Revisiting the Bias-Variance Tradeoff of Multi-Task Learning in High Dimensions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multi-task learning is a powerful approach in many applications such as image and text classification. Yet, there is little rigorous understanding of when multi-task learning outperforms single-task learning. In this work, we provide a rigorous study to anwer the question in the high-dimensional linear regression setting. We show that a bias-variance tradeoff of multi-task learning determines the effect of information transfer and develop new concentration bounds to analyze the tradeoff. Our key observation is that three properties of task data, namely *task similarity*, *sample size*, and *covariate shift* can affect transfer in the high-dimensional linear regression setting. We relate each property to the bias and variance of multi-task learning and explain three negative effects with decreased task similarity, increased source sample size, and covariate shift under increased source sample size. We validate the three effects on text classification tasks. Inspired by our theory, we show two practical connections of interest. First, single-task results can help understand when multi-task learning gives gains. Second, incrementally adding training data can mitigate negative transfer and improve multi-task training efficiency.

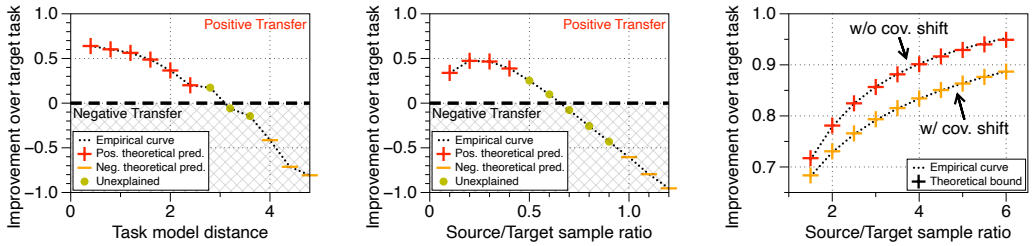
1 Introduction

Multi-task learning is a powerful approach to improve performance for many tasks in computer vision [1, 2], natural language processing [3, 4], and other areas [5]. In many settings, multiple source tasks are available to help predict a particular target task. The performance of multi-task learning depends on the relationship between the source and target tasks [6]. When the sources are relatively different from the target, multi-task learning (MTL) has often been observed to perform worse than single-task learning (STL) [7, 8], which is referred to as *negative transfer* [9]. While many empirical approaches have been proposed to mitigate negative transfer [5], a precise understanding of when negative transfer occurs remains elusive in the literature [10].

Understanding negative transfer requires developing generalization bounds that scale tightly with properties of each task data, such as its sample size. This presents a technical challenge in the multi-task setting because of the difference among task features, even for two tasks. For Rademacher complexity or VC-based techniques, the generalization error scales down as the sample sizes of all tasks increase, when applied to the multi-task setting [11, 12, 13, 14, 15]. Without a tight lower bound for multi-task learning, comparing its performance to single-task learning results in vacuous bounds. From a practical standpoint, developing a better understanding of multi-task learning in terms of properties of task data can provide guidance for downstream applications [16].

In this work, we study the bias and variance of multi-task learning in the high-dimensional linear regression setting [17, 18]. Our key observation is that three properties of task data, including *task similarity*, *sample size*, and *covariate shift*, can affect whether multi-task learning outperforms single-task learning (which we refer to as *positive transfer*). As an example, we vary each property in Figure 1 for two linear regression tasks and measure the improvement of multi-task learning over single-task learning for a particular task. We observe that the effect of transfer can be positive or negative as we vary each property. These phenomena cannot be explained using previous techniques [15]. The high-dimensional linear regression setting allows us to measure the three properties precisely. We

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.



(a) Task similarity (b) Sample size (c) Covariate shift

Figure 1: We observe a transition from positive to negative transfer as (a) *task model distance* increases and (b) source/target *sample ratio* increases. For the special case of having the same task model, we observe in (c) that as source/target *sample ratio* increases, having *covariate shift* worsens the performance of MTL. The y-axis measures the loss of STL minus MTL.

define each property for the case of two tasks and our definition applies to general settings. We refer to the first task as the source task and the second as the target task.

- **Task similarity:** Assume that both tasks follow a linear model with parameters $\beta_1, \beta_2 \in \mathbb{R}^p$, respectively. We measure the distance between them by $\|\beta_1 - \beta_2\|$.
- **Sample size:** Let $n_1 = \rho_1 \cdot p, n_2 = \rho_2 \cdot p$ be the sample size of each task, where $\rho_1, \rho_2 > 1$ are both fixed values that do not grow with p. We measure the source/target sample ratio by ρ_1/ρ_2.
- **Covariate shift:** Assume that the task features are random vectors with positive semidefinite covariance matrix $\Sigma_1, \Sigma_2 \in \mathbb{R}^{p \times p}$, respectively. We measure covariate shift with matrix $\Sigma_1^{1/2}\Sigma_2^{-1/2}$.

We consider a multi-task estimator obtained using a shared linear layer for all tasks and a separate output layer for each task [15]. This two-layer model is inspired by a commonly used idea of hard parameter sharing in multi-task learning [10, 19]. We consider the bias and variance of the multi-task estimator for predicting a target task and compare its performance to single-task learning.

