariate shift, which extends the classical result from Lemma 2.5 to the sum of two sample covariance matrices with covariate shift. Based on the characterization, We provide a detailed analysis of the impact of covariate shift on the excess risk of HPS. In Section 3.2, we characterize the bias of HPS under model shift. Combined with the variance estimates, we describe three regimes of information transfer in the random-effect model, depending on the sample size n_1 and the model shift parameter μ. Finally, Section 3.3 extends these results to settings under covariate and model shifts.

3.1 Covariate shift

We begin by considering the covariate shift setting where both tasks have the same linear model but different population covariance matrices. We show the exact asymptotic limit of the excess risk of the HPS estimator in the high-dimensional setting.

First, we consider the variance. Recall from equation (2.9) that the variance equation is equal to $\sigma^2\cdot\operatorname{Tr}\left[\Sigma^{(2)}(\hat{\Sigma}(a))^{-1}\right]$. In particular, the matrix $\hat{\Sigma}$ adds up both tasks' sample covariance matrices. Thus, the expectation of $\hat{\Sigma}(a)$ is equal to a mixture of both tasks' population covariance matrices, with mixing proportions determined by their sample sizes. Intuitively, the spectrum of $\hat{\Sigma}(a)^{-1}$ now not only depends on the sample sizes of both tasks, but also depends on how well "aligned" $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are. To capture this alignment quantitatively, we introduce the covariate shift matrix

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

Let $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$ be the singular values of M in descending order. Our first main result 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 3.1 (Variance estimates under covariate shift). *Under Assumption 2.1 (recalling* $\varphi > 4$*), for any small constant* $c > 0$*, there exists a high probability event* Ξ*, on which the following estimate holds for* $L_{\mathrm{Var}}(a)$ *in (2.9):*

$$\left|L_{\mathrm{Var}}(a) - \frac{\sigma^2}{n_1+n_2}\operatorname{Tr}\left[\frac{1}{\alpha_1 a^2\cdot M^\top M+\alpha_2}\right]\right| \leqslant \frac{(n_1+n_2)^{\frac{2}{\varphi}-\frac{1}{2}+c}}{p^{1/2}}\cdot\frac{p\sigma^2}{n_1+n_2}, \tag{3.1}$$

uniformly in all $a\in\mathbb{R}$. *Here* (α_1,α_2) *is the solution of the following system of equations*

$$\alpha_1+\alpha_2 = 1-\frac{p}{n_1+n_2},\quad \alpha_1+\frac{1}{n_1+n_2}\left(\sum_{i=1}^{p}\frac{(a\lambda_i)^2\alpha_1}{(a\lambda_i)^2\alpha_1+\alpha_2}\right) = \frac{n_1}{n_1+n_2}. \tag{3.2}$$

≪Todo notes: changed a_1, a_2 to α_1, α_2 since they conflict with a≫

Equation (3.1) thus characterizes the variance of the HPS estimator, up to an error term on the order of $O\big(\frac{\sigma^2}{p^{1/2}}\big)$ (recall that $\varphi > 4$, c is an arbitrarily small constant, and $\frac{p}{n_1+n_2} \leqslant \frac{1}{2+2\tau}$). Theorem 3.1 generalizes the classical result from Lemma 2.5 to the sum of two independent sample covariance matrices. To see this, we note that Lemma 2.5 corresponds to a setting where the singular values $\lambda_1, \ldots, \lambda_p$ (of the matrix M) are all equal to one, and $a = 1$. In this setting, the second part of equation (3.2) simplifies to

$$\alpha_1 + \frac{p}{n_1+n_2} \cdot \frac{\alpha_1}{\alpha_1+\alpha_2} = \frac{n_1}{n_1+n_2}.$$

By solving the above equation and the first part of equation (3.2), we obtain that

$$\alpha_1 = \frac{n_1}{n_1+n_2}, \quad \alpha_2 = \frac{n_2-p}{n_1+n_2}.$$

Plugging the solutions of α_1, α_2 back to the LHS of equation (3.1), we thus find that

$$\begin{aligned}
\frac{\sigma^2}{n_1+n_2} \operatorname{Tr}\left[\frac{1}{\alpha_1 M^\top M + \alpha_2}\right] &= \frac{\sigma^2}{n_1+n_2}\frac{p}{\alpha_1+\alpha_2} && \text{(since } \lambda_i = 1 \text{ for all } i = 1, \ldots, p\text{)} \\
&= \frac{\sigma^2}{n_1+n_2} \cdot \frac{p(n_1+n_2)}{n_1+n_2-p} && \text{(applying the solutions of } \alpha_1 \text{ and } \alpha_2\text{)} \\
&= \frac{\sigma^2 p}{n_1+n_2-p}, && (3.3)
\end{aligned}$$

which is precisely the limit in Lemma 2.5. Additionally, the convergence rate in equation (3.1) matches the rate in Lemma 2.5 for this particular setting. The proof of Theorem 3.1, which is based on recent developments in random matrix theory (Knowles and Yin, 2016), can be found in Section **??**.

Second, we consider the bias. Note that the setting where $\beta_1 = \beta_2$ can be viewed as a special case of the random-effect model with $d = 0$, thus, assuming that $\|\beta_1\|_2^2 \geqslant \sigma^2/p^{1/2-c_0}$ for a small constant $c_0 > 0$, we have that $\hat{B}_1/\hat{B}_2$ is equal to 1 plus lower-order terms that scale to zero as p goes to infinity (see equation (2.12) for the precise scaling). In Proposition 2.4, we have seen that the global minimizer $\hat{a}$ is close to 1 up to a small error. Thus, the bias equation $L_{\text{Bias}}(\hat{B}_1/\hat{B}_2)$ is equal to 0 plus the above lower-order terms. We summarize this discussion in the following corollary:

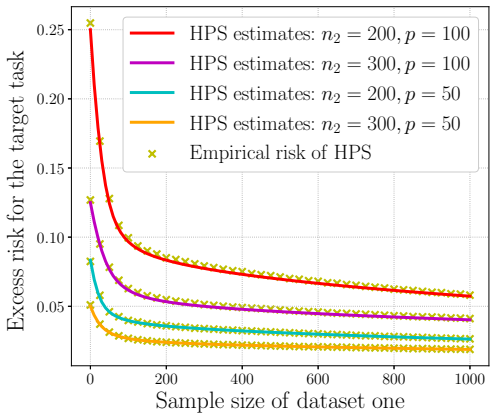
Corollary 3.2 (Excess risk of HPS under covariate shift)**.** *Under Assumption 2.1, suppose further that* $\beta_1 = \beta_2$ *and* $\|\beta_1\|_2^2 \geqslant \frac{\sigma^2}{p^{1/2-c_0}}$ *for a constant* $c_0 > 0$*. Then, for any constant* $c > 0$*, the following estimate on the excess risk of the HPS estimator holds w.h.p.*

$$\left| L(\hat{\beta}_2^{\text{HPS}}) - \frac{\sigma^2}{n_1+n_2} \operatorname{Tr}\left[\frac{1}{\alpha_1 M^\top M + \alpha_2}\right] \right| \leqslant \mathrm{O}\left(\frac{(n_1+n_2)^{\frac{2}{\varphi}-\frac{1}{2}+c}}{p^{1/2}} \cdot \frac{p\sigma^2}{n_1+n_2} + p^{-c_0+2c} \right), \tag{3.4}$$

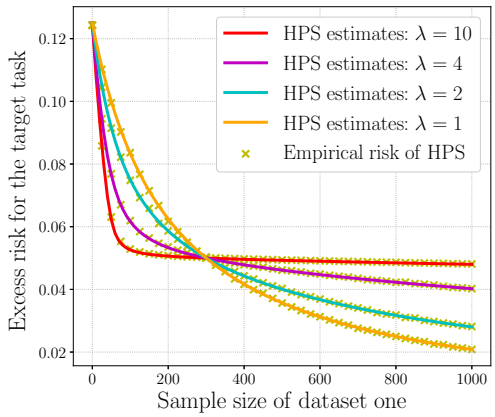
where α_1 *and* α_2 *are the solutions of equation* (3.2) *after taking* $a = 1$*.*

Illustrative examples. Next, we illustrate the result of Corollary 3.2 in several examples. We use our estimates to explore whether covariate shift helps or hurts information transfer. Our first example illustrates that the effect of covariate shift depends on the sample sizes of each dataset in an intricate manner. While the folklore belief is that transferring from a covariate-shifted distribution performs worse than an identical distribution, we show that, surprisingly, the former can sometimes outperform the latter.

