

# On Information Transfer in Learning Multiple High-Dimensional Regression Tasks with Covariate and Model Shifts

02/04/2020 at 8:12pm

## 1 Introduction

Multi-task learning that is applied on heterogeneous data can often result in suboptimal models (or negative transfer in more technical terms). In a previous work [5], we identified three factors that help determine when multi-task learning works, and when it doesn't. In this work, we zoom in to the task covariance part of [5] to further understand when multi-task learning works for heterogeneous data. We study a high-dimensional linear regression setting under covariate and model shifts. By using random matrix theory, we can explain several phenomena that are not explained by the techniques of [5]. These are achieved through tight generalization bounds established in the high-dimensional regression setting.

First, we provide a complete and precise picture for when there is positive vs negative transfer in the setting of learning two tasks.

- (i) **De-noising effect of MTL:** we show that multi-task learning has a de-noising effect of reducing the variance of the estimator. When this variance reduction effect is bigger than the bias caused by model shift, we get positive transfer.
- (ii) **Covariate shift affects the rate of transfer:** we further show that covariate shift can control the rate in which transfer occurs.
- (iii) **Negative transfer from model shift:** we establish a fundamental limit of multi-task learning compared to single-task learning in the presence of model shift. The phenomenon of negative transfer persists despite changing the model capacity or reweighting the tasks.

Secondly for the general setting with  $k$  tasks, we study the case when all tasks share the same covariates. This setting is prevalent in applications of multi-task learning to image classification, where there are multiple prediction labels/tasks for every image [1]. **We show the following results.**

## 2 Problem Setup

In the multi-task learning (MTL) problem, we are given the input of  $k$  tasks  $(X_1, Y_1), (X_2, Y_2), \dots, (X_k, Y_k)$ . We use a shared body (module)  $B \in \mathbb{R}^{p \times r}$  for all tasks and a separate head (module)  $\{W_i \in \mathbb{R}^r\}_{i=1}^k$  for each task. This corresponds to minimizing the following optimization objective.

$$f(B; W_1, \dots, W_k) = \sum_{i=1}^k \|X_i B W_i - Y_i\|^2. \quad (2.1)$$

Note that we consider the natural parameterization without reweighting the tasks above. The shared body  $B$  plays an important role because it allows information transfer between different task data. There are two ways to ensure the sharing of information between tasks.

- Adding a regularization over  $B$ , e.g. [3, 4].
- Controlling the capacity  $r$  of  $B$ , e.g. [2, 5]. Moreover, [2] observed that controlling the capacity can outperform the implicit capacity control of adding regularization over  $B$ .

**Data generation.** We shall assume that each task data follows a linear model. For every  $1 \leq i \leq k$ , we assume that

$$Y_i = X_i \beta_i + \varepsilon_i,$$

where  $\beta_i \in \mathbb{R}^p$  is the model parameter for the  $i$ -th task. Each row of  $X_i \in \mathbb{R}^{n_i \times p}$  is assumed to be drawn i.i.d. from a fixed distribution with covariance matrix  $\Sigma_i$ . We assume that for every row  $x$  of  $X_i$ , we have

$$\mathbb{E}[xx^\top] = \Sigma_i.$$

We also write  $x = \Sigma_i^{1/2} z_i$ , where  $z_i$  is a random vector with mean 0 and variance 1.

**Objectives.** We will designate the  $k$ -th task as the target. Our goal is to come up with an estimator  $\hat{\beta}$  to provide accurate predictions for the target task, provided with the other auxiliary task data. Concretely, we focus on two objectives.

- Estimation error for the target model  $\beta_t$ : we consider their distance

$$e(\hat{\beta}) := \mathbb{E}_{\varepsilon_i, \forall 1 \leq i \leq k} \left[ \left\| \hat{\beta} - \beta_t \right\|^2 \right].$$

- Test error for the target task:

$$\begin{aligned} te(\hat{\beta}) &:= \mathbb{E}_{x \sim \Sigma_k} \left[ \mathbb{E}_{\varepsilon_i, \forall 1 \leq i \leq k} \left[ (x^\top \hat{\beta} - x^\top \beta_t)^2 \right] \right] \\ &= \mathbb{E}_{\varepsilon_i, \forall 1 \leq i \leq k} \left[ (\hat{\beta} - \beta_t)^\top \Sigma_k (\hat{\beta} - \beta_t) \right]. \end{aligned}$$

## 2.1 Hypothesis

Our hypothesis is that the heterogeneity among the multiple tasks can be categorized into two classes, *covariate shift* and *model shift*. We consider two natural questions within each category.

- How does covariate shift affect the rate of information transfer? For example, is it better to have the same covariance matrix or not?
- Under model shift, when do we get positive vs. negative transfer? How does the type of transfer depend on the number of data points, the distance of the task models etc?

**Model shift.** In general the single-task models can also be different across different tasks. We shall argue that in addition to the bias and variance terms of generalization error, model shift introduces a third term which is the bias caused by model shift.

Here the hypothesis is that the optimal  $B$  is captured by a low-rank approximation of the single-task models?

**Covariate shift.** The covariance matrices  $\Sigma_i$  may be different across tasks, i.e. having different spectrum or singular vectors. This is also known as covariate shift in the literature. **Our hypothesis is that the covariate shift can slow down the convergence of learning the true  $\theta$  as a function of the number of data points.** A special case of this setting is that the single-task models are the same across all the tasks, i.e.  $\beta_i = \beta$ , for all  $1 \leq i \leq k$ .

Since all tasks share the same underlying model  $\beta$ , we use a simplified objective as follows.

$$f(w) = \sum_{i=1}^k \|X_i w - Y_i\|^2. \quad (2.2)$$

Equation (2.2) is simplified from equation (2.1) by setting  $A_i$  to be 1 for all tasks.