Main results. First, we develop tight bounds for the bias and variance of the multi-task estimator for two tasks by applying recent development in random matrix theory [20, 21, 22]. We observe that the variance of the multi-task estimator is *always smaller* than single-task learning, because of added source task samples. On the other hand, the bias of the multi-task estimator is *always larger* than single-task learning, because of model distances. Hence, the tradeoff between bias and variance determines whether the transfer is positive or negative. We provide a sharp analysis of the *variance* that scales with sample size and covariate shift. We extend the analysis to the bias, which *in addition* scales with task similarity. Combining both, we analyze the bias-variance tradeoff for two tasks in Theorem 3.2 and extend the analysis to many tasks with the same features in Theorem 3.6.

Second, we explain the phenomena in Figure 1 in isotropic and covariate shifted settings.

- We provide conditions to predict the effect of transfer as a parameter of model distance $\|\beta_1 - \beta_2\|$ (Section 3.2). As model distance increases, the bias becomes larger, resulting in negative transfer.
- We provide conditions to predict transfer as a parameter of sample ratio ρ_1/ρ_2 (Section 3.3). Adding source task samples helps initially by reducing variance, but hurts eventually due to bias.
- For a special case of $\beta_1 = \beta_2$, we show that MTL performs best when the singular values of $\Sigma_1^{1/2}\Sigma_2^{-1/2}$ are all equal (Section 3.4). Otherwise, the variance reduces less with covariate shift.

Along the way, we analyze the benefit of MTL for reducing labeled data to achieve comparable performance to STL, which has been empirically observed in Taskonomy by Zamir et al. [2].

Our study also leads to several algorithmic consequences with practical interest. First, we show that single-task learning results can help predict positive or negative transfer for multi-task learning. We validate this observation on ChestX-ray14 [1] and sentiment analysis datasets [23]. Second, we propose a new multi-task training schedule by incrementally adding task data batches to the training procedure. This is inspired by our observation in Figure 1b where adding more source task data helps initially, but hurts eventually. Using our incremental training schedule, we reduce the computational cost by 65% compared to baseline multi-task training over six sentiment analysis datasets while keeping the accuracy the same. Third, we provide a fine-grained insight on a covariance alignment procedure proposed in [15]. We show that the alignment procedure provides more significant improvement when the source/target sample ratio is large. Finally, we validate our three theoretical findings on sentiment analysis tasks.

2 Problem Formulation for Multi-Task Learning

We begin by defining our problem setup including the multi-task estimator we study. Then, we describe the bias-variance tradeoff of the multi-task estimator and connect the bias and variance of the estimator to *task similarity*, *sample size*, and *covariate shift*.

Problem setup. Suppose we have t datasets, where t is a fixed value that does not grow with the feature dimension p. In the high-dimensional linear regression setting (e.g. [17, 18]), the features of the k-th task, denoted by $X_k \in \mathbb{R}^{n_i \times p}$, consist of n_k feature vectors given by $x_1, x_2, \ldots, x_{n_k}$. And each feature $x_i = \Sigma^{1/2} z_i$, where $z_i \in \mathbb{R}^p$ consists of i.i.d. entries with mean zero and unit variance. The sample size n_k equals $\rho_k \cdot p$ for a fixed value ρ_k. The labels $Y_k = X_k \beta_k + \varepsilon_k$, where β_k denotes the linear model parameters and ε_k denotes i.i.d. noise with mean zero and variance σ^2.

We focus on the commonly used hard parameter sharing model for multi-task learning [10]. When specialized to the linear regression setting, the model consists of a linear layer $B \in \mathbb{R}^{p \times r}$ that is shared by all tasks and t output layers $W_1, \ldots, W_t$ that are in $\mathbb{R}^r$. The width of B, denoted by r, plays an important role in regularization. As observed in Proposition 1 of [15], if $r \geqslant t$, there is no regularization effect. Hence, we assume that $r < t$ in our study. For example, when there are only two tasks, $r = 1$ and B reduces to a vector whereas W_1, W_2 become scalars. We study the following procedure inspired by how hard parameter sharing models are trained in practice (e.g. [19]).

- Separate each dataset (X_i, Y_i) randomly into a training set (X_i^{tr}, Y_i^{tr}) and a validation set (X_i^{val}, Y_i^{val}). The size of each set is described below.
- Learn the shared layer B: minimize the training loss over B and $W_1, \ldots, W_t$, leading to a closed form equation for $\hat{B}$ that depends on $W_1, \ldots, W_k$.

$$f(B; W_1, \ldots, W_t) = \sum_{k=1}^{t} \|X_k^{tr} B W_k - Y_k^{tr}\|^2. \tag{2.1}$$

- Tune the output layers W_i: set $B = \hat{B}$ and minimize the validation loss over $W_1, \ldots, W_k$.

$$g(W_1, \ldots, W_t) = \sum_{k=1}^{t} \|X_k^{val} \hat{B} W_k - Y_k^{val}\|^2. \tag{2.2}$$

We make several remarks. In general, the objective $f(\cdot)$ is non-convex in B and the W_k's. Therefore, we first minimize B in equation (2.1) and then minimize W_k given B in equation (2.2). For our purpose, a validation set of size $\rho_i \cdot p^{0.99}$ that is much larger than the number of output layer parameters $r \cdot t$ suffices. The size of the training set is then $\rho_i(p - p^{0.99})$. The advantage of tuning the output layers on the validation set is to reduce the effect of noise from $\hat{B}$.

