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

Anonymous Author
Anonymous Institution

Abstract

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

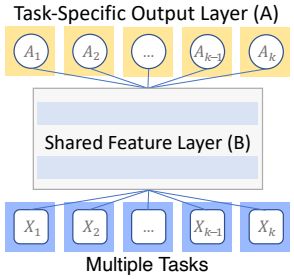
1 Introduction

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

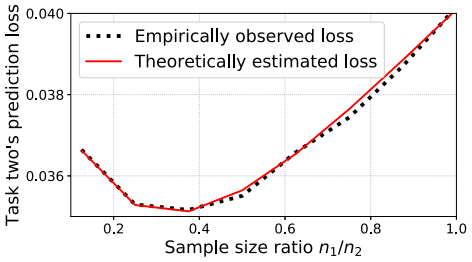
Hard parameter sharing works as an inductive transfer mechnism and a regularizer that reduces overfitting, both of which have great intuitive appeal (Ruder, 2017). For example, by restricting the shared space's size, HPS encourages information sharing among multiple tasks (Kumar and Daumé III, 2012).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

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

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



(a) A hard parameter sharing architecture



(b) Varying sample size ratio

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

1.1 Setup and Main Results

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

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

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

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

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

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

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

Covariate shift (Example 3.4): In addition to sample sizes, variance also scales with two datasets' covariate shifts. For a large sample size ratio, HPS's variance is smallest when there is no covariate shift. Counterintuitively, for a small sample size ratio, having covariate shifts reduces variance through a complementary spectrum. We achieve this result through a novel charac-

terization on the inverse of the sum of two sample covariance matrices with arbitrary covariate shifts. See our discussion of proof techniques below for details.

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

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

1.2 Related Work

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

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

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

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

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

2 A Bias-variance Decomposition for Multiple Tasks

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

Setting. Suppose we have t datasets whose sample sizes are all equal to n and whose features are all denoted by $X \in \mathbb{R}^{n \times p}$. The label vector of the i-th task follows a linear model $Y^{(i)} = X\beta^{(i)} + \varepsilon^{(i)}$. We assume: (i) $X = Z\Sigma^{1/2} \in \mathbb{R}^{n \times p}$ for a positive semidefinite

matrix $\Sigma \in \mathbb{R}^{p\times p}$, and every entry of $Z \in \mathbb{R}^{n\times p}$ is drawn independently from a one dimensional distribution with zero mean, unit variance, and constant φ-th moment for a fixed $\varphi > 4$.

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

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

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

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

Our first main result shows that hard parameter sharing essentially approximates all tasks through a rank-r subspace. To formalize this geometric intuition, we introduce the matrix $B^\star := [\beta_1, \beta_2, \ldots, \beta_t] \in \mathbb{R}^{p\times t}$ which contains all the linear model parameters. Let $A^\star A^{\star\top}$ denote the best rank-r subspace approximation of $B^{\star\top}\Sigma B^\star$ (which is task labels' "covariance").[2]

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

To see this, recall that r is one and $A^\star A^{\star\top}$ is the best rank-1 approximation of $B^{\star\top}\Sigma B^\star = B^{\star\top}B^\star$. Hence, the above expression is equal to the sum of $B^{\star\top}B^\star$'s bottom $r-1$ singular values. Based on the definition of the random-effect model, the (i,j)-th entry of $B^{\star\top}B^\star$ is equal to (ignoring lower order terms)

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

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

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

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

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

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

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

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

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

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

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

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

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

where $Y = [Y^{(1)}, Y^{(2)}, \ldots, Y^{(t)}]$. Here we have used that $X^\top X$ is invertible since $n > \rho\cdot p$ and $\rho > 1$ (cf. Fact E.3). Plugging $\hat{B}(A)$ into equation (2.2), we obtain the following objective that only depends on A (in matrix notation):

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

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

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

The previous section assumes that all tasks have the same sample size and feature vectors. In this section, we study how different sample sizes and different features impact hard parameter sharing. The setting where features differ across tasks is often called "covariate shift".

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

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

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

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

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

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

To derive the bias and variance equation, we consider the expected loss conditional on the covariates as follows (the empirical loss is close to this expectation as will be shown in equation (D.21)):

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

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

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

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

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

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

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

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

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

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

Applying Theorem 3.1 to the above setting, we get that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

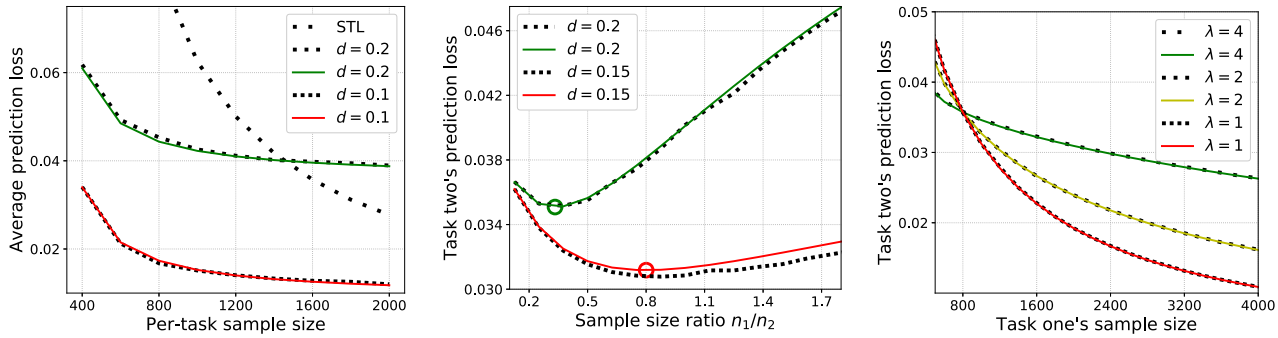
Our proof uses the Stieltjes transform or the resolvent method in random matrix theory. We briefly describe the key ideas and refer the interested readers to classical texts such as Bai and Silverstein (2010); Tao (2012); Erdos and Yau (2017). For any probability measure μ supported on $[0, \infty)$,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

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

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

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

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



(a) Example 2.4 (b) Example 3.2 (c) Example 3.4

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

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

4 Simulation Studies

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

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

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

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

5 Conclusions and Discussions

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

References

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

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

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

Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 2020.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.

Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.*, 19(33):1–53, 2014.

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

Xiucai Ding and Fan Yang. A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *Ann. Appl. Probab.*, 28(3):1679–1738, 2018.

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

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

L. Erdős, A. Knowles, and H.-T. Yau. Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré*, 14:1837–1926, 2013a.

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

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

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

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

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

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

Abhishek Kumar and Hal Daumé III. Learning task grouping and overlap in multi-task learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012.

Tao Lei, Yu Zhang, Sida I Wang, Hui Dai, and Yoav Artzi. Simple recurrent units for highly parallelizable recurrence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4470–4481, 2018.

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

Xin Li and Dan Roth. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019.

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

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884, 2016.

Alexandru Nica and Roland Speicher. *Lectures on the combinatorics of free probability*, volume 13. Cambridge University Press, 2006.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

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

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

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

Terence Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Soc., 2012.

Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Revisiting multi-task learning in the deep learning era. *arXiv preprint arXiv:2004.13379*, 2020.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

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

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

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018.

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

Shurong Zheng, Zhidong Bai, and Jianfeng Yao. CLT for eigenvalue statistics of large-dimensional general Fisher matrices with applications. *Bernoulli*, 23(2): 1130–1178, 2017.