Representation Transfer Learning via Multiple Pre-trained models for Linear Regression

Navjot Singh and Suhas Diggavi

Department of Electrical and Computer Engineering,
University of California, Los Angeles, USA
navjotsingh@ucla.edu, suhas@ee.ucla.edu

***Abstract*—In this paper, we consider the problem of learning a linear regression model on a data domain of interest (target) given few samples. To aid learning, we are provided with a set of pre-trained regression models that are trained on potentially different data domains (sources). Assuming a representation structure for the data generating linear models at the sources and the target domains, we propose a representation transfer based learning method for constructing the target model. The proposed scheme is comprised of two phases: (i) utilizing the different source representations to construct a representation that is adapted to the target data, and (ii) using the obtained model as an initialization to a fine-tuning procedure that re-trains the entire (over-parameterized) regression model on the target data. For each phase of the training method, we provide excess risk bounds for the learned model compared to the true data generating target model. The derived bounds show a gain in sample complexity for our proposed method compared to the baseline method of not leveraging source representations when achieving the same excess risk, therefore, theoretically demonstrating the effectiveness of transfer learning for linear regression.**

I. Introduction

A critical challenge for Deep Learning applications is the scarcity of available labeled data to train large scale models that generalize well to the data distribution. This is captured under the framework of *Few-Shot Learning* where Transfer Learning has emerged as an attractive framework to address this issue [1]. In transfer learning, one typically has access to a model trained on some data domain (hereby called *source domain*) that can be adapted to the data domain of interest (*target domain*). Within this context, a recently proposed strategy is that of *representation transfer learning* [2], [3], where one typically assumes a shared structure between the source and target learning tasks. The idea is to then learn a feature mapping for the underlying model (e.g. Neural Network representations) using the sample rich source domain that can be utilized directly on the target domain, for e.g, by training a few layers on top of the obtained network representation. This adaptation utilizes much fewer samples than what is required for training the entire model from scratch, while achieving good generalization performance which has been empirically observed for various large-scale machine learning application including image, speech and language [1], [4], [5] tasks.

A defining factor in the need for representation transfer methods is that the source and target domains have different distributions. Learning across different domains has been studied extensively in the context of *Domain Adaptation* (see for e.g. [6], [7]) where it is usually assumed that source and target domain data can be accessed simultaneously. However, in many important practical scenarios of interest, the target data samples (labeled or unlabeled) are not available when training the source models. Transferring the source dataset to the target deployment scenario is infeasible for modern large-scale applications and violates data privacy. Thus there has been an increasing interest in transferring pre-trained source models to the target domain for sample efficient learning.

Despite the immense empirical success of representation transfer learning, development of a theory for understanding the generalization of representation learning methods and the sample complexity requirements is still in its infancy. Recent efforts in this direction have been made in understanding generalization for the simpler case of linear models [8]–[11]. Within these works, [8], [10], [11] consider a common low-dimensional representation in the data generation process for the source and target domains, while [9] allows for the general case of data-generating representations being different. However, the analysis presented in that work requires the number of samples at the target to scale with the dimension of the model (see [9, Theorem 3.1]), which is impractical for few-shot learning scenarios.

A related line of work for understanding generalization of large scale models in the small sample regime is through the lens of *benign overfitting*. This is inspired by the surprising (empirical) observation that many large models, even when they overfit, tend to generalize well on the data distribution [12], [13]. In this context, [14]–[16] study this phenomena for linear models and analyze the generalization properties of the *min-norm solution*, where optimization methods like Gradient Descent are known to converge to in this setting [17], [18]. Specifically, these works seek to understand how the data distribution affects the excess population risk of the min-norm solution relative to the true data-generating linear model.

In this paper, we make efforts to understanding the generalization of linear models while leveraging pre-trained models inspired by the notions of representation transfer learning and benign-overfitting discussed above. These ideas lend themselves organically to the construction of a sample efficient training method for the target which we describe below briefly, along with our contributions.