Problem statement. We focus on predicting a particular task, say the t-th task, without loss of generality. Let $\hat{\beta}_t^{\text{MTL}}$ denote the multi-task estimator obtained from the procedure above. Our goal is to compare the prediction loss of $\hat{\beta}_t^{\text{MTL}}$, defined by

$$L(\hat{\beta}_t^{\text{MTL}}) = \mathop{\mathbb{E}}_{\{\varepsilon_i\}_{i=1}^t} \mathop{\mathbb{E}}_{x = \Sigma_t^{1/2} z} \left[(x^\top \hat{\beta} - x^\top \beta_t)^2 \right] = \mathop{\mathbb{E}}_{\{\varepsilon_i\}_i^t} \left\| \Sigma_2^{1/2} (\hat{\beta}_t^{\text{MTL}} - \beta_t) \right\|^2,$$

to the prediction loss $L(\hat{\beta}_t^{\text{STL}})$ of the single-task estimator $\hat{\beta}_t^{\text{STL}} = (X_t^\top X_t)^{-1} X_t^\top Y_t$. We say there is negative transfer if $L(\hat{\beta}_t^{\text{MTL}}) > L(\hat{\beta}_t^{\text{STL}})$ and positive transfer otherwise.

As an example, for the setting of two tasks, we can decompose $L(\hat{\beta}_t^{\text{MTL}}) - L(\hat{\beta}_t^{\text{STL}})$ into a bias term and a variance term as follows (derived in Appendix B).

$$L(\hat{\beta}_t^{\text{MTL}}) - L(\hat{\beta}_t^{\text{STL}}) = \hat{v}^2 \left\| \Sigma_2^{1/2} (\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 (\beta_1 - \hat{v}\beta_2) \right\|^2 \tag{2.3}$$

$$+ \sigma^2 \left(\operatorname{Tr}\left[(\hat{v}^2 X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_2 \right] - \operatorname{Tr}\left[(X_2^\top X_2)^{-1} \Sigma_2 \right] \right). \tag{2.4}$$

In the above, $\hat{v} = W_1 / W_2$ where W_1, W_2 are obtained from solving equation (2.2) (recalling that W_1, W_2 are scalars for two tasks). The role of $\hat{v}$ is to scale the shared subspace B to fit each task.

Equation (2.3) corresponds to the bias of $\hat{\beta}_t^{\text{MTL}}$. Hence, the bias term introduces a negative effect that depends on the *similarity* between β_1 and β_2. Equation (2.4) corresponds to the variance of $\hat{\beta}_t^{\text{MTL}}$ minus the variance of $\hat{\beta}_t^{\text{STL}}$, which is always negative. Intuitively, the more *samples* we have, the smaller the variance is. Meanwhile, *covariate shift* also affects how small the variance can be.

3 Comparing Multi-Task Learning to Single-Task Learning

We provide tight bounds on the bias and variance of the multi-task estimator for two tasks. We show theoretical implications for understanding the performance of multi-task learning. (a) *Task similarity*: we explain the phenomenon of negative transfer precisely as tasks become more different. (b) *Sample size*: we further explain a curious phenomenon where increasing the source sample size helps initially, but hurts eventually. (c) *Covariate shift*: as the source sample size increases, we show that the covariate shift worsens the performance of the multi-task estimator. Finally, we extend our results from two tasks to many tasks with the same features.

3.1 Analyzing the Bias-Variance Tradeoff using Random Matrix Theory

A well-known result in the high-dimensional linear regression setting states that $\mathrm{Tr}[(X_2^\top X_2)^{-1}\Sigma_2]$ is concentrated around $1/(\rho_2 - 1)$ (e.g. Chapter 6 of [24]), which scales with the sample size of the target task. Our main technical contribution is to extend this result to two tasks. We show how the variance of the multi-task estimator scales with sample size and covariate shift in the following result.

Lemma 3.1 (Variance bound)**.** *In the setting of two tasks, let $n_1 = \rho_1 \cdot p$ and $n_2 = \rho_2 \cdot$ be the sample size of the two tasks. Let $\lambda_1, \ldots, \lambda_p$ be the singular values of the covariate shift matrix $\Sigma_1^{1/2}\Sigma_2^{-1/2}$ in decreasing order. With high probability, the variance of the multi-task estimator $\hat{\beta}_t^{MTL}$ equals*

$$\frac{\sigma^2}{n_1+n_2} \cdot \mathrm{Tr}\left[(\hat{v}^2 a_1 \Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} + a_2\,\mathrm{Id})^{-1}\right] + \mathrm{O}\left(p^{-1/2+o(1)}\right),$$

where a_1, a_2 are solutions of the following equations:

$$a_1 + a_2 = 1 - \frac{1}{\rho_1+\rho_2}, \quad a_1 + \frac{1}{\rho_1+\rho_2} \cdot \frac{1}{p}\sum_{i=1}^{p} \frac{\hat{v}\lambda_i^2 a_1}{\hat{v}\lambda_i^2 a_1 + a_2} = \frac{\rho_1}{\rho_1+\rho_2}.$$

Lemma 3.1 allows us to get a tight bound on equation (2.4), that only depends on *sample size*, *covariate shift* and the scalar $\hat{v}$. As a remark, the concentration error $\mathrm{O}(p^{-1/2+o(1)})$ of our result is nearly optimal. For the bias term of equation (2.3), a similar result that scales with task model distance in addition to sample size and covariate shift holds (cf. Lemma C.3 in Appendix C). Combining the two lemmas, we provide a sharp analysis of the bias-variance tradeoff of the multi-task estimator. For a matrix X, let $\lambda_{\min}(X)$ denote its smallest singular value and $\|X\|$ denote its spectral norm.