**Proposition 2.1.** *Suppose that  $n > p$ . When there is no model shift, adding the source task data always reduces the estimation error and the test error for the target task, i.e.*

$$e(\hat{\beta}_{s,t}) \leq e(\hat{\beta}_t), \text{ and} \quad (2.3)$$

$$te(\hat{\beta}_{s,t}) \leq te(\hat{\beta}_t) \quad (2.4)$$

*Proof.* Equation (2.3) is simply because

$$e(\hat{\beta}_{s,t}) = \text{Tr} [(X_1^\top X_1 + X_2^\top X_2)^{-1}] \leq \text{Tr} [(X_1^\top X_1)^{-1}] = e(\hat{\beta}_t).$$

Equation (2.4) follows because

$$\begin{aligned} te(\hat{\beta}_{s,t}) &= \text{Tr} [(X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_2] = \text{Tr} \left[ \left( \Sigma_2^{-1/2} X_1^\top X_1 \Sigma_2^{-1/2} + \Sigma_2^{-1/2} X_2^\top X_2 \Sigma_2^{-1/2} \right)^{-1} \right] \\ &\leq \text{Tr} \left[ \left( \Sigma_2^{-1/2} X_2^\top X_2 \Sigma_2^{-1/2} \right)^{-1} \right] = \text{Tr} [(X_2^\top X_2)^{-1} \Sigma_2] = te(\hat{\beta}_t). \end{aligned}$$

□

## 2.2 The High-Dimensional Setting

We would like to get insight on how covariate and model shifts affect the rate of transfer. We will consider the high-dimensional setting where for the target task, its number of data points is a small constant times  $p$ . This setting captures a wide range of applications of multi-task learning where we would like to use auxiliary task data to help train tasks with limited labeled data.

**The case of two tasks.** We can get precise rates using random matrix theory. For the sake of clarity, we call task 1 the source task and task 2 the target task, i.e.  $\beta_1 = \beta_s$  and  $\beta_2 = \beta_t$ . We introduce the following notations for the high-dimensional setting

$$c_{n_1} := \frac{n_1}{p} \rightarrow c_1, \quad c_{n_2} := \frac{n_2}{p} \rightarrow c_2, \quad \text{as } n_1, n_2 \rightarrow \infty,$$

for some constants  $c_1, c_2 \in (1, \infty)$ .

## 3 Warm Up: The Simplified Objective with Shared Weights Only

We begin by considering the simplified equation (2.2) to solve for the target task. This is equivalent to setting  $W_i = 1$  for the two tasks in equation (2.1). The result of this simplified setting will set up the ground

for solving equation (2.1) when we also optimize  $W_1$  and  $W_2$ . By putting the two tasks together, we get

$$\begin{aligned}\hat{\beta}_{s,t} &= (X_1^\top X_1 + X_2^\top X_2)^{-1} (X_1^\top Y_1 + X_2^\top Y_2) \\ &= (X_1^\top X_1 + X_2^\top X_2)^{-1} ((X_1^\top X_1 \beta_s + X_2^\top X_2 \beta_t) + (X_1^\top \varepsilon_1 + X_2^\top \varepsilon_2)) \\ &= \beta_t + (X_1^\top X_1 + X_2^\top X_2)^{-1} (X_1^\top X_1 (\beta_s - \beta_t) + (X_1^\top \varepsilon_1 + X_2^\top \varepsilon_2))\end{aligned}$$

Hence

$$e(\hat{\beta}_{s,t}) = \|(X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 (\beta_s - \beta_t)\|^2 + \sigma^2 \text{Tr}[(X_1^\top X_1 + X_2^\top X_2)^{-1}], \text{ and} \quad (3.1)$$

$$te(\hat{\beta}_{s,t}) = \left\| \Sigma_2^{1/2} (X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 (\beta_s - \beta_t) \right\|^2 + \sigma^2 \cdot \text{Tr}[(X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_2] \quad (3.2)$$

We can see that because of model shift, i.e.  $\beta_s \neq \beta_t$ . We can no longer guarantee that  $te(\hat{\beta}_{s,t}) \leq te(\hat{\beta}_t)$ . The goal of this part is to show conditions under which we get positive vs. negative transfer.

**For technical reasons, we will consider the test error.** From Lemma 3.9, we know that

$$\text{Tr}[(X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_2] = \frac{1}{n_1 + n_2} \cdot \text{Tr} \left[ \left( a_1 \Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2} + a_2 \text{Id} \right)^{-1} \right],$$

where  $a_1, a_2$  are specified in Theorem 3.5. Hence the variance part is reduced by the following amount

$$\Delta_{\text{var}} := \sigma^2 \cdot \left( \frac{p}{n_2 - p} - \frac{1}{n_1 + n_2} \text{Tr} \left[ (a_1 M^\top M + a_2 \text{Id})^{-1} \right] \right) \quad (3.3)$$

It remains to consider the increment from the first term in  $te(\hat{\beta}_{s,t})$ , which is bounded using the following lemma.