Key Contributions: Our work provides a method for leveraging multiple pre-trained models for linear regression objectives (of dimension d) on a target task of interest in the small sample regime (samples $n_T \ll d$). The proposed two-phase approach leverages representation transfer (Phase 1) and over-parameterized training (Phase 2) to construct the target model, and we provide theoretical bounds for the excess risk for each phase of the training process (Theorem 1 and Theorem 2). In particular, we show that the learned model after the first phase has an excess risk of $\mathcal{O}(q/n_T)+\epsilon$, where q is dimension of the subspace spanned by learned source representations and ϵ is a constant that captures the approximation error when utilizing source representations for the target model (c.f. Assumption 1). This provides a gain in sample complexity compared to the baseline $\mathcal{O}(d/n_T)$ when learning the target model from scratch when the given source representations span a subspace of dimension much smaller than d (i.e. $q \ll d$). For the case when all representations are the same ($\epsilon = 0$), we recover the result of [8] for a *single* common representation. Similarly, for the overall model obtained after the second phase, we provide conditions on the target data distribution and the source/target representations that lead to an excess risk much smaller than $\mathcal{O}(d/n_T)$. Thus, we theoretically demonstrate the benefit of leveraging pre-trained models for linear regression.

A. Related Work

The problem of learning with few samples has been studied under the framework of *Few-Shot learning*, where *Meta-learning*– using experiences from previously encountered learning tasks to adapt to new tasks quickly [19], and *Transfer-learning*– transferring model parameters and employing pre-training or fine-tuning methods [3], are two major approaches. Theoretical works on Meta-learning algorithms typically assume some relation between the distribution of source and target tasks, for e.g., being sampled from the same task distribution. A more general framework is that of *Out-of-Domain (OOD)* generalization, where the goal is to learn models in a manner that generalize well to unseen data domains [20].

Transfer learning, especially, representation transfer learning has shown empirical success for large-scale machine learning [2], however, theoretical works on understanding generalization in this setting are few; see [11], [21], [22]. A related line of work is representation learning in context of *Domain Adaptation (DA)*, see for e.g. [23]–[26]. However, this usually assumes that source and data domains can be accessed simultaneously. There are deviations from this theme in *Multi-Source DA* where the goal is to understand how multiple source models can be combined to generalize well on a target domain of interest, although without changing the learned model based on the target samples [27]–[29].

In context of leveraging pre-trained models for linear regression, our work is most closely related to [8], [9] that theoretically analyze representation transfer for linear models. In contrast to [8], we allow for the true target representations to be different among the source models as well as the target, and introduce a notion of closeness between these representations (c.f. Assumption 1). Although a similar setting was considered in [9] where representations are assumed to be close in the ℓ_2 norm, their resulting bound for the fine-tuned model risk shows that the required number of target samples scale with the dimension of the learned model for efficient transfer [9, Theorem 3.1]. In contrast, the proposed method in our work provides analysis relating these bounds to the properties of the target data distribution taking inspiration from works on benign overfitting for linear regression [14], [18], [30]. This enables us to identify conditions on the target data distribution that allow the required target samples to be much smaller than the overall model dimension (see Theorem 2).

B. Paper Organization

We set up the problem and define the notation we use throughout the paper in Section II. In Section III we describe our training method for the target task model when given access to multiple pre-trained source models. Section IV establishes excess risk bounds of our proposed scheme and we provide concluding remarks in Section V. Omitted proof details can be found in the arxiv version of the paper.