We first describe a setting for modeling covariate shift. Let $\mathcal{S}$ be the set of positive definite matrices $M^\top M$ such that for $i = 1, 2, \ldots, \lfloor \frac{p}{2} \rfloor$, $\lambda_{p-i} = \frac{1}{\lambda_i}$, where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$ are the eigenvalues of M. Under this setting, we ask: suppose that the sample sizes n_1 and n_2 are both fixed, which M provides the best "transfer" in the sense of HPS estimator's excess risk? We observe a dichotomy that depends on whether or not n_1 is greater than n_2.



(a) Illustrating HPS estimates in finite dimensions



(b) Varying the covariate shift parameter λ

Figure 2: We verify that Theorem 3.1 provides incredibly accurate estimates of the empirical variance under various finite sample sizes and dimensions. Furthermore, we illustrate an intriguing dichotomy between sample sizes and covariate (cf. Claim 3.3). When $n_1 \geqslant n_2$, the lowest risk for predicting task two's labels is achieved by transferring from a dataset that has the same covariance matrix as task two. When $n_1 < n_2$, the lowest risk is instead achieved by transferring a covariate-shifted dataset. Both simulation uses $\sigma = 1/2$. Figure 2a fixes $\lambda = 4$ and varies n_1, p. Figure 2b fixes $p = 100, n_2 = 300$ while varying n_1, λ.

Proposition 3.3 (Sample sizes vs. covariate shift)**.** *Let* $g(M) = \frac{\sigma^2}{n_1+n_2} \operatorname{Tr}\left[\frac{1}{\alpha_1 M^\top M + \alpha_2}\right]$. *Suppose that* n_1 *and* n_2 *are both fixed. Within the set of all possible covariate shift matrix* $M \in \mathcal{S}$*, the following dichotomy holds:*

i) If $n_1 \geqslant n_2$*, then* $g(M)$ *is minimized in* $\mathcal{S}$ *when* $\lambda_i = 1$*, for any* $i = 1, \ldots, p$.

ii) If $n_1 < n_2$*, then* $g(M)$ *is maximized in* $\mathcal{S}$ *when* $\lambda_i = 1$*, for any* $i = 1, \ldots, p$.

Proof. For any $M \in \mathcal{S}$, we can write $g(M)$ as

$$g(M) = \frac{\sigma^2}{n_1 + n_2} \sum_{i=1}^{p/2} \left(\frac{1}{\lambda_i^2 \alpha_1 + \alpha_2} + \frac{1}{\lambda_i^{-2} \alpha_1 + \alpha_2} \right).$$

When $M = \mathrm{Id}_{p \times p}$, by the first equation of (3.2), we have

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

where we abbreviate $\gamma := p/(n_1 + n_2)$. Using $\alpha_1 + \alpha_2 = 1 - \gamma$, we further have that

$$g(M) - g(\mathrm{Id}_{p\times p}) = \frac{\sigma^2}{n_1 + n_2 - p} \sum_{i=1}^{p/2} \frac{(\lambda_i^2 - 1)^2 \alpha_1 \left[\alpha_1 - \alpha_2\right]}{[\alpha_1 + \lambda_i^2 \alpha_2][\lambda_i^2 \alpha_1 + \alpha_2]}.$$

We claim that $\alpha_1 > \alpha_2$ if and only if $n_1 > n_2$, which proves claim i). In fact, if $\alpha_1 > \alpha_2$, then the first part of equation (3.2) gives that $\alpha_1 > (1-\gamma)/2$. The second part of equation (3.2) gives that

$$\frac{n_1}{n_1 + n_2} > \alpha_1 + \frac{1}{n_1 + n_2} \sum_{i=1}^{p/2} \left(\frac{\lambda_i^2}{\lambda_i^2 + 1} + \frac{\lambda_i^{-2}}{\lambda_i^{-2} + 1} \right) = \frac{1-\gamma}{2} + \frac{\gamma}{2} = \frac{1}{2},$$

which is equivalent to $n_1 > n_2$. The proof of claim ii) follows from a similar argument. □

Figure 2b illustrates a special case where $\lambda_1 = \cdots = \lambda_{\lfloor p/2 \rfloor} = \lambda > 0$ and the rest of the eigenvalues are all equal to $1/\lambda$. Thus, $\lambda > 0$ captures the degree of covariate shift: higher λ implies worse covariate shift. We observe that our theoretical estimates using Corollary 3.2 matches the empirical risks incredibly accurately. As a result, we indeed observe the dichotomy shown in Claim 3.3. Furthermore, for higher λ, HPS' excess risk decreases slower–indicating a worse rate of "transfer" from task one. As a remark, the classical work of David et al. (2010) have likewise shown impossibility results for transfer learning under covariate shift in a classification setting. By contrast, our analysis applies to a (high-dimensional) regression setting.

Our second example illustrates that the effect of covariate shift worsens as the sample size of task one increases. We consider a set of covariate shift matrices $M^\top M$ whose determinant are all equal to one, and the eigenvalues are bounded between τ^2 and $1/\tau^2$. Let $\mathcal{S}$ denote the set of $M \in \mathbb{R}^{p \times p}$ such that $M^\top M$ satisfies both conditions.

Proposition 3.4 (Identical covariance provides approximately optimal transfer under imbalanced dataset sizes)**.** *Recall that* $g(M) = \frac{\sigma^2}{n_1+n_2} \operatorname{Tr}\left[\frac{1}{\alpha_1 M^\top M + \alpha_2}\right]$. *We have the following result regarding* $g(\mathrm{Id}_{p\times p})$*:*

$$g(\mathrm{Id}_{p\times p}) \leqslant \left(1 + \frac{n_2}{\tau^2 n_1}\right) g(M), \text{ for any } M \in \mathcal{S}. \tag{3.5}$$

Proof. We can write the trace of $(\alpha_1 M^\top M + \alpha_2)^{-1}$ using the eigenvalues of $M^\top M$ as follows:

$$\begin{aligned}
\operatorname{Tr}\left[(\alpha_1 M^\top M + \alpha_2)^{-1}\right] &= \sum_{i=1}^{p} \frac{1}{\alpha_1 \lambda_i^2 + \alpha_2} \\
&\geqslant \sum_{i=1}^{p} \frac{1}{\alpha_1 \lambda_i^2 + \alpha_2 \cdot \lambda_i^2/\tau^2} && \text{(since } \lambda_i \geqslant \tau \text{, for any } i\text{)} \\
&= \frac{1}{\alpha_1 + \alpha_2/\tau^2} \sum_{i=1}^{p} \frac{1}{\lambda_i^2} \\
&\geqslant \frac{1}{\alpha_1 + \alpha_2/\tau^2} p \cdot \left(\prod_{i=1}^{p} \frac{1}{\lambda_i^2}\right)^{1/p} && \text{(by the AM-GM inequality)} \\
&= \frac{p}{\alpha_1 + \alpha_2/\tau^2} && \text{(since } \textstyle\prod_{i=1}^{p} \lambda_i^2 = \det(M^\top M) = 1\text{)}
\end{aligned}$$

Next, recall from equation (3.3) that $g(\mathrm{Id}_{p\times p}) = \frac{\sigma^2 p}{n_1+n_2-p}$. Thus, equation (3.5) follows if we show that (after rearranging terms)

$$\frac{n_1+n_2}{n_1+n_2-p} \cdot \frac{1}{\alpha_1 + \frac{\alpha_2}{\tau^2}} \leqslant 1 + \frac{n_2}{\tau^2 \cdot n_1}. \tag{3.6}$$

From the second part of equation (3.2), we have that

$$\begin{aligned}
\frac{n_1}{n_1+n_2} &= \alpha_1 + \frac{1}{n_1+n_2}\left(\sum_{i=1}^{p} \frac{\lambda_i^2 \alpha_1}{\lambda_i^2 \alpha_1 + \alpha_2}\right) \\
&< \alpha_1 + \frac{p}{n_1+n_2}.
\end{aligned}$$

Thus, $\alpha_1 > \frac{n_1-p}{n_1+n_2}$, which implies that $\alpha_2 < \frac{n_2}{n_1+n_2}$. Hence, $\alpha_1 + \frac{\alpha_2}{\tau^2} = 1 - \frac{p}{n_1+n_2} - \alpha_2 + \frac{\alpha_2}{\tau^2} \leqslant 1 - \frac{p}{n_1+n_2} + \frac{n_2}{n_1+n_2}\frac{1}{\tau^2}$, which implies equation (3.6) by straightforward calculations. Thus, the proof is complete. □

As a corollary of Claim 3.4, when n_1 is much larger than n_2/τ^2, the excess risk of HPS is approximately optimal in the set of all possible $M \in \mathcal{S}$ when there is no covariate shift between dataset one and two.