Theorem 3.2 (Two tasks)**.** *For the setting of two tasks, let $\delta > 0$ be a fixed error margin, $\rho_2 > 1$ and $\rho_1 \gtrsim \delta^{-2} \cdot \lambda_{\min}(\Sigma_1^{1/2}\Sigma_2^{-1/2})^{-4}\|\Sigma_1\| \max(\|\beta_1\|^2, \|\beta_2\|^2)$. There exists two deterministic functions Δ_{bias} and Δ_{var} that only depend on $\{\hat{v}, \Sigma_1, \Sigma_2, \rho_1, \rho_2, \beta_1, \beta_2\}$ such that*

- *If $\Delta_{bias} - \Delta_{var} < -\delta$, then w.h.p. over the randomness of X_1, X_2, we have $L(\hat{\beta}_t^{MTL}) < L(\hat{\beta}_t^{STL})$.*
- *If $\Delta_{bias} - \Delta_{var} > \delta$, then w.h.p. over the randomness of X_1, X_2, we have $L(\hat{\beta}_t^{MTL}) > L(\hat{\beta}_t^{STL})$.*

Theorem 3.2 applies to settings where large amounts of source task data are available but the target sample size is small. For such settings, we obtain a sharp transition from positive transfer to negative transfer determined by $\Delta_{\text{bias}} - \Delta_{\text{var}}$. While the general form of these functions can be complex (as are previous generalization bounds for MTL), they admit interpretable forms for simplified settings.

The proof of Theorem 3.2 is presented in Appendix C and the proof of Lemma 3.1 is in Appendix F.

3.2 Task Similarity

It is well-known since the seminal work of Caruana [6] that how well multi-task learning performs depends on task relatedness. We formalize this connection in the following simplified setting, where we can perform explicit calculations. We show that as we increase the distance between β_1 and β_2, there is a transition from positive transfer to negative transfer in MTL.

The isotropic model. Consider two tasks with isotropic covariances $\Sigma_1 = \Sigma_2 = \mathrm{Id}$. Each task has sample size $n_1 = \rho_1 \cdot p$ and $n_2 = \rho_2 \cdot p$. Assume that for task two, β_2 has i.i.d. entries with mean zero and variance κ^2. For the source task, β_1 equals β_2 plus i.i.d. entries with mean 0 and variance d^2. The labels are $Y_i = X_i\beta_i + \varepsilon_i$, where ε_i consists of i.i.d. entries with mean zero and variance σ^2. For our purpose, it is enough to think of the order of d being $1/\sqrt{p}$ and pd^2/σ^2 being constant.

We introduce the following notations.

$$\Psi(\beta_1, \beta_2) = \mathbb{E}\left[\|\beta_1 - \beta_2\|^2\right]/\sigma^2, \quad \Phi(\rho_1, \rho_2) = \frac{(\rho_1+\rho_2-1)^2}{\rho_1(\rho_1+\rho_2)(\rho_2-1)}.$$

Proposition 3.3 (Task model distance)**.** *In the isotropic model, suppose that ρ_1 and $\rho_2 > 1$. Then*

- *If $\Psi(\beta_1, \beta_2) < \frac{1}{\nu} \cdot \Phi(\rho_1, \rho_2)$, then w.h.p. over the randomness of X_1, X_2, $L(\hat{\beta}_t^{MTL}) < L(\hat{\beta}_t^{STL})$.*
- *If $\Psi(\beta_1, \beta_2) > \nu \cdot \Phi(\rho_1, \rho_2)$, then w.h.p. over the randomness of X_1, X_2, $L(\hat{\beta}_t^{MTL}) > L(\hat{\beta}_t^{STL})$.*

Here $\nu = (1 - \mathrm{o}(1)) \min((1 - 1/\sqrt{\rho_1})^{-4}, (1 + 1/\sqrt{\rho_1})^4)$. Concretely, if $\rho_1 > 40$, then $\nu \in (1, 2)$.

Proposition 3.3 simplifies Theorem 3.2 in the isotropic model, allowing for a more explicit statement of the bias-variance tradeoff. Concretely, $\Psi(\beta_1, \beta)$ and $\Phi(\rho_1, \rho_2)$ corresponds to Δ_{bias} and Δ_{var}, respectively. Roughly speaking, the transition threshold scales as $\frac{pd^2}{\sigma^2} - \frac{1}{\rho_1} - \frac{1}{\rho_2}$. We apply Proposition 3.3 to the parameter setting of Figure 1a (the details are left to Appendix G.1). We can see that our result is able to predict positive or negative transfer accurately and matches the empirical curve. There are several unexplained observations near the transition threshold 0, which are caused by the concentration error ν. The proof of Proposition 3.3 can be found in Appendix D.1. A key part of the analysis shows that $\hat{v} \approx 1$ in the isotropic model, thus simplifyling the result of Theorem 3.2.

Algorithmic consequence. We can in fact extend the result to the cases where the noise variances are different. In this case, we will see that MTL is particularly effective. Concretely, suppose the noise variance σ_1^2 of task 1 differs from the noise variance σ_2^2 of task 2. If σ_1^2 is too large, the source task provides a negative transfer to the target. If σ_1^2 is small, the source task is more helpful. We leave the result to Proposition D.2 in Appendix D.1. Inspired by the observation, we propose a single-task based metric to help understand MTL results using STL results.