**Theorem 3.1** (Positive transfer under model shift). *Let  $a_1, a_2$  be the solutions from equations (3.13) and (3.14). Denote by  $M := \Sigma_1^{1/2} \Sigma_2^{-1/2}$ . We have  $te(\hat{\beta}_{s,t}) \leq te(\hat{\beta}_t)$  when*

$$\Delta_{\text{var}} \geq \left( 1 + \sqrt{\frac{p}{n_1}} \right)^2 \Delta_\beta, \quad (3.4)$$

where

$$\Delta_\beta := \frac{n_1^2}{(n_1 + n_2)^2} (\beta_s - \beta_t)^\top \Sigma_1^{1/2} M \frac{(1 + a_3) \text{Id} + a_4 M^\top M}{(a_2 + a_1 M^\top M)^2} M^\top \Sigma_1^{1/2} (\beta_s - \beta_t),$$

and  $a_3, a_4$  are the solutions of the following linear equations

$$\left( \frac{n_2}{a_2^2} - \sum_{i=1}^p \frac{1}{(a_2 + \lambda_i^2 a_1)^2} \right) a_3 - \left( \sum_{i=1}^p \frac{\lambda_i^2}{(a_2 + \lambda_i^2 a_1)^2} \right) a_4 = \sum_{i=1}^p \frac{1}{(a_2 + \lambda_i^2 a_1)^2}, \quad (3.5)$$

$$\left( \frac{n_1}{a_1^2} - \sum_{i=1}^p \frac{\lambda_i^4}{(a_2 + \lambda_i^2 a_1)^2} \right) a_4 - \left( \sum_{i=1}^p \frac{\lambda_i^2}{(a_2 + \lambda_i^2 a_1)^2} \right) a_3 = \sum_{i=1}^p \frac{\lambda_i^2}{(a_2 + \lambda_i^2 a_1)^2}. \quad (3.6)$$

Theorem 3.1 shows upper bounds that guarantee positive transfer, which is determined by the change of variance  $\Delta_{\text{var}}$  and a certain model shift bias parameter  $\Delta_\beta$  determined by the covariate shift matrix and the model shift.

**Negative transfer: the Limit caused by model shifts.** Next we describe lower bounds that guarantee negative transfer to complement Theorem 3.1.

**Theorem 3.2** (Negative transfer under model shift). *In the setting of Theorem 3.1, we have  $te(\hat{\beta}_{s,t}) \geq te(\hat{\beta}_t)$  when*

$$\Delta_{var} \leq \left(1 - 4\sqrt{\frac{p}{n_1}} - \frac{2p}{n_1}\right) \Delta_\beta$$

*In the special case that  $\beta_s - \beta_t$  is i.i.d. with mean 0 and variance  $d^2$  and  $\Sigma_1 = \text{Id}$ , we can get a tighter lower bound that guarantees negative transfer when*

$$\Delta_{var} \leq \left(1 - \sqrt{\frac{p}{n_1}}\right)^4 \Delta_\beta. \quad (3.7)$$

We describe examples based on Theorem 3.1 and Theorem 3.2 to show several conceptual insights.

**Example 3.3 (The effect of  $d$ , the distance of task models).** We consider a simple setting where  $\Sigma_1 = \text{Id}$ . Suppose that  $\beta_s - \beta_t$  is a vector with mean zero and variance  $d^2$ . Hence the task models have distance  $d^2 \cdot p$  in expectation.

We first consider  $\Sigma_2 = \text{Id}$ . In this case, we can simplify  $\Delta_\beta$  as follows

$$\Delta_\beta := d^2 \cdot \sum_{i=1}^p \frac{(1 + a_3)\lambda_i^2 + a_4\lambda_i^4}{(a_1\lambda_i^2 + a_2)^2}. \quad (3.8)$$

Now we solve the equations (3.13), (3.14), (3.5), (3.6) to get

$$a_1 = \frac{c_1(c_1 + c_2 - 1)}{(c_1 + c_2)^2}, a_2 = \frac{c_2(c_1 + c_2 - 1)}{(c_1 + c_2)^2}, a_3 = \frac{c_2}{(c_1 + c_2)(c_1 + c_2 - 1)}, a_4 = \frac{c_1}{(c_1 + c_2)(c_1 + c_2 - 1)}. \quad (3.9)$$

Then we obtain

$$\Delta_\beta = p \cdot d^2 \cdot \frac{c_1^2(c_1 + c_2)}{(c_1 + c_2 - 1)^3}, \Delta_{var} = \sigma^2 \cdot \frac{c_1}{(c_2 - 1)(c_1 + c_2 - 1)}. \quad (3.10)$$

We demonstrate our result with a simulation. We consider a setting where  $p = 200$ ,  $n_1 = 90p$ ,  $n_2 = 30p$ . **Fill in other params.** We fix the target task and vary the source task, by varying the task model distance parameter  $d$ . We show that Theorem 3.1 predicts whether we can get positive or negative transfer. Figure 1 shows the result. We obtain the following insight from the simulation.

- Adding the source task has the effect of reducing the variance of the estimator, independent of the model shift.
- Model shift introduces an additional bias term, which scales with  $d^2$ , the distance of the two task models. Hence, the type of transfer is determined by the tradeoff between the bias caused by model shift and the reduction of variance.

**Example 3.4 (The effect of  $n_1$ , the source task data size).**

### 3.1 The Effect of Covariate Shift on the Rate of Transfer

As Proposition 2.1 shows, if  $\beta_s$  and  $\beta_t$  are equal, then adding the source task dataset always helps learn the target task. The goal of this section is to understand how covariate shift affects the rate of transfer. **add conceptual msg**

The estimator using the source and target together from minimizing (2.2) is

$$\hat{\beta}_{s,t} = (X_1^\top X_1 + X_2^\top X_2)^{-1}(X_1^\top Y_1 + X_2^\top Y_2)$$

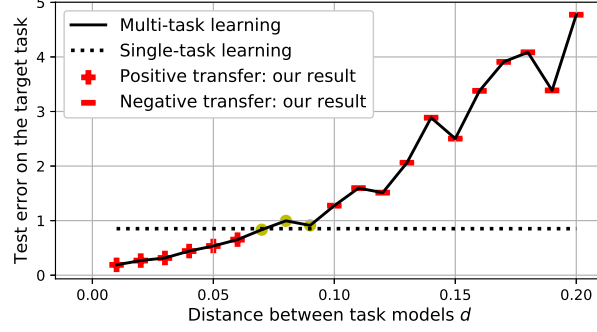


Figure 1: Positive vs negative transfer as a parameter of the task model distances.