II. Problem Setup and Notation

Notation: We use boldface for vectors and matrices, with matrices in uppercase. For a matrix $\mathbf{A}$, we denote the projection matrix onto its column space by $\mathbf{P}_{\mathbf{A}} := \mathbf{A}(\mathbf{A}^\top\mathbf{A})^\dagger\mathbf{A}^\top$ where $\mathbf{W}^\dagger$ denotes the Moore-Penrose pseudo-inverse of the matrix $\mathbf{W}$. We define $\mathbf{P}_{\mathbf{A}}^\perp := \mathbf{I} - \mathbf{P}_{\mathbf{A}}$, where $\mathbf{I}$ denotes the identity matrix of appropriate dimensions. We denote by $\mathcal{C}(\mathbf{A})$ the column space of a matrix $\mathbf{A}$ and by $\sigma_i(\mathbf{A})$, $\lambda_i(\mathbf{A})$ its i^{th} largest singular value and eigenvalue, respectively. $\|.\|_F$ denotes the Frobenius norm. For a vector $\mathbf{v}$, $\|\mathbf{v}\|_2$ denotes the ℓ_2 norm, while for a matrix $\mathbf{V}$, $\|\mathbf{V}\|_2$ denotes the spectral norm. $\mathrm{Tr}[\,.\,]$ denotes the trace operation. $\lesssim$ denotes the inequality sign where we ignore the constant factors. The notation $\mathcal{O}$ is the 'big-O' notation and we define $[m] = \{1, 2, \ldots, m\}$.

Setup: We consider m number of source tasks and a single target task. We denote by $\mathcal{X} \subseteq \mathbb{R}^d$ the space of inputs and $\mathcal{Y} \subseteq \mathbb{R}$ the output space. The source and target tasks are associated with data distributions p_i, $i \in [m]$ and p_T, respectively, over the space $\mathcal{X}$. We assume a linear relationship between the input and output pairs for source task $i \in [m]$ given by:

$$y_i = \mathbf{x}_i^\top \mathbf{B}_i^* \mathbf{w}_i^* + z_i, \quad \boldsymbol{\theta}_i^* := \mathbf{B}_i^* \mathbf{w}_i^* \tag{1}$$

where $\mathbf{x}_i \in \mathcal{X}$ denotes an input feature vector drawn from distribution p_i, $y_i \in \mathcal{Y}$ is the output, and $z_i \sim \mathcal{N}(0, \sigma^2)$ denotes Gaussian noise. The associated true task parameter $\boldsymbol{\theta}_i^* := \mathbf{B}_i^*\mathbf{w}_i^*$ is comprised of the representation matrix $\mathbf{B}_i^* \in \mathbb{R}^{d\times k}$ which maps the input to a lower $k-$dimensional space (where $k \ll d$) and a head vector $\mathbf{w}_i^* \in \mathbb{R}^k$ mapping the intermediate sample representation to the output. The data generation process for the target task is defined similarly with distribution p_T and associated target task parameter given by $\boldsymbol{\theta}_T^* = \mathbf{B}_T^*\mathbf{w}_T^*$. For sources $i \in [m]$, we define the input covariance matrix $\boldsymbol{\Sigma}_i = \mathbb{E}_{\mathbf{x}_i\sim p_i}[\mathbf{x}_i\mathbf{x}_i^\top]$ and similarly for the target distribution, $\boldsymbol{\Sigma}_T = \mathbb{E}_{\mathbf{x}_T\sim p_T}[\mathbf{x}_T\mathbf{x}_T^\top]$.

In our scenario of interest, the pre-trained models are trained 'offline' on source distributions and are made available to the target task during deployment. That is, for training the target task, we have access to only the models learned by the source tasks and not the source datasets themselves. For learning the pre-trained source models, we assume n_S number of samples for each of the source tasks (thus, mn_S source task samples in total) denoted by the pair $(\mathbf{X}_i, \mathbf{y}_i)$ for source $i \in [m]$ where $\mathbf{X}_i \in \mathbb{R}^{n_S \times d}$ contains row-wise input feature vectors and $\mathbf{y}_i \in \mathbb{R}^{n_S}$ is the vector of corresponding outputs. We similarly have n_T samples $(\mathbf{X}_T, \mathbf{y}_T)$ for the target task where $n_T \ll n_S$. We also assume $n_T \ll d$.