- For each task, we train a single-task model. Let z_s and z_t be the prediction accuracy of each task, respectively. Let $\tau \in (0, 1)$ be a fixed threshold.
- If $z_s - z_t > \tau$, then we predict that there will be positive transfer when combining the two tasks using MTL. If $z_s - z_t < -\tau$, then we predict negative transfer.

3.3 Sample Size

In classical Rademacher or VC based theory of multi-task learning, the generalization bounds are usually presented for settings where the sample sizes are equal for all tasks [11, 13, 14]. On the other hand, uneven sample sizes between different tasks (or even dominating tasks) have been empirically observed as a cause of negative transfer [25]. For such settings, we have also observed that adding more labeled data from one task does not always help. In the isotropic model, we consider what happens if we vary the source task sample size. Our theory accurately predicts a curious phenomenon, where increasing the sample size of the source task results in negative transfer!

Proposition 3.4 (Source/target sample ratio)**.** *In the isotropic model, suppose that $\rho_1 > 40$ and $\rho_2 > 110$ are fixed constants, and $\Psi(\beta_1, \beta_2) > 2/(\rho_2 - 1)$. Then we have that*

- *If $\frac{n_1}{n_2} = \frac{\rho_1}{\rho_2} < \frac{1}{\nu} \cdot \frac{1 - 2\rho_2^{-1}}{\Psi(\beta_1, \beta_2)(\rho_2 - 1) - \nu^{-1}}$, then w.h.p. $L(\hat{\beta}_t^{MTL}) < L(\hat{\beta}_t^{STL})$.*
- *If $\frac{n_1}{n_2} = \frac{\rho_1}{\rho_2} > \nu \cdot \frac{1 - 2\rho_2^{-1}}{\Psi(\beta_1, \beta_2)(\rho_2 - 1.5) - \nu}$, then w.h.p. $L(\hat{\beta}_t^{MTL}) > L(\hat{\beta}_t^{STL})$.*

Proposition 3.4 describes the bias-variance tradeoff in terms of the sample ratio ρ_1/ρ_2. We apply the result to the setting of Figure 1b (described in Appendix G.1). There are several unexplained observations near $y = 0$ caused by ν. The proof of Proposition 3.4 can be found in Appendix D.2.

Connection to Taskonomy. We use our tools to explain a key result of Taskonomy by Zamir et al. [2], which shows that MTL can reduce the amount of labeled data needed to achieve comparable performance to STL. For $i = 1, 2$, let $\hat{\beta}_i^{\text{MTL}}(x)$ denote the estimator trained using $x \cdot n_i$ datapoints from every task. The data efficiency ratio is defined as

$$\arg\min_{x \in (0,1)} \; L_1(\hat{\beta}_1^{\text{MTL}}(x)) + L_2(\hat{\beta}_2^{\text{MTL}}(x)) \leqslant L_1(\hat{\beta}_1^{\text{STL}}) + L_2(\hat{\beta}_2^{\text{STL}}).$$

For example, the data efficiency ratio is 1 if there is negative transfer. Using our tools, we show that in the isotropic model, the data efficiency ratio is roughly

$$\frac{1}{\rho_1 + \rho_2} + \frac{2}{(\rho_1 + \rho_2)(\rho_1^{-1} + \rho_2^{-1} - \Theta(\Psi(\beta_1, \beta_2)))}.$$

Compared with Proposition 3.3, we see that when $\Psi(\beta_1, \beta_2)$ is smaller than $\rho_1^{-1} + \rho_2^{-1}$ (up to a constant multiple), the transfer is positive. Moreover, the data efficiency ratio quantifies how effective the positive transfer is using MTL. The result can be found in Proposition D.3 in Appendix D.2.

Algorithmic consequence. An interesting consequence of Proposition 3.4 is that $L(\hat{\beta}_t^{\text{MTL}})$ is not monotone in ρ_1. In particular, Figure 1b (and our analysis) shows that $L(\hat{\beta}_t^{\text{MTL}})$ behaves as a quadratic function over ρ_1. More generally, depending on how large $\Psi(\beta_1, \beta_2)$ is, $L(\hat{\beta}_t^{\text{MTL}})$ may also be monotonically increasing or decreasing. Based on this insight, we propose an incremental optimization schedule to improve MTL training efficiency.

- We divide the source task data into S batches. For S rounds, we incrementally add the source task data by adding one batch at a time.
- After training T epochs, if the validation accuracy becomes worse than the previous round's result, we terminate. Algorithm 1 in Appendix G describes the procedure in detail.

3.4 Covariate Shift

So far we have considered the isotropic model where $\Sigma_1 = \Sigma_2$. This setting is relevant for settings where different tasks share the same input features such as multi-class image classification. In general, the covariance matrices of the two tasks may be different such as in text classification. In this part, we consider what happens when $\Sigma_1 \neq \Sigma_2$. We show that when n_1/n_2 is large, MTL with covariate shift can be suboptimal compared to MTL without covariate shift.

Example. We measure covariate shift by $M = \Sigma_1^{1/2}\Sigma_2^{-1/2}$. Assume that $\Psi(\beta_1, \beta_2) = 0$ for simplicity. We compare two cases: (i) when $M = \mathrm{Id}$; (ii) when M has $p/2$ singular values that are equal to λ and $p/2$ singular values that are equal to $1/\lambda$. Hence, λ measures the severity of the covariate shift. Figure 1c shows a simulation of this setting by varying λ. We observe that as source/target sample ratio increases, the performance gap between the two cases increases.