The estimation error of  $\hat{\beta}_{s,t}$  is

$$e(\hat{\beta}_{s,t}) = \sigma^2 \cdot \text{Tr}[(X_1^\top X_1 + X_2^\top X_2)^{-1}]. \quad (3.11)$$

The estimation error using the target alone is

$$e(\hat{\beta}_t) = \sigma^2 \cdot \text{Tr}[(X_2^\top X_2)^{-1}]. \quad (3.12)$$

The improvement of estimation error from adding the source task is then given by  $e(\hat{\beta}_t) - e(\hat{\beta}_{s,t})$ . For the test error on the target task, the improvement from adding the source task is

$$te(\hat{\beta}_t) - te(\hat{\beta}_{s,t}) = \sigma^2 \cdot \text{Tr}[(X_2^\top X_2)^{-1} - (X_1^\top X_1 + X_2^\top X_2)^{-1} \cdot \Sigma_2].$$

We can get a precise result on the improvement of adding the source task data that only depends on the covariance matrices  $\Sigma_1, \Sigma_2$  and the number of data points  $n_1, n_2$ .

**Theorem 3.5** (Positive transfer rate under covariate shift). *Let  $\lambda_1, \lambda_2, \dots, \lambda_p$  denote the singular values of  $\Sigma_1^{1/2} \Sigma_2^{-1/2}$ . When there is no model shift, we have that*

$$e(\hat{\beta}_{s,t}) = \sigma^2 \cdot \text{Tr} \left[ \frac{1}{(n_1 + n_2)a_1 \Sigma_1 + (n_1 + n_2)a_2 \Sigma_2} \right]$$

$$te(\hat{\beta}_{s,t}) = \sigma^2 \cdot \text{Tr} \left[ \frac{1}{(n_1 + n_2)a_1 \Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2} + (n_1 + n_2)a_2 \text{Id}} \right]$$

where  $a_1, a_2$  are the solutions of the following equations

$$a_1 + a_2 = 1 - \frac{p}{n_1 + n_2} \quad (3.13)$$

$$a_1 + \sum_{i=1}^p \frac{a_1}{(n_1 + n_2)a_1 + (n_1 + n_2)a_2/\lambda_i^2} = \frac{n_1}{n_1 + n_2}. \quad (3.14)$$

**redo the following** As a remark, we see that Proposition 2.1 follows from Theorem 3.5. The amount of reduction on estimation error and test error for the target task is given as

$$e(\hat{\beta}_t) - e(\hat{\beta}_{s,t}) = \sigma^2 p \cdot \text{Tr} \left[ \frac{1}{(n_2 - p)\Sigma_2} - \frac{1}{(n_1 + n_2)a_1 \Sigma_1 + (n_1 + n_2)a_2 \Sigma_2} \right],$$

$$te(\hat{\beta}_t) - te(\hat{\beta}_{s,t}) = \sigma^2 p \cdot \text{Tr} \left[ \frac{1}{n_2 - p} - \frac{1}{(n_1 + n_2)a_1 \Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2} + (n_1 + n_2)a_2 \text{Id}} \right].$$

Because

$$\begin{aligned} e(\hat{\beta}_{s,t}) &\leq e(\hat{\beta}_t) \Leftarrow (n_2 - p)\Sigma_2 \preceq (n_1 + n_2)a_1\Sigma_1 + (n_1 + n_2)a_2\Sigma_2 \\ &\Leftrightarrow \mathbf{0} \preceq (n_1 + n_2)a_1\Sigma_1 + (n_1 - (n_1 + n_2) \cdot a_1)\Sigma_2, \end{aligned}$$

which is true since  $a_1 \leq n_1/(n_1 + n_2)$  by equation (3.14). The proof for  $te(\hat{\beta}_{s,t}) \leq te(\hat{\beta}_t)$  follows by multiplying  $\Sigma_2^{-1/2}$  on both sides of the inequalities above.

Now we apply Theorem 3.5 to show how covariate shift affects the rate of transfer.

*Example 3.6 (When  $\Sigma_1 = \Sigma_2$ ).* In this case, we have  $\lambda_i = 1$  for all  $1 \leq i \leq p$ . And  $a_1 + a_2 = 1 - p/(n_1 + n_2)$ . Hence

$$te(\hat{\beta}_{s,t}) = \frac{\sigma^2 p^2}{n_1 + n_2 - p} \text{ and } e(\hat{\beta}_{s,t}) = \frac{\sigma^2 p}{n_1 + n_2 - p} \text{Tr}[\Sigma_2^{-1}].$$

*Example 3.7 (When  $\Sigma_1 = \Sigma_2/\lambda$ ).* In this case, equations (3.13) and (3.14) become

$$a_1 + a_2 = 1 - p/(n_1 + n_2), a_1 + \frac{p}{n_1 + n_2} \cdot \frac{a_1}{a_1 + \lambda^2 a_2} = \frac{n_1}{n_1 + n_2}.$$

By solving these, we can get the test errors (the estimation error behaves similarly). Figure 2 shows how they grow as we increase the number of source task data points. Here  $n_2 = 4p$  and  $n_1$  ranges from  $p$  to  $20p$ . We can see that the smaller  $\lambda$  is, the lower the test errors will be.