3.2 Model shift

Next, we consider a model shift setting where task one and two have the same linear model but different population covariance matrices. We show the exact asymptotic limits of the bias and variance of HPS.

Theorem 3.5 (Excess risk of HPS under model shift)**.** *Under Assumption 2.1, suppose that* $\Sigma^{(1)} = \Sigma^{(2)}$ *and the entries of* $Z^{(1)}$ *and* $Z^{(2)}$ *are i.i.d. Gaussian random variables. Denote by* $\xi_1 := p/n_1$ *and* $\xi_2 := p/n_2$*. 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* $a \in \mathbb{R}$*:*

$$L_{\mathrm{Var}}(a) = \sigma^2 \mathcal{L}_1(a) + \mathrm{O}\left(\frac{\sigma^2 p^c}{n_1}\right), \tag{3.7}$$

$$L_{\mathrm{Bias}}(a) = \left\|\beta^{(1)} - a\beta^{(2)}\right\|^2 \mathcal{L}_2(a) + \mathrm{O}\left(p^{-\frac{1}{2}+c}\|\beta^{(1)} - a\beta^{(2)}\|^2 + p^{-C}\left(\|\beta^{(1)}\|^2 + \|\beta^{(2)}\|^2\right)\right), \tag{3.8}$$

where $\mathcal{L}_1(a)$ *and* $\mathcal{L}_2(a)$ *are defined as follows*

$$\mathcal{L}_1(a) := 2p \cdot \left((n_2 - p) + a^2(n_1 - p) + \sqrt{\left((n_2 - p) + a^2(n_1 - p)\right)^2 + 4a^2(n_1 p + n_2 p - n_1 n_2)}\right)^{-1},$$

$$\mathcal{L}_2(a) := \frac{1}{a^2} \cdot \frac{1 - 2\frac{\mathcal{L}_1(a)}{\xi_2(1+\mathcal{L}_1(a))} + \kappa(a)}{1 - \xi_2\kappa(a)}, \text{ in which } \kappa(a) := \frac{\mathcal{L}_1(a)^2}{\xi_2^2(1+\mathcal{L}_1(a))^2}\left(1 - \frac{a^4\mathcal{L}_1(a)^2}{\xi_1(1+a^2\mathcal{L}_1(a))^2}\right)^{-1}.$$

Combining equation (3.7) and (3.8), we thus obtain an estimate for the excess risk of HPS. We provide some intuition behind both estimates. First, the variance estimate is a special case of Theorem 3.1 where $\Sigma^{(1)} = \Sigma^{(2)}$ but $n_1 \neq n_2$. In fact, the result can be obtained by solving α_1, α_2 in equation (3.2) using $\lambda_i = 1$ for all $i = 1, \ldots, p$. Second, the bias estimate is obtained based on a sharp convergence estimate proved in Bao et al. (2017a,b) for the free addition of two probability measures. To illustrate, denote by $\mathbf{v}(a) := (\Sigma^{(1)})^{1/2}\left(a\beta^{(1)} - a^2\beta^{(2)}\right)$. We can write the bias equation as

$$L_{\mathrm{Bias}}(a) = \mathbf{v}_a^\top (Z^{(1)})^\top Z^{(1)}\left(a^2(Z^{(1)})^\top Z^{(1)} + (Z^{(2)})^\top Z^{(2)}\right)^{-2}(Z^{(1)})^\top Z^{(1)}\,\mathbf{v}_a,$$

By the rotational invariance of $(Z^{(1)})^\top Z^{(1)}$ and $(Z^{(2)})^\top Z^{(2)}$, we have that

$$L_{\mathrm{Bias}}(a) \approx \|\,\mathbf{v}_a\,\|^2 \cdot \frac{1}{p}\,\mathrm{Tr}\left[\left((Z^{(1)})^\top Z^{(1)}\right)^2\left(a^2(Z^{(1)})^\top Z^{(1)} + (Z^{(2)})^\top Z^{(2)}\right)^{-2}\right] \tag{3.9}$$

up to a small error.

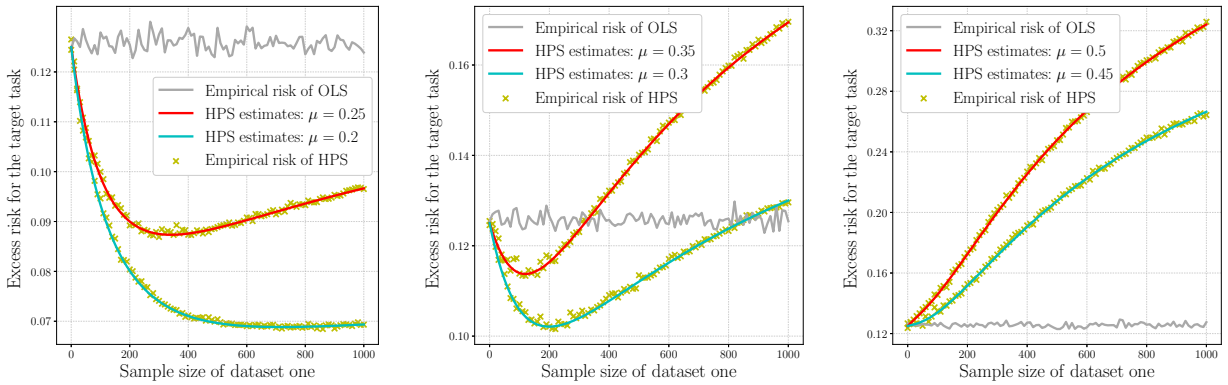
Our key idea is to write equation (3.9) as the derivative of a certain polynomial with respect to x at $x = 0$:

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

It is well-known that the empirical spectral distributions (ESD) of $(Z^{(i)})^\top Z^{(i)}$, $i = 1, 2$, satisfy the famous Marchenko-Pastur (MP) law asymptotically (Marčenko and Pastur, 1967). 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 rotational invariance of multivariate Gaussian distributions, 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 $(Z^{(2)})^\top Z^{(2)}$(Nica and Speicher, 2006). Finally, We use the result of Bao et al. (2017a,b) to obtain a sharp estimate of

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

which implies a sharp estimate of $L_{\mathrm{Bias}}(a)$. ≪Todo notes: can we add a few sentences to describe any technical challenge here? right now it sounds like our result is a direct corollary, which is not good?≫ The complete proof of Theorem 3.5 can be found in Section **??**.



(a) Positive transfer (b) From positive to negative transfer (c) Negative transfer

Figure 3: We illustrate three regimes of information transfer in HPS under model shift. When μ is small, HPS outperforms the OLS estimator irrespective of n_1. When μ is large, HPS performs worse than OLS. For an intermediate range of μ, HPS outperforms OLS only for a restricted range of n_1. See Claim 3.6 for the precise range of each regime. This simulation fixes $p = 100, n_2 = 300, \sigma = 1/2$ while varying n_1, μ.

Remarks. First, we believe the convergence rates p^c/n_1 and $p^{-1/2+c}$ in (3.7) and (3.8) are both sharp up to the p^c factor. ≪Todo notes: the rate in (3.7) seems different from Thm 3.1?≫ Second, we believe that the above argument can be extended to the case without the Gaussian assumption. For example, instead of using the results in (Bao et al., 2017a,b) on free addition, we can use the sharp local laws on polynomials of random matrices in (Erdős et al., 2020). However, applying the result of Erdős et al. (2020) requires checking certain technical regularity conditions for our setting. We leave this as an open question for future work.

Illustrative examples. While the estimates from Theorem 3.5 are generally complex, we describe a simplified result for the random-effect model (cf. Section 2.1). Under the assumptions stated in Proposition 2.4, $\hat{a}$ is approximately equal to one. Then, we can calculate that

$$\mathcal{L}_1(1) = \frac{p}{n_1 + n_2 - p} \text{ and } \mathcal{L}_2(1) = \frac{n_1^2(n_1 + n_2 - p) + pn_1n_2}{(n_1 + n_2)^2(n_1 + n_2 - p)}, \text{ in which } \kappa(1) = \frac{n_2^2}{(n_1 + n_2)^2 - n_1p}.$$

Thus, using that $\|\beta^{(1)} - \beta^{(2)}\|^2 = (2 + \mathrm{o}(1))\mu^2$ w.h.p., we conclude that for the random-effect model, the excess risk $L(\hat{\beta}_2^{\mathrm{HPS}})$ is approximately equal to

$$g(n_1) := \sigma^2\mathcal{L}_1(1) + 2\mu^2\mathcal{L}_2(1) = \frac{p\sigma^2}{n_1 + n_2 - p} + 2\mu^2 \cdot \frac{n_1^2(n_1 + n_2 - p) + pn_1n_2}{(n_1 + n_2)^2(n_1 + n_2 - p)}$$

plus lower order terms (that vanishes to zero as p goes to infinity). Next, we analyze when combining task one using HPS transfers positively to task two, depending on sample sizes n_1, n_2, and the model shift parameter μ in the random-effect model.