We compare different choices of M that belong to the following bounded set. Let λ_i be the i-th singular value of M. Let $\mu_{\min} < \mu < \mu_{\max}$ be fixed values that do not grow with p.

$$\mathcal{S}_\mu := \left\{ M \,\middle|\, \prod_{i=1}^{p} \lambda_i \leqslant \mu^p, \mu_{\min} \leqslant \lambda_i \leqslant \mu_{\max}, \text{ for all } 1 \leqslant i \leqslant p \right\},$$

Proposition 3.5 (Covariate shift)**.** *Assume that $\Psi(\beta_1, \beta_2) = 0$ and $\rho_1, \rho_2 > 1$. Let $g(M)$ denote the prediction loss of $\hat{\beta}_t^{MTL}$ when $M = \Sigma_1^{1/2}\Sigma_2^{-1/2} \in \mathcal{S}_\mu$. We have that*

$$g(\mu\,\mathrm{Id}) \leqslant (1 + \mathrm{O}\left(\rho_2/\rho_1\right)) \min_{M \in \mathcal{S}_\mu} g(M).$$

This proposition shows that when source/target sample ratio is large, then having no covariate shift is optimal. The proof of Proposition 3.5 is left to Appendix D.3.

Algorithmic consequence. Our observation highlights the need to correct covariate shift when n_1/n_2 is large. Hence for such settings, we expect procedures that aim at correcting covariate shift to provide more significant gains. We consider a covariance alignment procedure proposed in [15], which is designed for the purpose of correcting covariate shift. The idea is to add an alignment module between the input and the shared module B. This new module is then trained together with B and the output layers. We validate our insight on this procedure in the experiments.

3.5 Extensions

Next, we describe our result for more than two tasks with same features, i.e. $X_i = X$ for any i. This setting is prevalent in applications of multi-task learning to image classification, where there are multiple prediction labels/tasks for every image [1, 26].

Theorem 3.6 (Many tasks)**.** *For the setting of t tasks where $X_i = X$, for all $1 \leqslant i \leqslant t$, let $B^\star := [\beta_1, \beta_2, \ldots, \beta_t]$ and $U_r \in \mathbb{R}^{t \times r}$ denote the linear model parameters. Let $U_r U_r^\top$ denote the best rank-r subspace approximation of $(B^\star)^\top \Sigma B^\star$. Assume that $\lambda_{\min}(B^{\star\top}\Sigma B^\star) \gtrsim \sigma^2$. Let v_i denote the i-th row vector of U_r. There exists a value $\delta = \mathrm{o}\left(\|B^\star\|^2 + \sigma^2\right)$ such that*

- *If $\left(1 - \|v_t\|^2\right) \frac{\sigma^2}{\rho - 1} - \|\Sigma(B^\star U_r v_t - \beta_t)\|^2 > \delta$, then w.h.p $L(\hat{\beta}_t^{MTL}) < L(\hat{\beta}_t^{STL})$.*
- *If $\left(1 - \|v_t\|^2\right) \frac{\sigma^2}{\rho - 1} - \|\Sigma(B^\star U_r v_t - \beta_t)\|^2 < -\delta$, then w.h.p. $L(\hat{\beta}_t^{MTL}) > L(\hat{\beta}_t^{STL})$.*

Theorem 3.6 provides a sharp analysis of the bias-variance tradeoff beyond two tasks. Specifically, $(1 - \|v_t\|^2)\sigma^2/(\rho - 1)$ shows the amount of reduced variance and $\|\Sigma(B^\star U_r v_t - \beta_t)\|$ shows the bias of the multi-task estimator. The proof of 3.6 can be found in Appendix E.

4 Experiments

We validate our algorithmic insights and then our theory. In Section 4.2, we first show that single-task learning results can help predict positive or negative transfer. Second, our proposed incremental training schedule improves the training efficiency of standard multi-task training on sentiment analysis tasks. In Section 4.3, we validate our theoretical results. We further show that when the sample ratio is large, performing the alignment procedure of [15] provides more improvement for MTL.

4.1 Experimental Setup

We consider a text classification task and an image classification task as follows.

Sentiment Analysis. We consider six tasks: movie review sentiment (MR), sentence subjectivity (SUBJ), customer reviews polarity (CR), question type (TREC), opinion polarity (MPQA), and the Stanford sentiment treebank (SST) tasks. The question is to predict positive or negative sentiment expressed in the text. We use an embedding layer with GloVe embeddings followed by an LSTM, MLP or CNN layer proposed by [27].

ChestX-ray14. This dataset contains 112,120 frontal-view X-ray images. There are 14 diseases (tasks) for every image that we would like to predict. We use densenet121 as the shared module.

For all models, we use a shared module for all tasks and assign a separate output layer on top of the shared module for each task. The baseline training schedule for MTL is the round-robin training schedule. We measure the test accuracy of predicting a target task. We measure computional cost by summing over all epochs the number of samples used in every epoch.

4.2 Experimental Results

Predicting transfer effect via STL results. We show that the single-task based metric proposed in Section 3.2 can predict positive or negative transfer in MTL. A common challenge in the study of MTL is that the results can be hard to understand. It is difficult to predict when MTL performs well without running extensive trials. Our insight is that we can use STL results to help understand MTL results. Table 1 shows the result on both the sentiment analysis and the ChestX-ray14 tasks. We find that using a threshold of $\tau = 0.1$, the STL results correctly predict positive or negative transfer with 75.6% precision and 38.8% recall among 30 times 5 (random seeds) task pairs! We observe similar results for 91 task pairs from the ChestX-ray14 dataset.