*Example 3.8 (When  $\Sigma_1$  and  $\Sigma_2$  are complementary).* Here we have that for  $\Sigma_1^{-1/2}\Sigma_2^{1/2}$ , the first  $p/2$  singular values of  $\Sigma_1$  is  $\lambda$  and the rest is  $1/\lambda$ . For this case, equations (3.13) and (3.14) become

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

It's not hard to verify that there is only one valid solution from the above. After solving these, we get the test error for the target task as follows.

$$\frac{p}{2(n_1 + n_2)} \cdot \left( \frac{1}{\frac{a_1}{\lambda^2} + a_2} + \frac{1}{a_1 \lambda^2 + a_2} \right).$$

In Figure 3, we plot the test error of the target task for  $n_2 = 4p$  and  $n_1$  ranging from  $p$  to  $20p$ . We observe the following two phases as we increase  $n_1/p$ .

- When  $n_1 \leq n_2$ , having complementary covariance matrices leads to lower test error compared to the case when  $\Sigma_1 = \Sigma_2$ .
- When  $n_1 > n_2$ , having complementary covariance matrices leads to higher test error compared to the case when  $\Sigma_1 = \Sigma_2$ .

To prove Theorem 3.5, we study the spectrum of the random matrix model:

$$Q = \Sigma_1^{1/2} Z_1^T Z_1 \Sigma_1^{1/2} + \Sigma_2^{1/2} Z_2^T Z_2 \Sigma_2^{1/2},$$

where  $\Sigma_{1,2}$  are  $p \times p$  deterministic covariance matrices, and  $X_1 = (x_{ij})_{1 \leq i \leq n_1, 1 \leq j \leq p}$  and  $X_2 = (x_{ij})_{n_1+1 \leq i \leq n_1+n_2, 1 \leq j \leq p}$  are  $n_1 \times p$  and  $n_2 \times p$  random matrices, respectively, where the entries  $x_{ij}$ ,  $1 \leq i \leq n_1 + n_2 \equiv n$ ,  $1 \leq j \leq p$ , are real independent random variables satisfying

$$\mathbb{E}z_{ij} = 0, \quad \mathbb{E}|z_{ij}|^2 = n^{-1}. \quad (3.15)$$

For now, we assume that the random variables  $x_{ij}$  are i.i.d. Gaussian, but we know that universality holds for generally distributed entries.

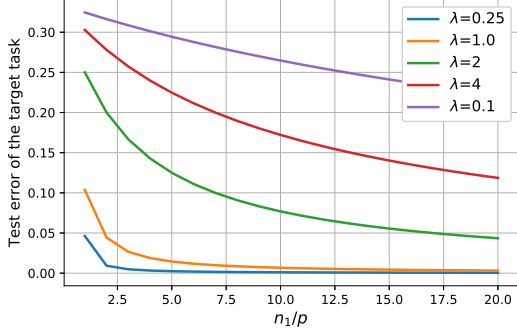


Figure 2: When  $\Sigma_1 = \Sigma_2/\lambda$ .

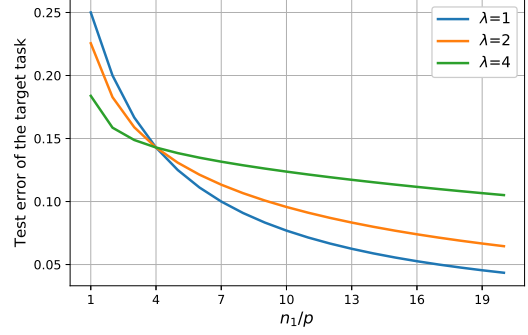


Figure 3: When  $\Sigma_1$  and  $\Sigma_2$  are complementary.

**Lemma 3.9.** *In the setting of Theorem 3.5, we have with high probability  $1 - o(1)$ ,*

$$\text{Tr}((X_1^\top X_1 + X_2^\top X_2)^{-1}) = \frac{1}{n_1 + n_2} \cdot \text{Tr} \left[ \frac{1}{a_1 \Sigma_1 + a_2 \Sigma_2} \right] + O \left( \frac{1}{n^{1-\varepsilon}} \right).$$

for any constant  $\varepsilon > 0$ . In particular, when  $n_1 = 0$ , we have that  $a_1 = 0$  and  $a_2 = (n_2 - p)/n_2$ , hence

$$\text{Tr}[(X_2^\top X_2)^{-1}] = \frac{1}{n_2 - p} \text{Tr} \left[ \frac{1}{\Sigma_2} \right] + O \left( \frac{1}{n_2^{1-\varepsilon}} \right).$$

## 4 The Multi-Task Learning Objective

In this part, we shall establish the intuition that adding more data helps in multi-task learning by reducing the variance of the estimator. We achieve this through tight generalization bounds obtained from random matrix theory. We focus the case of two tasks to provide a complete picture on how covariate and model shifts can affect the type of transfer. From [5], we know that either we need to explicitly restrict the capacity  $r$  of  $B$  so that there is transfer between the two tasks. Following [5], for the rest of the section, we shall consider the case when  $r = 1$  since there are only two tasks.

## 5 The Case with $k$ Tasks of the Same Covariates

## 6 Extensions to the Sparse Setting

**Describe a straightforward extension of our results here through isometry.** One important case is the sparse case, where we assume that the  $\beta_i$ 's are sparse in the sense that each  $\beta_i$  is in a subspace of dimension  $k_i \ll p$ . In other words,

$$\beta_i = \sum_{j=1}^{k_i} a_j^{(i)} v_j^{(i)},$$

where  $\{v_j^{(i)}\}$  is a set of  $k_i$  orthonormal vectors. For  $X_i = Z_i \Sigma_i^{1/2}$ , we have

$$X_i \beta_i = Z_i \Sigma_i^{1/2} \beta_i = Z_i \Sigma_i^{1/2} P_i \beta_i, \quad P_i := \sum_{j=1}^{k_i} v_j^{(i)} (v_j^{(i)})^\top.$$

Then we have

$$\frac{1}{n_i} \mathbb{E} \left[ P_i \Sigma_i^{1/2} Z_i^T Z_i \Sigma_i^{1/2} P_i \right] = P_i \Sigma_i P_i.$$



Hence our analysis will be valid with projected covariance matrices  $P_i \Sigma_i P_i$ . Moreover, due to the sparsity assumption, it is natural to have restricted isometry property:

$$\frac{1}{n_i} P_i \Sigma_i^{1/2} Z_i^T Z_i \Sigma_i^{1/2} P_i \approx P_i \Sigma_i P_i.$$

Finally, for our analysis it is not necessary to assume that  $\beta_T$  is sparse for the target.