Proposition 3.6 (Sample sizes vs. model shift)**.** *Suppose Assumption 2.1 and the random-effect model under equation (2.5) and (2.11) holds. Suppose further that $n_2 \geqslant 3p$. Then, there exists a large constant $C > 0$ such that the following holds w.h.p.,*

i) If $\mu^2 \leqslant \frac{\sigma^2 p}{2(n_2-p)}$, then $L(\hat{\beta}_2^{\mathrm{HPS}}) \leqslant L(\hat{\beta}_2^{\mathrm{OLS}}) + \mathrm{O}(p^{-C})$.

ii) If $\frac{\sigma^2 p}{2(n_2-p)} < \mu^2 < \frac{\sigma^2 n_2}{2(n_2-p)}$, then there exists a deterministic constant ρ such that if $n_1 \leqslant \rho \cdot p$, then $L(\hat{\beta}_2^{\mathrm{HPS}}) \leqslant L(\hat{\beta}_2^{\mathrm{OLS}}) + \mathrm{O}(p^{-C})$, else $L(\hat{\beta}_2^{\mathrm{OLS}}) \leqslant L(\hat{\beta}_2^{\mathrm{HPS}}) + \mathrm{O}(p^{-C})$.

iii) If $\frac{\sigma^2 n_2}{2(n_2-p)} \leqslant \mu^2$, then $L(\hat{\beta}_2^{\mathrm{OLS}}) \leqslant L(\hat{\beta}_2^{\mathrm{HPS}}) + \mathrm{O}(p^{-C})$.

Proof. By Theorem 3.5 and the discussion above, we have that $L(\hat{\beta}_2^{\mathrm{HPS}}) = g(n) + \mathrm{O}(p^{-C})$ w.h.p. By Lemma 2.5, we have that the excess risk of the OLS estimator for task two satisfies

$$L(\hat{\beta}_2^{\mathrm{OLS}}) = \sigma^2 \cdot \mathrm{Tr}\left[\Sigma^{(2)}\left((X^{(2)})^\top X^{(2)}\right)^{-1}\right] = \frac{\sigma^2 p}{n_2 - p} + \mathrm{O}(p^{-C}).$$

Thus, whether or not $L(\hat{\beta}_2^{\mathrm{HPS}}) \leqslant L(\hat{\beta}_2^{\mathrm{OLS}})$ reduces to comparing $g(n_1)$ and $\frac{\sigma^2 p}{n_2 - p}$—let $h(n_1)$ be their difference. We can write $h(n_1)$ as

$$h(n_1) = 2\mu^2 \cdot \frac{n_1^2(n_1 + n_2 - p) + p n_1 n_2}{(n_1 + n_2)^2(n_1 + n_2 - p)} - \frac{\sigma^2 p n_1}{(n_1 + n_2 - p)(n_2 - p)}.$$

We observe that the sign of $h(n_1)$ is the same as the sign of the following simplified polynomial

$$\begin{aligned}\tilde{h}(n_1) &= 2\mu^2(n_2 - p)(n_1(n_1 + n_2 - p) + p n_2) - \sigma^2(n_1 + n_2)^2 \\ &= \left(2\mu^2(n_2 - p) - \sigma^2\right)n_1^2 + \left(2\mu^2(n_2 - p)^2 - 2\sigma^2 p n_2\right)n_1 + \left(2\mu^2(n_2 - p)p n_2 - \sigma^2 p n_2^2\right).\end{aligned}$$

Let C_0, C_1, C_2 be the coefficient of n_1^0, n_1^1, n_1^2 above, respectively. We argue about each claim as follows.

For claim i), if $\mu^2 \leqslant \frac{\sigma^2 p}{2(n_2 - p)}$, then C_0, C_1, C_2 are all at most zero. Thus, $\tilde{h}(n_1) \leqslant 0$, which implies that $h(n_1) \leqslant 0$.

For claim ii), if $\frac{\sigma^2 p}{2(n_2 - p)} < \mu^2 < \frac{\sigma^2 n_2}{2(n_2 - p)}$ (recall that $n_2 > p$ by Assumption 2.1), then $C_2 > 0$ and $C_0 < 0$. Thus, $\tilde{h}(n_1)$ has a positive root and a negative root. Let ρ be the positive root. Hence, if $n_1 \leqslant \rho \cdot p$, then $\tilde{h}(n_1) \leqslant 0$. Otherwise, $\tilde{h}(n_1) \geqslant 0$.

For claim iii), if $\frac{\sigma^2 n_2}{2(n_2 - p)} \leqslant \mu^2$, then $C_1 \geqslant 0$ and $C_2 \geqslant 0$. Furthermore, $\frac{\sigma^2 p n_2}{(n_2 - p)^2} \leqslant \frac{\sigma^2 n_2}{2(n_2 - p)}$ because $n_2 \geqslant 3p$ by our assumption. Thus, C_2 is non-negative as well, which implies $\tilde{h}(n_1) \geqslant 0$. □

As a remark, when $p < n_2 < 3p$, similar results can be derived using the above arguments (details omitted). Figure 3 illustrates Claim 3.6 for different regimes of model shift parameter μ in the random-effect model. This simulation affirms that our estimates accurately match with the empirical bias equation (2.8) plus the variance equation (2.9). Furthermore, we observe the three information transfer regimes shown in Claim 3.6 by varying μ in the random-effect model.

3.3 Covariate and model shifts

≪Todo notes: write down the formal statement≫

Consider the special case with $\Sigma^{(2)} = \mathrm{Id}$, $a = 1$ and the random effect model. Let $\sigma_i^{(1)}$ be the eigenvalues of $\Sigma^{(1)}$. Define the function

$$g_0(x) = \frac{1}{n_1}\sum_{i=1}^{p} \frac{\sigma_i^{(1)}}{1 + \sigma_i^{(1)} x} - \frac{1}{x}. \tag{3.10}$$

Let y be the unique positive solution to the equation

$$\frac{n_1 + n_2 - p}{n_1} + g_0(y) \cdot (1 + y) = 0.$$

Define the following three functions of y:

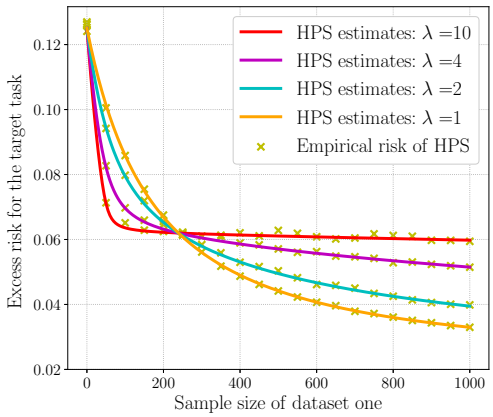
$$f_1 = \frac{n_1}{p} y + \frac{n_1 - p}{p}\frac{1}{g_0(y)}, \quad f_2 = \frac{n_1}{p}\frac{1}{g_0'(y)} - \frac{n_1 - p}{p}\frac{1}{[g_0(y)]^2}, \quad f_3 = -g_0(y),$$

where $g_0'(y)$ is the derivative

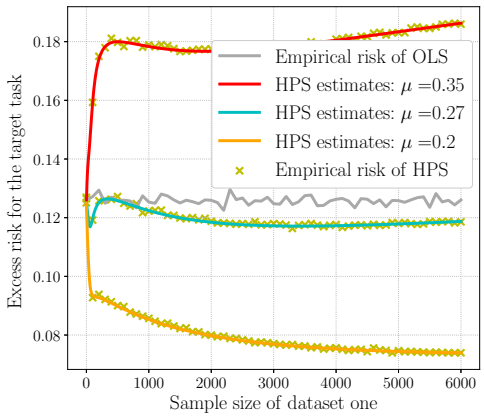
$$g_0'(y) = -\frac{1}{n_1}\sum_{i=1}^{p} \frac{(\sigma_i^{(1)})^2}{(1 + \sigma_i^{(1)} y)^2} + \frac{1}{y^2}.$$

Then the variance limit is given by

$$L_{\mathrm{Var}}(1) = \sigma^2 \frac{p}{n_1} f_1,$$



(a) Fixed μ, different λ



(b) Fixed λ, different μ

Figure 4: We further show that our observations in Figures 2 and 3 extend to settings under covariate and model shifts. Both simulation use $p = 100, n_2 = 300, \sigma = 1/2$. Figure 4a fixes $\mu = 0.1$ while varying λ, n_1. Figure 4b fixes $\lambda = 4$ while varying μ, n_1.

and the bias limit is given by

$$L_{\text{Bias}}(1) = 2\mu^2 \frac{1 - 2f_1 f_3 + f_2 f_3^2}{1 - \frac{p}{n_2} f_2 f_3^2}.$$

(Here $2\mu^2$ represents $\|\beta^{(1)} - \beta^{(2)}\|^2$.)