Mitigating negative transfer via incremental training. First, we show that our proposed incremental training schedule (Algorithm 1) can help mitigate negative transfer for predicting a particular target task. Over all 15 pairs from the sentiment analysis tasks, we find that Algorithm 1 requires only 45% of the computational cost to achieve similar performance on the target task, compared to the MTL baseline. Our insight is that since adding more samples from the source task does not always help, we can improve efficiency by adding source samples *incrementally* during training.

Our next result shows the incremental training schedule applies to multiple tasks as well. In Table 2, we find that over all six sentiment analysis tasks, incremental training requires less than 35% of the computational cost compared to baseline MTL training, while achieving the same accuracy averaged over all six tasks. As a further validation, excluding TREC, we observe similar comparative results.

4.3 Validating the Theoretical Results

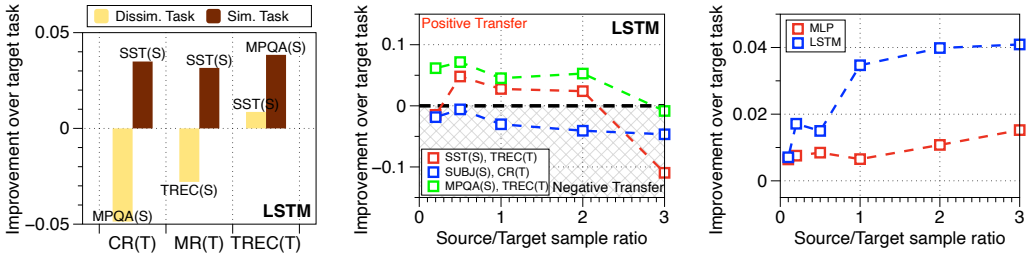
We first validate our theoretical results in Section 3.2 and 3.3. In Figure 2a, we compare the performance training with a semantically similar task versus a dissimilar task with a target task. We select each task pair based on our domain knowledge. We observe that adding a similar task helps the target task whereas adding a dissimilar task hurts. In Figure 2b, we validate that adding more source

Threshold	Sentiment analysis		ChestX-ray14	
	Precision	Recall	Precision	Recall
0.0	0.596	1.000	0.593	1.000
0.1	**0.756**	**0.388**	**0.738**	**0.462**
0.2	0.919	0.065	0.875	0.044

Table 1: Single-task learning results can help predict postive or negative transfer in multi-task learning.

Models	Sentiment analysis	
	all tasks	w/o TREC
MLP	31%	29%
LSTM	35%	34%
CNN	30%	28%

Table 2: Efficiency of incremental training compared to baseline MTL.



(a) Task similarity (b) Data size (c) Covariate shift

Figure 2: Validating the three results of Section 3 on sentiment analysis tasks. (a) Adding a semantically similar source task in MTL performs better than adding a dissimilar task. (b) As source/target sample ratio increases, we observe a transition from positive to negative transfer. (c) As source/target sample ratio increases, aligning task covariances [15] improves more over the baseline. Note: (S) denotes the source task and (T) denotes the target task.

samples does not always improve performance on the target task. Finally, we validate the algorithmic consequence of Section 3.4. In Figure 2c, we measure the performance gains from performing the alignment procedure proposed in [15] minus baseline MTL. We average the results over all 15 task pairs. The result shows that as the source samples increases, the alignment procedure shows a bigger improvement over MTL. The rest of experimental procedures are left to Appendix G.2.

5 Related Work

We refer the interested readers to several excellent surveys on multi-task learning for a comprehensive survey [9, 10, 5, 28]. Below, we describe several lines of work that are most related to this work.

Multi-task learning theory. Some of the earliest works on multi-task learning are Baxter [11], Ben-David and Schuller [29]. Mauer [13] studies generalization bounds for linear separation settings of MTL. Ben-David et al. [30] provides uniform convergence bounds that combines source and target errors in an optimal way. The benefit of learning multi-task representations is studied for learning certain half-spaces [14] and sparse regression [31, 32]. Our work is closely related to Wu et al. [15]. While Wu et al. provide generalization bounds to show that adding more labeled helps learn the target task more accurately, their techniques do not explain the phenomena of negative transfer.

Multi-task learning methodology. Ando and Zhang [12] introduces an alternating minimization framework for learning multiple tasks. Argyriou et al. [33] present a convex algorithm which learns common sparse representations across a pool of related tasks. Evgeniou et al. [34] develop a framework for multi-task learning in the context of kernel methods. The multi-task learning model that we have focused on is based on the idea of hard parameter sharing [35, 36, 10]. We believe that the technical tools we have developed can also be applied to many other multi-task learning models.

Random matrix theory. The random matrix theory tool and related proof of our work fall into a paradigm of the so-called local law of random matrices [20]. For a sample covariance matrix $X^\top X$ with $\Sigma = \mathrm{Id}$, such a local law was proved in [21]. It was later extended to sample covariance matrices with non-identity Σ [22], and separable covariance matrices [37]. On the other hand, one may derive the asymptotic result in Theorem 3.1 with error $\mathrm{o}(1)$ using the free addition of two independent random matrices in free probability theory [38]. To the best of my knowledge, we do not find an *explicit result* for the sum of two sample covariance matrices with general covariates in the literature.