## References

- [1] Sabri Eyuboglu, Geoffrey Angus, Bhavik N. Patel, Anuj Pareek, Guido Davidzon, JaredDunnmon, and Matthew P. Lungren. Multi-task weak supervision enables automated abnormality localization in whole-body fdg-pet/ct. 2020.
- [2] Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.
- [3] 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.
- [4] 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.
- [5] Sen Wu, Hongyang Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*, 2020.

## 7 Proofs of Theorem bla

The proofs of Theorem 3.1 and Theorem 3.2 involve two parts.

**Part I: Bounding the bias from model shift.** We relate the first term in equation (3.2) to  $\Delta_\beta$ .

**Proposition 7.1.** *In the setting of Theorem 3.1, we have that*

$$\left(1 - 4\sqrt{\frac{p}{n_1}} - 2\frac{p}{n_1}\right) \Delta_\beta \leq \left\| \Sigma_2^{1/2} (X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 (\beta_s - \beta_t) \right\|^2 \leq \left(1 + \sqrt{\frac{p}{n_1}}\right)^4 \Delta_\beta.$$

For the special case when  $\Sigma_1 = \text{Id}$  and  $\beta_s - \beta_t$  is i.i.d. with mean 0 and variance  $d^2$ , we further have

$$\left(1 - \sqrt{\frac{p}{n_1}}\right)^4 \Delta_\beta \leq \left\| \Sigma_2^{1/2} (X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 (\beta_s - \beta_t) \right\|^2.$$

*Proof.* The proof follows by applying equation (6.2). Recall that  $X_1^\top X_1 = \Sigma_1^{1/2} Z_1^\top Z_1 \Sigma_1^{1/2}$ . Denote by  $\mathcal{E} = Z_1^\top Z_1 - n_1 \text{Id}$ . Let  $K = (X_1^\top X_1 + X_2^\top X_2)^{-1}$ . Let  $\alpha = \left\| \Sigma_2^{1/2} K \Sigma_1 (\beta_s - \beta_t) \right\|^2$ . We have

$$\begin{aligned} & \left\| \Sigma_2^{1/2} (X_1^\top X_1 + X_2^\top X_2)^{-1} X_1^\top X_1 (\beta_s - \beta_t) \right\|^2 \\ &= n_1^2 \alpha + 2n_1 (\beta_s - \beta_t)^\top \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \beta_t) + \left\| \Sigma_2^{1/2} K \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} (\beta_s - \beta_t) \right\|^2 \\ &\leq n_1 \left( n_1^2 + \frac{2n_1}{p} (p + 2\sqrt{n_1 p}) + (p + 2\sqrt{n_1 p})^2 \right) \alpha = n_1^2 \left( 1 + \sqrt{\frac{p}{n_1}} \right)^4 \alpha. \end{aligned} \tag{7.1}$$

Here we use the following on the second term in equation (6.1)

$$\begin{aligned}
& \left| (\beta_s - \beta_t)^\top \Sigma_1^{1/2} \mathcal{E} \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \beta_t) \right| \\
&= \left| \text{Tr} \left[ \mathcal{E} \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \beta_t) (\beta_s - \beta_t)^\top \Sigma_1^{1/2} \right] \right| \\
&\leq \|\mathcal{E}\| \cdot \left\| \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \beta_t) (\beta_s - \beta_t)^\top \Sigma_1^{1/2} \right\|_* \\
&\leq n_1 \left( 2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1} \right) \cdot \left\| \Sigma_1^{1/2} K \Sigma_2 K \Sigma_1 (\beta_s - \beta_t) (\beta_s - \beta_t)^\top \Sigma_1^{1/2} \right\|_* \quad (\text{by equation (6.2)}) \\
&= n_1 \left( 2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1} \right) \alpha \quad (\text{since the matrix inside is rank 1})
\end{aligned}$$

The third term in equation (6.1) can be bounded similarly.

For the other direction, we simply note that the third term in equation (6.1) is positive. And the second term is bigger than  $-2n_1^2(2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1})\alpha$  using equation (6.2).  $\square$

**Part II: The limit of  $\left\| \Sigma_2^{1/2} (X_1^\top X_1 + X_2^\top X_2)^{-1} \Sigma_1 (\beta_s - \beta_t) \right\|^2$  using random matrix theory.** We consider the same setting as in previous subsection:

$$X_1^\top X_1 := \Sigma_1^{1/2} Z_1^\top Z_1 \Sigma_1^{1/2}, \quad X_2^\top X_2 = \Sigma_2^{1/2} Z_2^\top Z_2 \Sigma_2^{1/2},$$

where  $z_{ij}$ ,  $1 \leq i \leq n_1 + n_2 \equiv n$ ,  $1 \leq j \leq p$ , are real independent random variables satisfying (3.15). For now, we assume that the random variables  $z_{ij}$  are i.i.d. Gaussian, but we know that universality holds for generally distributed entries. Assume that  $p/n_1$  is a small number such that  $Z_1^\top Z_1$  is roughly an isometry, that is, under (3.15), [If we assume the variances of the entries of  \$Z\_1\$  are 1, then we have](#)

$$-n_1 \left( 2\sqrt{\frac{p}{n_1}} - \frac{p}{n_1} \right) \leq Z_1^\top Z_1 - n_1 \text{Id} \leq n_1 \left( 2\sqrt{\frac{p}{n_1}} + \frac{p}{n_1} \right). \quad (7.2)$$