For the bias limit, we have the following proposition.

Theorem 3.7 (Bias estimates under covariate and model shifts)**.** *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 $L_{\text{Bias}}(a)$ in (2.8):*

$$\begin{aligned}
&\left| L_{\text{Bias}}(a) - (\beta^{(1)} - a\beta^{(2)})^\top (\Sigma^{(1)})^{1/2} \Pi(a) (\Sigma^{(1)})^{1/2} (\beta^{(1)} - a\beta^{(2)}) \right| \\
&\leqslant \left[\left(1 + \sqrt{\frac{p}{n_1}}\right)^4 - 1 + \mathrm{O}\left(n_1^{-1/2+2/\varphi+c}\right) \right] \frac{n_1^2 \lambda_1^2 \left\| (\Sigma^{(1)})^{1/2} \left(\beta^{(1)} - a\beta^{(2)}\right) \right\|^2}{[(\sqrt{n_1} - \sqrt{p})^2 \lambda_p^2 + (\sqrt{n_2} - \sqrt{p})^2]^2} \\
&+ p^{-C} \left[\|\beta^{(1)}\|^2 + \|\beta^{(2)}\|^2 \right],
\end{aligned} \tag{3.11}$$

uniformly in all $a \in \mathbb{R}$. Here λ_1 and λ_p are respectively the largest and smallest singular values of $M(a)$, $\Pi(a)$ is a $p \times p$ matrix defined as

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

and (a_3, a_4) is the solution of the following system of equations

$$\begin{aligned}
a_3 + a_4 &= \frac{1}{n_1 + n_2} \sum_{i=1}^{p} \frac{1}{\lambda_i^2 a_1 + a_2}, \\
a_3 + \frac{1}{n_1 + n_2} \sum_{i=1}^{p} \frac{\lambda_i^2 (a_2 a_3 - a_1 a_4)}{(\lambda_i^2 a_1 + a_2)^2} &= \frac{1}{n_1 + n_2} \sum_{i=1}^{p} \frac{\lambda_i^2 a_1}{(\lambda_i^2 a_1 + a_2)^2},
\end{aligned} \tag{3.12}$$

where we recall that (a_1, a_2) is the solution of (3.2).

≪Todo notes: consider changing a_3, a_4 to α_3, α_4 because of conflict with a≫

Note that the first error term on the right-hand side of (3.11) 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 (3.11) only gives an exact asymptotic limit in the regime $n_1 \gg p$. Moreover, by equations (3.2) and (3.12) 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

$$\begin{aligned} &(\beta^{(1)} - a\beta^{(2)})^\top (\Sigma^{(1)})^{1/2} \Pi(a) (\Sigma^{(1)})^{1/2} (\beta^{(1)} - a\beta^{(2)}) \\ &= \left\| (\Sigma^{(2)})^{1/2} \frac{1}{a^2\Sigma^{(1)} + \Sigma^{(2)}} a\Sigma^{(1)} \left(\beta^{(1)} - a\beta^{(2)}\right) \right\|^2 + \mathrm{O}\left(\frac{p \left\|\beta^{(1)} - a\beta^{(2)}\right\|^2}{n_1+n_2}\right). \end{aligned} \tag{3.13}$$

Hence (3.11) 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 $L_{\mathrm{Bias}}(a)$ using the law of large numbers in the regime $n_1 \gg p$. However, simulations show that our estimate (3.11) is more precise than the first term on the right-hand side of (3.13).

Remark 3.8. The main error in Proposition 3.7 comes from approximating $(Z^{(1)})^\top Z^{(1)}$ by $n_1 \mathrm{Id}_{n_1 \times n_2}$ using Corollary **??** in the supplement (Yang et al., 2020). 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 (Zheng et al., 2017). 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 3.2 for the bias term also fails, because we cannot reduce the problem into the free addition of two random matrices that are asymptotically freely independent.

4 Extension to multiple sources

Our setup and the results in Section 3 are both for transferring from one data source. This section extends our setup to transfer learning from multiple data sources. We focus on a natural setting where all the tasks have the same covariates but different labels.

Data model. 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

$$Y^{(i)} = X\beta^{(i)} + \varepsilon^{(i)}, \text{ for } i = 1, 2, \cdots, t. \tag{4.1}$$

Similar to Section 3, we use the first $(t-1)$ datasets as sources to predict the t-th task. However, there is a model shift between the data sources and the task we would like to predict. We make several standard assumptions on X and each of $\varepsilon^{(i)}$. First, $X = Z\Sigma^{1/2}$ is a random matrix satisfying Assumption 2.1 (same as $X^{(2)}$). In particular, the sample size n is greater than the dimension p. Second, every $\varepsilon^{(i)} \in \mathbb{R}^n$ is a random vector with i.i.d entries of mean zero, variance σ^2, and bounded moments up to any order (cf. equation (2.3)). Finally, $\beta^{(i)} \in \mathbb{R}^p$ is a fixed vector independent from any other $\beta^{(j)}$ for $j \neq i$, the matrix X, and $\varepsilon^{(j)}$ for any $j = 1, \ldots, t$. As a remark, the above data model is similar in spirit to the setting of Lounici et al. (2011). While that work assumes all of the model vectors $\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(t)}$ share the same subset of non-zero coordinates, our result below will apply even when these vectors do not satisfy such conditions.

Estimator. We combine multiple data sources by extending the two-layer linear neural network from equation (1.2) as follows:

$$f(A, B) = \sum_{j=1}^{t} \left\| XBA_j - Y^{(j)} \right\|^2, \tag{4.2}$$

where $A = [A_1, A_2, \ldots, A_t] \in \mathbb{R}^{r \times t}$ denotes the output layer and $B \in \mathbb{R}^{p \times r}$ denotes the (shared) feature layer. We set the width of the feature layer r less than the number of tasks t. Otherwise, when $r \geqslant t$, the global minimum of $f(A, B)$ reduces to single-task learning, similar to Proposition 2.2 (details omitted).

Let $(\hat{A}, \hat{B})$ denote a global minimizer of $f(A, B)$. We define the HPS estimator for task i as $\hat{\beta}_i^{\mathrm{HPS}} := \hat{B}\hat{A}_i$, where $\hat{A}_i$ denotes the i-th column of $\hat{A}$. We evaluate the performance of $\hat{\beta}_i^{\mathrm{HPS}}$ according to its excess risk:

$$L_i(\hat{\beta}_i^{\mathrm{HPS}}) = \left\| \Sigma^{1/2} \left(\hat{\beta}_i^{\mathrm{HPS}} - \beta^{(i)} \right) \right\|^2. \tag{4.3}$$

Result. We show that in the multi-task setting, hard parameter sharing finds the "best" rank-r approximation to all tasks. To formally describe our result, we introduce several notations. Let $B^\star := [\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(t)}] \in \mathbb{R}^{p \times t}$ be the concatenated model vectors of all tasks. Let $A^\star A^{\star\top}$ be the best approximation of ${B^\star}^\top \Sigma B^\star$ in the set of rank-r subspaces:

$$A^\star := \operatorname*{argmax}_{U \in \mathbb{R}^{t \times r} : U^\top U = \mathrm{Id}_{r \times r}} \langle UU^\top, {B^\star}^\top \Sigma B^\star \rangle, \tag{4.4}$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product between two matrices. Let $a_i^\star \in \mathbb{R}^r$ be the i-th column vector of $A^\star A^{\star\top}$. We show a precise estimate of $L_i(\hat{\beta}_i^{\mathrm{HPS}})$ based on $a_i^\star$ as follows.