6 Conclusions and Open Problems

In this work, we analyzed the bias and variance of multi-task learning versus single-task learning. We provided tight concentration bounds for the bias and the variance. Based on these bounds, we analyzed the impact of three properties, including task similarity, sample size, and covariate shift on the bias and variance, to derive conditions for transfer. We validated our theoretical results. Based on the theory, we proposed to train multi-task models by incrementally adding labeled data and showed encouraging results inspired by our theory. We describe several open questions for future work. First, our bound on the bias term (cf. Lemma C.3) involves an error term that scales down with ρ_1. Tightening this error bound can potentially cover the unexplained observations in Figure 1. Second, it would be interesting to extend our results to non-linear settings. We remark that this likely requires addressing significant technical challenges to deal with non-linearity.

Broader Impacts

In this work, we provide a theoretical study to help understand when multi-task learning performs well. We approach this question by studying the bias-variance tradeoff of multi-task learning. We provide new technical tools to bound the bias and variance. We relate the bounds to three properties of task data. We further provide guidance for detecting and mitigating negative transfer on image and text classification tasks.

Our theoretical framework has the potential to impact many other neighboring areas in the ML community. Multi-task learning connects to a wide range of areas [28]. To name a few, transfer learning, meta learning, multimodal learning, semi-supervised learning, and representation learning are all closely related areas to multi-task learning. Any learning scenario such as reinforcement learning [25] that combines multiple datasets to supervise a model is using multi-task learning. While the theoretical results that we have provided are not directly applicable to these different settings, we believe that the tools we have developed and the framework we have provided can inspire followup works in different settings. For one specific example, we have developed new concentration bounds that may apply to many settings such as soft parameter sharing [10] , kernel methods [34], and convex formulation of multi-task learning [39]. For another example, our results also allow extensions to transfer learning and domain adaptation [40]. The insights we have developed on positive and negative transfer can potentially find applications in multimodal learning, where the data sources are usually heterogeneous. Our fine-gained study on sample sizes have the potential to provide new insight in meta learning, where scarse labeled samples presents a significant challenge.

Our algorithmic consequences of our theory have the potential to impact downstream applications of multi-task learning. For example, many medical applications use multi-task learning to train large-scale image classification models by combining multiple datasets [1, 26]. Unlike the applications of multi-task learning in text classification where large amounts of labeled data are collected [3], in medical applications it is typically difficult to acquire large amounts of labeled data. For such settings, training multi-task models can be very challenging. Our insight on using single-task learning results to help understand multi-task learning can be valuable for helping practitioners understand their results.

References

[1] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[2] 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.

[3] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

[4] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275, 2019.

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

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

[7] Héctor Martínez Alonso and Barbara Plank. When is multitask learning effective? semantic sequence prediction under varying data conditions. *arXiv preprint arXiv:1612.02251*, 2016.

[8] Joachim Bingel and Anders Søgaard. Identifying beneficial task relations for multi-task learning in deep neural networks. *arXiv preprint arXiv:1702.08303*, 2017.

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

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

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

[12] 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.

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

[14] 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.

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

[16] Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4763–4771, 2019.

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

[18] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

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

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

[21] 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.

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

[23] 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.

[24] Vadim Ivanovich Serdobolskii. *Multiparametric statistics*. Elsevier, 2007.

[25] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020.

[26] Sabri Eyuboglu, Geoffrey Angus, Bhavik N. Patel, Anuj Pareek, Guido Davidzon, JaredDunnmon, and Matthew P. Lungren. Multi-task weak supervision enables automatedabnormality localization in whole-body fdg-pet/ct. 2020.

[27] 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.

[28] 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.

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

[30] 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.

[31] Karim Lounici, Massimiliano Pontil, Alexandre B Tsybakov, and Sara Van De Geer. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, 2009.

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

[33] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

[34] Theodoros Evgeniou, Charles A Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of machine learning research*, 6(Apr):615–637, 2005.

[35] Rich Caruana. Multitask learning: A knowledge-based source of inductive bias. *Proceedings of the Tenth International Conference on Machine Learning*.

[36] 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.

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

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

[39] Yu Zhang and Dit-Yan Yeung. A regularization approach to learning task relationships in multitask learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(3):1–31, 2014.

[40] Wouter M Kouw. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.

[41] Daniel McNamara and Maria-Florina Balcan. Risk bounds for transferring representations with and without fine-tuning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2373–2381. JMLR. org, 2017.

[42] Brian Bullins, Elad Hazan, Adam Kalai, and Roi Livni. Generalize across tasks: Efficient algorithms for linear representation learning. In *Algorithmic Learning Theory*, pages 235–246, 2019.

[43] Giovanni Cavallanti, Nicolo Cesa-Bianchi, and Claudio Gentile. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11(Oct):2901–2934, 2010.

[44] Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

[45] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532, 2019.

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

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

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

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

[50] Z. D. Bai and Jack W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.*, 26(1):316–345, 1998.

[51] 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.

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

[53] Johannes Alt, L. Erdős, and Torben Krüger. Local law for random Gram matrices. *Electron. J. Probab.*, 22:41 pp., 2017.

[54] Haokai Xi, Fan Yang, and Jun Yin. Local circular law for the product of a deterministic matrix with a random matrix. *Electron. J. Probab.*, 22:77 pp., 2017.

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

[56] 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.

[57] 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, 2013.