[\(revise the following proof\)](#) It remain to study the following expression

$$\begin{aligned}
\frac{1}{X_1^\top X_1 + X_2^\top X_2} \Sigma_2 \frac{1}{X_1^\top X_1 + X_2^\top X_2} &= \Sigma_2^{-1/2} \left( \frac{1}{A^\top Z_1^\top Z_1 A + Z_2^\top Z_2} \right)^2 \Sigma_2^{-1/2} \\
&\stackrel{d}{=} \Sigma_2^{-1/2} V \left( \frac{1}{\Lambda Z_1^\top Z_1 \Lambda + Z_2^\top Z_2} \right)^2 V^\top \Sigma_2^{-1/2},
\end{aligned}$$

where

$$A := \Sigma_1^{1/2} \Sigma_2^{-1/2} = U \Lambda V^\top, \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_p). \quad (7.3)$$

Using

$$\left( \frac{1}{\Lambda Z_1^\top Z_1 \Lambda + Z_2^\top Z_2} \right)^2 = \frac{d}{dz} \Big|_{z=0} \frac{1}{\Lambda Z_1^\top Z_1 \Lambda + Z_2^\top Z_2 - z},$$

we need to study the resolvent of

$$G(z) = (\Lambda Z_1^\top Z_1 \Lambda + Z_2^\top Z_2 - z)^{-1}.$$

Its local law can be studied as in previous subsection (be careful we need to switch the roles of  $Z_1$  and  $Z_2$ ). More precisely, we have that

$$G(z) \approx \text{Diag} \left( \frac{1}{-z(1 + m_3(z) + \lambda_i^2 m_4(z))} \right)_{1 \leq i \leq p} = \frac{1}{-z(1 + m_3(z) + \Lambda^2 m_4(z))}.$$

Here  $m_{3,4}(z)$  satisfy the following self-consistent equations

$$\frac{n_2}{n} \frac{1}{m_3} = -z + \frac{1}{n} \sum_{i=1}^p \frac{1}{1 + m_3 + \lambda_i^2 m_4}, \quad \frac{n_1}{n} \frac{1}{m_4} = -z + \frac{1}{n} \sum_{i=1}^p \frac{\lambda_i^2}{1 + m_3 + \lambda_i^2 m_4}. \quad (7.4)$$

Then we calculate the derivatives of  $u_3 := zm_3$  and  $u_4 := zm_4$  with respect to  $z$ :

$$\frac{n_2}{n} \frac{1}{u_3^2} u_3' = \frac{1}{n} \sum_{i=1}^p \frac{1 + u_3' + \lambda_i^2 u_4'}{(z + u_3 + \lambda_i^2 u_4)^2}, \quad \frac{n_1}{n} \frac{1}{u_4^2} u_4' = \frac{1}{n} \sum_{i=1}^p \frac{\lambda_i^2 (1 + u_3' + \lambda_i^2 u_4')}{(z + u_3 + \lambda_i^2 u_4)^2}. \quad (7.5)$$

We can solve the above equations to get  $u_3'$  and  $u_4'$ . Then we have

$$G^2(z) \approx \frac{1 + u_3'(z) + \Lambda^2 u_4'(z)}{(z + u_3(z) + \Lambda^2 u_4(z))^2}$$

in certain sense.

$$a_3 \rightarrow a_2, a_4 \rightarrow a_1$$

We now simplify the expressions for  $z \rightarrow 0$  case. When  $z \rightarrow 0$ , we shall have

$$u_3(z) = -a_3 + O(z), \quad u_4(z) = -a_4 + O(z), \quad a_3, a_4 > 0.$$

For  $z \rightarrow 0$ , the equations in (6.4) are reduced to

$$\frac{n_2}{n} \frac{1}{a_2} = 1 + \frac{1}{n} \sum_{i=1}^p \frac{1}{a_2 + \lambda_i^2 a_1}, \quad \frac{n_1}{n} \frac{1}{a_1} = 1 + \frac{1}{n} \sum_{i=1}^p \frac{\lambda_i^2}{a_2 + \lambda_i^2 a_1}. \quad (7.6)$$

It is easy to see that these equations are equivalent to

$$a_1 + a_2 = 1 - \gamma_n, \quad a_1 + \frac{1}{n} \sum_{i=1}^p \frac{a_1}{a_1 + a_2 / \lambda_i^2} = \frac{n_1}{n}. \quad (7.7)$$

The equations in (6.5) reduce to

$$\begin{aligned} \left( \frac{n_2}{n} \frac{1}{a_2^2} - \frac{1}{n} \sum_{i=1}^p \frac{1}{(a_2 + \lambda_i^2 a_1)^2} \right) a_3 - \left( \frac{1}{n} \sum_{i=1}^p \frac{\lambda_i^2}{(a_2 + \lambda_i^2 a_1)^2} \right) a_4 &= \frac{1}{n} \sum_{i=1}^p \frac{1}{(a_2 + \lambda_i^2 a_1)^2}, \\ \left( \frac{n_1}{n} \frac{1}{a_1^2} - \frac{1}{n} \sum_{i=1}^p \frac{\lambda_i^4}{(a_2 + \lambda_i^2 a_1)^2} \right) a_4 - \left( \frac{1}{n} \sum_{i=1}^p \frac{\lambda_i^2}{(a_2 + \lambda_i^2 a_1)^2} \right) a_3 &= \frac{1}{n} \sum_{i=1}^p \frac{\lambda_i^2}{(a_2 + \lambda_i^2 a_1)^2}. \end{aligned} \quad (7.8)$$

where we denote  $b_3 := u_3'(0)$  and  $b_4 := u_4'(0)$ . Thus we have

$$\left( \frac{1}{\Lambda Z_1^\top Z_1 \Lambda + Z_2^\top Z_2} \right)^2 = G^2(0) \approx \frac{1 + b_3 + \Lambda^2 b_4}{(a_3 + \Lambda^2 a_4)^2}.$$