Theorem 4.1 (Excess risk of HPS for multiple tasks under model shift)**.** *Suppose the multi-task setting according to equation (4.1) holds. Let $r < t$ be a positive integer. Suppose the r-th largest eigenvalue of ${B^\star}^\top \Sigma B^\star$ is strictly larger than its $(r+1)$-th largest eigenvalue. Let $c > 0$ be an arbitrarily (small) constant. The following estimate of $L_i(\hat{\beta}_i^{\mathrm{HPS}})$ holds w.h.p., for any $i = 1, \ldots, t$:*

$$\left| L_i(\hat{\beta}_i^{\mathrm{HPS}}) - L_i(B^\star a_i^\star) - \frac{p\sigma^2}{n-p} \cdot \|a_i^\star\|^2 \right| \leqslant \sqrt{\|{B^\star}^\top \Sigma B^\star\|_2 \cdot n^{-\frac{1}{2}+\frac{2}{\varphi}+c} + \sigma^2 n^{-\frac{1}{2}+c}} \cdot \frac{\|{B^\star}^\top \Sigma B^\star\|_2 + \sigma^2}{\lambda_r - \lambda_{r+1}}. \tag{4.5}$$

Recall that $\varphi > 4$ according to Assumption 2.1, thus, $\frac{2}{\varphi} \leqslant \frac{1}{2}$ and $n^{-\frac{1}{2}+\frac{2}{\varphi}+c}$ vanishes to zero for a small enough constant c. Thus, $L_i(B^\star a_i^\star) + \frac{p\sigma^2}{n-p}\|a_i^\star\|^2$ is the limit of $L_i(\hat{\beta}_i^{\mathrm{HPS}})$ as p approaches infinity—$L_i(B^\star a_i^\star)$ is the limiting bias of $\hat{\beta}_i^{\mathrm{HPS}}$ and $\frac{p\sigma^2}{n-p}\|a_i^\star\|^2$ is its limiting variance. The proof of Theorem 4.1, which characterizes the global minimum of problem (4.2) through a connection to PCA, can be found in Section **??**.

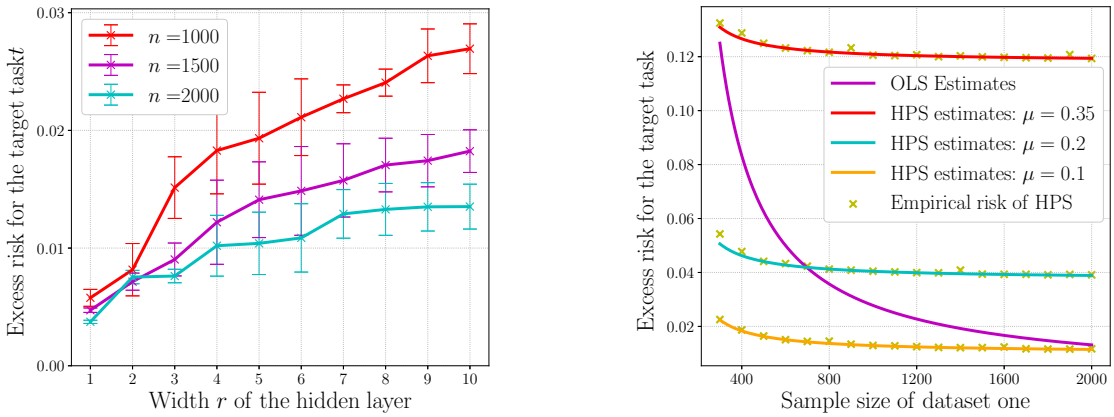
Illustrative examples. We give an example of our estimate in the random-effect model. In this setting, the model vector of every task is the sum of a shared vector β_0 plus a task-specific component $\widetilde{\beta}^{(i)}$:

$$\beta^{(i)} = \beta_0 + \widetilde{\beta}^{(i)}, \text{ for every } i = 1, 2, \cdots, t, \tag{4.6}$$

where every entry of $\widetilde{\beta}^{(i)}$ is drawn i.i.d. from a Gaussian distribution with zero mean and μ^2/p variance. Thus, μ is a parameter for modeling the difference between different tasks. We study two fundamental questions in this setting. First, what should the width (r) of the feature layer B be? Second, when does HPS transfer positively to a particular task, depending the sample size n and the (model shift) parameter μ? Since all tasks are symmetric in this setting, we analyze the average (limiting) risk

$$g_r(n, \mu) = \frac{1}{t} \sum_{i=1}^{t} \left(L_i(B^\star a_i^\star) + \frac{p\sigma^2}{n-p} \cdot \|a_i^\star\|^2 \right).$$

We address both questions by characterizing $g_r(n, \mu)$, which accurately approximates $\frac{1}{t}\sum_{i=1}^{t} L_i(\hat{\beta}_i^{\mathrm{HPS}})$ according to Theorem 4.1. Recall from Lemma 2.4 that the excess risk of the OLS estimator satisfies that $L_i(\hat{\beta}_i^{\mathrm{OLS}}) = \frac{\sigma^2 p}{n-p} + o_p(1)$, in which $o_p(1)$ vanishes to zero as p approaches infinity. We compare $g_r(n, \mu)$ with the OLS estimate in the following claim.



(a) Varying width

(b) Varying sample size and model shift parameter

Figure 5: We illustrate the performance of HPS in the random-effect model with $p = 100, t = 10, \sigma = 1/2$. Figure 5a shows that from 1 to $t = 10$, setting width $r = 1$ gives the lowest excess risk for task t. The result is averaged over three random seeds. This simulation uses $\mu = 0.05$. Figure 5b fixes $r = 1$ while varying μ and n. We observe that depending on μ and n, HPS may provide a positive transfer or a negative transfer to task t (similar to Figure 3). The precise condition for determining the transfer effect can be found in Claim 4.2.

Claim 4.2 (Choosing r for HPS in the random-effect model)**.** *Suppose the multi-task setting with random-effect linear models under equation (4.1) and (4.6) holds. Let $\varepsilon = o_p(\|\beta_0\|^2 + \mu^2)$ be a small error term that decreases to zero as p goes to infinity. Let $r \in [1, r]$. W.h.p. over the randomness of $B^\star$, the following holds:*

i) If $\mu^2 \geqslant \frac{\sigma^2 p^2}{(n-p)\,\mathrm{Tr}[\Sigma]}$, then $g_r(n,\mu) \geqslant \frac{\sigma^2 p}{n-p} + \varepsilon$, for any $1 \leqslant r < t$.

i) If $\mu^2 < \frac{\sigma^2 p^2}{(n-p)\,\mathrm{Tr}[\Sigma]} - \frac{tp}{\mathrm{Tr}[\Sigma]}\varepsilon$, then $g_r(n,\mu)$ is minimized when $r = 1$. Additionally, $g_1(n,\mu) \leqslant \frac{\sigma^2 p}{n-p} + \varepsilon$.

Proof. We first simplify the expression of $g_r(n,\mu)$ using its definition

$$\begin{aligned} g_r(n,\mu) &= \frac{1}{t}\sum_{i=1}^{t}\left(L_i(B^\star a_i^\star) + \frac{p\sigma^2}{n-p}\cdot\|a_i^\star\|^2\right) \\ &= \left(\frac{1}{t}\sum_{i=1}^{t}\left\|\Sigma^{1/2}(B^\star a_i^\star - \beta^{(i)})\right\|^2\right) + \left(\frac{1}{t}\sum_{i=1}^{t}\frac{p\sigma^2}{n-p}\cdot\|a_i^\star\|^2\right) \\ &= \frac{1}{t}\left\|\Sigma^{1/2}(B^\star A^\star - B^\star)\right\|_F^2 + \frac{r}{t}\cdot\frac{p\sigma^2}{n-p}. \end{aligned}$$