This gives that

$$\begin{aligned} \frac{1}{X_1^\top X_1 + X_2^\top X_2} \Sigma_2 \frac{1}{X_1^\top X_1 + X_2^\top X_2} &= \Sigma_2^{-1/2} V \left( \frac{1}{\Lambda Z_1^\top Z_1 \Lambda + Z_2^\top Z_2} \right)^2 V^\top \Sigma_2^{-1/2} \\ &\approx \Sigma_2^{-1/2} V \frac{1 + b_3 + \Lambda^2 b_4}{(a_3 + \Lambda^2 a_4)^2} V^\top \Sigma_2^{-1/2} = \Sigma_2^{-1/2} \frac{1 + b_3 + b_4 A^\top A}{(a_3 + a_4 A^\top A)^2} \Sigma_2^{-1/2}. \end{aligned} \quad (7.9)$$

Hence we have

$$\begin{aligned} (\beta_s - \beta_t)^\top X_1^\top X_1 \frac{1}{X_1^\top X_1 + X_2^\top X_2} \Sigma_2 \frac{1}{X_1^\top X_1 + X_2^\top X_2} X_1^\top X_1 (\beta_s - \beta_t) \\ \approx (\beta_s - \beta_t)^\top \Sigma_1 \Sigma_2^{-1/2} \frac{1 + a_3 + a_4 A^\top A}{(a_2 + a_1 A^\top A)^2} \Sigma_2^{-1/2} \Sigma_1 (\beta_s - \beta_t). \end{aligned}$$

**Proof of Lemma 3.9.**

*Proof.* (revise the following proof) We assume that  $\Sigma_1^{-1/2}\Sigma_2$  has eigendecomposition

$$\Sigma_1^{-1/2}\Sigma_2^{1/2} = ODO^T, \quad D = \text{diag}(d_1, \dots, d_p). \quad (7.10)$$

Then by the rotational invariance of Gaussian matrices, we have

$$\tilde{Q} \stackrel{d}{=} \Sigma_1^{1/2} O \tilde{Q} O^T \Sigma_1^{1/2}, \quad \tilde{Q} := X_1^T X_1 + D X_2^T X_2 D.$$

Thus we study the spectrum of  $\tilde{Q}$  instead. We define  $\mathcal{G}(z) := (\tilde{Q} - z)^{-1}$  for  $z \in \mathbb{C}_+$ . With some random matrix tools, we have that

$$\mathcal{G}(z) \approx \text{Diag} \left( \frac{1}{-z(1 + m_3(z) + d_i^2 m_4(z))} \right)_{1 \leq i \leq p} = \frac{1}{-z(1 + m_3(z) + D^2 m_4(z))}$$

in certain sense. Here  $m_{3,4}(z)$  satisfy the following self-consistent equations

$$\frac{n_1}{n} \frac{1}{m_3} = -z + \frac{1}{n} \sum_{i=1}^p \frac{1}{1 + m_3 + d_i^2 m_4}, \quad \frac{n_2}{n} \frac{1}{m_4} = -z + \frac{1}{n} \sum_{i=1}^p \frac{d_i^2}{1 + m_3 + d_i^2 m_4} \quad (7.11)$$

When  $z \rightarrow 0$ , we shall have

$$m_3(z) = -\frac{a_3}{z} + O(1), \quad m_4(z) = -\frac{a_4}{z} + O(1), \quad a_3, a_4 > 0.$$

Then for  $z \rightarrow 0$ , the equations in (6.11) are reduced to

$$\frac{n_1}{n} \frac{1}{a_3} = 1 + \frac{1}{n} \sum_{i=1}^p \frac{1}{a_3 + d_i^2 a_4}, \quad \frac{n_2}{n} \frac{1}{a_4} = 1 + \frac{1}{n} \sum_{i=1}^p \frac{d_i^2}{a_3 + d_i^2 a_4}. \quad (7.12)$$

First, it is easy to see that these equations are equivalent to

$$a_3 + a_4 = 1 - \gamma_n, \quad a_3 + \frac{1}{n} \sum_{i=1}^p \frac{a_3}{a_3 + d_i^2 [(1 - \gamma_n) - a_3]} = c_n. \quad (7.13)$$

Furthermore, we have

$$\begin{aligned} \text{Tr}(Q^{-1}) &= \lim_{z \rightarrow 0} \text{Tr} \left[ \Sigma_1^{-1/2} O \mathcal{G}(z) O^T \Sigma_1^{-1/2} \right] = \text{Tr} \left[ \Sigma_1^{-1/2} O \left( \frac{1}{a_3 + D^2 a_4} \right) O^T \Sigma_1^{-1/2} \right] \\ &= \text{Tr} \left[ \Sigma_1^{-1/2} \frac{1}{a_3 + \Sigma_1^{-1} \Sigma_2 a_4} \Sigma_1^{-1/2} \right] = \text{Tr} \left[ \frac{1}{a_3 \Sigma_1 + a_4 \Sigma_2} \right]. \end{aligned}$$

□