In the last step, we use the matrix notation to write the sum for the first part. We use the fact that $\sum_{i=1}^{t}\|a_i^\star\|^2 = r$ because $A^\star = [a_1^\star, \ldots, a_t^\star]$ satisfies that ${A^\star}^\top A^\star = \mathrm{Id}_{p\times p}$ following its definition in equation (4.4). Using this condition on $A^\star$, we further get that

$$\left\|\Sigma^{1/2}(B^\star A^\star - B^\star)\right\|_F^2 = \mathrm{Tr}\left[{B^\star}^\top \Sigma B^\star(\mathrm{Id}_{p\times p} - A^\star {A^\star}^\top)\right],$$

which is precisely equal to the smallest $(t-r)$ eigenvalues of ${B^\star}^\top \Sigma B^\star$. Using the concentration of Gaussian random vectors in Lemma **??**, w.h.p. the (i,j)-th entry of ${B^\star}^\top \Sigma B^\star$ is equal to

$$\beta^{(i)\top}\Sigma\beta^{(j)} = \beta_0^\top\Sigma\beta_0 + \delta_{i,j}\frac{\mu^2}{p}\,\mathrm{Tr}\left[\Sigma\right] + \mathrm{O}\left(p^{-1/2+c}\|\beta_0\|^2 + p^{-1/2+c}\mu^2\right) \tag{4.7}$$

for any constant $c > 0$, where $\delta_{i,j} = 1$ if and only if $i = j$. Thus, ${B^\star}^\top \Sigma B^\star$ is equal to a rank-1 matrix with (spectral) norm $t \cdot \beta_0^\top\Sigma\beta_0$ plus Gaussian perturbation. Using the concentration result of Lemma **??** ≪Todo

notes: check≫, we have that

$$\lambda_1 = \left(1 + \mathrm{O}(p^{-1/2+c})\right) \cdot \left(t \cdot \beta_0^\top \Sigma \beta_0 + \frac{\mu^2}{p} \operatorname{Tr}[\Sigma]\right), \text{ and}$$
$$\lambda_i = \left(1 + \mathrm{O}(p^{-1/2+c})\right) \cdot \frac{\mu^2}{p} \operatorname{Tr}[\Sigma], \text{ for } i = 2, \cdots, t.$$

Thus, w.h.p. the sum of smallest $(t-r)$ eigenvalues of ${B^\star}^\top \Sigma B^\star$ is equal to $(1 + \mathrm{O}(p^{-1/2+c})) \cdot (t-r)\frac{\mu^2 \operatorname{Tr}[\Sigma]}{p}$. Let $\varepsilon = o_p(1)$ denote the error term. We conclude that w.h.p.,

$$\begin{aligned} g_r(n, \mu) - \frac{\sigma^2 p}{n-p} &= \left(1 - \frac{r}{t}\right) \frac{\mu^2}{p} \operatorname{Tr}[\Sigma] + \frac{r}{t} \cdot \frac{p\sigma^2}{n-p} + \varepsilon - \frac{\sigma^2 p}{n-p} \\ &= \left(1 - \frac{r}{t}\right) \cdot \left(\frac{\mu^2 \operatorname{Tr}[\Sigma]}{p} - \frac{p\sigma^2}{n-p}\right) + \varepsilon. \end{aligned} \tag{4.8}$$

Now we are ready to finish the proof. For claim i), if $\mu^2 \geqslant \frac{\sigma^2 p^2}{(n-p)\operatorname{Tr}[\Sigma]}$, the coefficient of $(1 - \frac{r}{t})$ in equation (4.8) is non-negative, and claim i) follows. For claim ii), if $\mu^2 < \frac{\sigma^2 p^2}{(n-p)\operatorname{Tr}[\Sigma]} - \frac{pt}{\operatorname{Tr}[\Sigma]}\varepsilon$, the coefficient of $(1 - \frac{r}{t})$ in equation (4.8) is negative. Furthermore, equation (4.8) is minimized when $r = 1$, and claim ii) follows since the RHS of equation (4.8) is at most ε. □

Figure 5 validates the result of Claim 4.2 in finite dimensions. First, we see that the empirical excess risk (measured on a particular task) is indeed smallest when the width r is equal to 1. Second, we see that HPS provides a positive transfer (to a particular task) depending on the sample size n and the model shift parameter μ. The results under different μ values match with the condition in Claim 4.2.

5 Empirical results

This section complements our theoretical analysis of HPS with empirical evaluations. First, we evaluate the performance of HPS estimators against several natural transfer learning estimators. We show that HPS achieves superior performance under various settings of covariate and model shifts. Second, we provide empirical implications that are inspired by our theoretical insights for mitigating covariate and model shifts. One concerns a certain covariate alignment procedure for mitigating covariate shift. We show that such alignment procedures provide greater gains for larger regimes of n_1/n_2. The other is a progressive training procedure for mitigating model shift. We show that progressively increasing the sample sizes of data sources significantly reduces the computational cost of learning HPS neural networks while achieving the same accuracy for predicting the target task.

5.1 Evaluating HPS under distribution shift

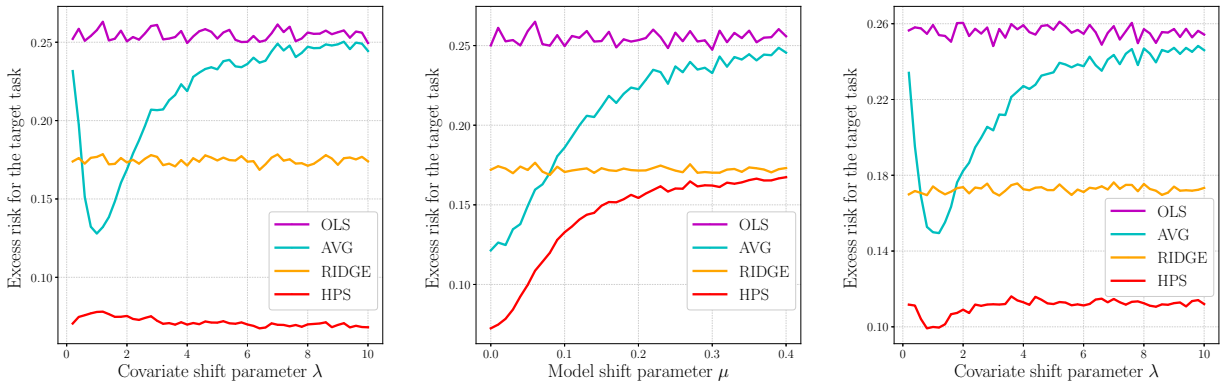
We show that HPS estimators enjoy superior empirical performance compared to several natural transfer learning estimators. We consider the following estimators for this comparative study:

i) STL estimator (OLS): $\hat{\beta}_2^{\mathrm{OLS}} = \left((X^{(2)})^\top X^{(2)}\right)^{-1} (X^{(2)})^\top Y^{(2)}$.

ii) Averaging estimator (AVG): given two tasks, take a convex combination of their OLS estimators $b \cdot \hat{\beta}_1^{\mathrm{OLS}} + (1-b) \cdot \hat{\beta}_2^{\mathrm{OLS}}$.

iii) Ridge estimator (RIDGE): $\hat{\beta}_2^{\mathrm{RIDGE}} = \left((X^{(2)})^\top X^{(2)} + \lambda \cdot \mathrm{Id}_{p\times p}\right)^{-1} (X^{(2)})^\top Y^{(2)}$.

The parameter b in the averaging estimator is optimized using a validation set of the same size as the training set. Additionally, we extend the two-layer neural network formulation of HPS (cf. equation (1.2)) to a weighted and regularized objective: $b \cdot \left\|X^{(1)} B A_1 - Y^{(1)}\right\|^2 + (1-b) \cdot \left\|X^2 B A_2 - Y^{(2)}\right\|^2 + \frac{\lambda}{2} \cdot \|B\|^2$, and optimize b, λ using the same validation set.



(a) Covariate shift (b) Model shift (c) Covariate and model shifts

Figure 6: We compare the empirical excess risk of HPS and several natural transfer learning estimators. We find that HPS achieves the lowest excess risk for predicting task two under covariate and model shifts. This simulation uses $p = 50, n_1 = n_2 = 100, \sigma = 1/2$. Figure 6c uses $\lambda = 4$ for generating covariate-shifted features.

Figure 6 shows the result under various settings of covariate shift and model shift. Figure 6a uses the same data generating process as Figure 2. Figure 6b uses the same data generating process as Figure 3. Figure 6c combines both by generating covariate-shifted features and different linear models for each dataset. For reach run, we average the result over 100 random seeds because of high variance from the small sample sizes. We find that HPS outperforms all the other estimators in this simulation (the result for the TL estimator is worse than STL and is omitted from the figure).

5.2 Implications for neural networks

We conduct further studies of HPS for text classification tasks. We consider six datasets for predicting movie review sentiment (MR and SST), sentence subjectivity (SUBJ), customer review sentiment (CR), question types (TREC), and polarity of a phrase (MPQA).[2] We learn a hard parameter sharing model that consists of a word embedding layer using GloVe (Pennington et al., 2014), followed by a shared feature representation layer.[3] A separate output layer is used for predicting the labels of each dataset. We evaluate on three possibilities of feature representation layers including LSTM, MLP, and CNN, all of which are implemented using PyTorch.[4] See Figure 1 for an illustration of the network architecture.

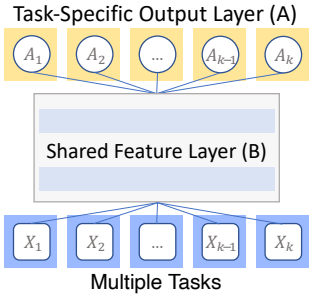


Figure 7: A hard parameter sharing neural network.

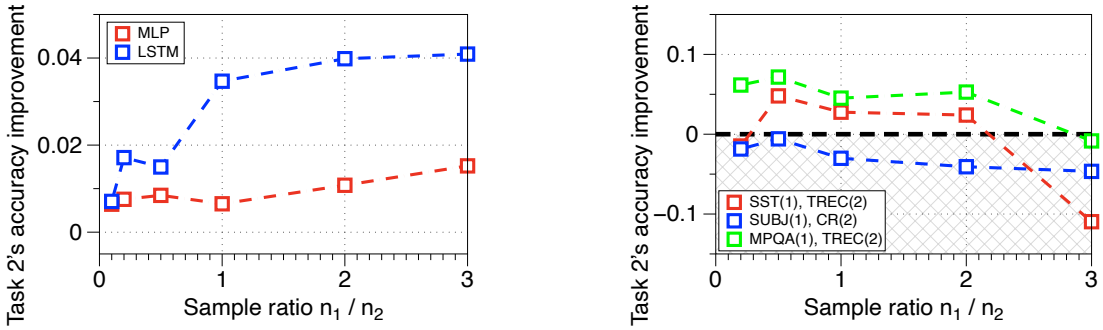
Covariate alignment for mitigating covariate shift. We apply our theoretical insights to conduct detailed analysis of a covariate alignment procedure proposed by Wu et al. (2020). The idea of this procedure is to insert an "alignment" module (or matrix) between every input X_i and the shared module B. During training, the entire network is optimized together with this alignment module (see Wu et al. (2020) for more details about the implementation).

Our hypothesis is that as n_1/n_2 increases, performing covariate alignment leads to larger accuracy improvement over the baseline HPS. Recall from Claim 3.4 that covariate shift between the data source and the target task worsens the performance of HPS. The effect is further exacerbated when the sample size of the data source is larger than the target task. To verify the hypothesis, we conduct multi-task training over all 15 pairwise tasks (among the six datasets). We measure the average accuracy improvement from

[2]The datasets can be downloaded at https://github.com/harvardnlp/sent-conv-torch/tree/master/data. Further statistics of each dataset can also be found following the link.

[3]The GloVe word vectors can be downloaded at https://nlp.stanford.edu/projects/glove/.

[4]The PyTorch neural network modules are available at https://pytorch.org/docs/stable/nn.html.



(a) Covariate alignment under varying sample sizes

(b) Information transfer under varying sample sizes

Figure 8: We extend some of our theoretical observations to text classification tasks using HPS neural networks. Figure 8a shows that covariance alignment improves the task two's test accuracy more for larger sample size ratio n_1/n_2. Figure 8b shows that task one provides a positive transfer to task two in HPS only for a restricted range of n_1 (similar to Figure 3b). Both experiments fixes $n_2 = 1000$ while varying n_1 from 100 to 3000. The y-axis measures task two's test accuracy using HPS subtracted by its test accuracy without using task one's data.

performing covariance alignment vs. HPS (trained using stochastic gradient descent (SGD)) over the 15 task pairs. Figure 8a confirms our hypothesis. We observe that covariate alignment achieves up to 4% accuracy improvement as n_1/n_2 increases.

Progressive training for mitigating model shift. Inspired by our theoretical analysis, we propose a progressive training procedure that reduces the computational cost of learning HPS networks. To motivate this procedure, we first conduct an experiment similar to Figure 3b) but using the text classification datasets instead. In Figure 8b, we find that for multiple pairs of datasets, increasing n_1 improves task two's test accuracy initially, but hurts eventually.

These examples and the ones in Figure 3b suggest a natural progressive training procedure that increases n_1 progressively until performance drops. Concretely, we first divide the training data into S batches. Then, we progress in S stages during training. During each stage, we progressively add one more minibatch of data from the data source(s). During each stage, we run SGD for T epochs using only the available minibatches of data. We terminate once task two's validation accuracy drops compared to the previous round's result or reaches a desired threshold τ. See Algorithm 1 for the complete procedure. As an example, this procedure will terminate at the optimal value of n_1 if applied to the settings of Figures 3b and 8b. In light of these observations, our hypothesis is that this procedure requires a lower computational cost compared to SGD.

We evaluate Algorithm 1 in two scenarios. First, we consider transferring from single data source to a target task. We evaluate over all 15 two-task pairs. We find that when averaged over all the pairs, our procedure requires less than 35% of the computational cost relative to SGD while reaching the same level of test accuracy for predicting task two. Second, we consider transferring from multiple data sources. We use all six datasets as data sources to help predict every other task—that is, we measure the average accuracy of predicting all six tasks. We find that a similar progressive training style procedure requires less than 35% of the computational cost (relative to the cost of running SGD) for reaching the same average test accuracy of SGD.

Models	Relative cost
CNN	30%
LSTM	35%
MLP	31%

Table 1: The computational cost of running Algorithm 1 relative to the cost of running SGD.

Further details. Algorithm 1 uses SGD during the minimization step. We randomly shuffle the data of both tasks and apply SGD on the shuffled data. For the two-task experiment, we set τ to be task two's (STL) test accuracy trained without using any other task's data. For the six-task experiment, We set τ to be the average test accuracy of all six tasks trained using all tasks' data. We include the data from all tasks except SST. For SST, we progressively increasing its sample size similar to Algorithm 1.

Algorithm 1 Progressive training of hard parameter sharing networks

Input: Two tasks $(X^{(1)}, Y^{(1)})$ and $(X^{(2)}, Y^{(2)})$, divided into a training, validation, and test set.
Parameter: A shared module B, and output layers A_1, A_2.
Require: Number of batches S, epochs T, a threshold $\tau \in (0, 1)$.
Output: The trained modules B, A_2 optimized for task two.

```
Divide the training set of (X^(1), Y^(1)) randomly into S batches: (x^(1), y^(1)), ..., (x^(S), y^(S)).
for i = 1, ..., S do
    for j = 1, ..., T do
        Minimize the cross-entropy loss of B, A_1, A_2 on {(x^(k), y^(k))}_{k=1}^{i} and the training set of (X^(2), Y^(2))
    using SGD.
    end for
    Let a_i be the current validation accuracy for task two.
    if a_i < a_{i-1} or a_i > τ then
        break
    end if
end for
```

6 Conclusions

Distribution shift is a fundamental challenge in applying transfer learning. This work formulated a theoretical setup in which questions related to the transfer effect can be formally analyzed. The setup can reproduce several interesting phenomena in the context of transfer learning. Precise asympotics for the risk of a two-layer linear neural network are shown under various kinds of distribution shift. Such analysis has led to a number of theoretical insights and practical implications for applying transfer learning to text classification tasks.

It is a very interesting question to derive the asymptotic limit under general covariate and model shift. We remark that likely such a result will require studying the asymptotic distribution of the singular values of asymmetric matrices, which is technically challenging. Another interesting question is to study the impact of distribution shift in other settings such as logistic regression (Sur and Candès, 2019).

References

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

Z. D. Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*, volume 2 of *Mathematics Monograph Series*. Science Press, Beijing, 2006.

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

Zhigang Bao, László Erdős, and Kevin Schnelli. Local law of addition of random matrices on optimal scale. *Communications in Mathematical Physics*, 349(3):947–990, 2017a.

Zhigang Bao, László Erdős, and Kevin Schnelli. Convergence rate for spectral distribution of addition of random matrices. *Advances in Mathematics*, 319:251 – 291, 2017b.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

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.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32): 15849–15854, 2019.

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.

Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(8), 2008.

Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010.

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 and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

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

László Erdős, Torben Krüger, and Yuriy Nemish. Local laws for polynomials of Wigner matrices. *Journal of Functional Analysis*, 278(12):108507, 2020.

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

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.

Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, and Mahdi Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

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

Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

Wouter M Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.

Qi Lei, Wei Hu, and Jason Lee. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, pages 6164–6174. PMLR, 2021.

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.

Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

Tengyuan Liang, Alexander Rakhlin, et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

Karim Lounici, Massimiliano Pontil, Sara Van De Geer, and Alexandre B Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The annals of statistics*, 39(4):2164–2204, 2011.

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

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

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

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.

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.

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

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.

Antonia M Tulino and Sergio Verdú. *Random matrix theory and wireless communications*. Now Publishers Inc, 2004.

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, Hongyang R. Zhang, Sen Wu, Weijie J. Su, and Christopher Ré. Supplement to "sharp bias-variance tradeoffs of hard parameter sharing in high-dimensional linear regression". 2020.

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.