With the data generation process defined above, we now define the expected population risk on the target distribution for $\hat{\boldsymbol{\theta}}$:

$$R(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{x}\sim p_T}\mathbb{E}_{\mathbf{y}|\mathbf{x}^\top \boldsymbol{\theta}_T^*}[(\mathbf{y} - \mathbf{x}^\top \hat{\boldsymbol{\theta}})^2]$$

Our goal is to learn a model $\hat{\boldsymbol{\theta}}$ for the target task that generalizes well to the target data distribution. Thus, we want $\hat{\boldsymbol{\theta}}$ that minimizes the *Expected Excess Risk* defined by:

$$\text{EER}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_T^*) := R(\hat{\boldsymbol{\theta}}) - R(\boldsymbol{\theta}_T^*) \tag{2}$$

Since we are given access to only $n_T \ll d$ target samples, it is infeasible to learn a predictor from scratch that performs well for the excess risk defined in (2).

To aid learning on the target, we have access to models learned on the source tasks. Specifically, the target has access to the trained source models representations $\{\widehat{\mathbf{B}}_i\}_{i=1}^m$ that are solutions of the following empirical minimization problem:

$$\{\widehat{\mathbf{B}}_i, \widehat{\mathbf{w}}_i\}_{i=1}^m \leftarrow \min_{\{\mathbf{B}_i\},\{\mathbf{w}_i\}} \frac{1}{mn_S}\sum_{i=1}^m \|\mathbf{y}_i - \mathbf{X}_i\mathbf{B}_i\mathbf{w}_i\|_2^2 \tag{3}$$

Since we have data rich source domains ($n_S \gg d$), we expect the obtained source models $\widehat{\mathbf{B}}_i\widehat{\mathbf{w}}_i$ to be close to $\boldsymbol{\theta}_i^*$ for $i \in [m]$ (c.f. Equation (1)). For effective representation transfer, we also want the representations $\{\widehat{\mathbf{B}}_i\}$ to be close to the true representations $\{\mathbf{B}_i^*\}$, in the sense that they approximately span the same subspace. We make this notion precise in Lemma 1 stated with our main results in Section IV. Given access to the source model representations, our proposed method for training the target model leverages the representation maps $\{\widehat{\mathbf{B}}_i\}_{i=1}^m$ to drastically reduce the sample complexity. We describe our training method in Section III and provide the excess risk bounds for the resulting target model in Section IV.

III. Learning with Multiple Pre-trained models

To leverage source representations for training the target model, it is instinctive that there should be a notion of closeness between the true source and target model representation that can be exploited for target task training. We now make this notion precise. We first define as $\mathbf{V}^* \in \mathbb{R}^{d\times l}$ the matrix whose columns are an orthonormal basis of the set of columns of all the source representation matrices $\{\mathbf{B}_i^*\}_{i=1}^m$. The individual source models can be represented by $\boldsymbol{\theta}_i^* = \mathbf{V}^*\widetilde{\mathbf{w}}_i^*$ for some $\widetilde{\mathbf{w}}_i^* \in \mathbb{R}^l$ for all $i \in [m]$. The target model $\boldsymbol{\theta}_T^* = \mathbf{B}_T^*\mathbf{w}_T^*$ governing the target data generation is assumed to satisfy:

Assumption 1: Consider the projection of the target model $\mathbf{B}_T^*\mathbf{w}_T^*$ to space $\mathcal{C}(\mathbf{V}^*)$ given by $\mathbf{V}^*\widetilde{\mathbf{w}}_T^*$ for some $\widetilde{\mathbf{w}}_T^* \in \mathbb{R}^l$. Then for some $\epsilon > 0$, we have:

$$\mathbb{E}_{\mathbf{x}\sim p_T}\left[\mathbf{x}^\top\mathbf{V}^*\widetilde{\mathbf{w}}_T^* - \mathbf{x}^\top\mathbf{B}_T^*\mathbf{w}_T^*\right]^2 \le \epsilon^2$$

The value of ϵ in Assumption 1 above captures how far away the output of the true target model is to a model learned using the true source representations. Note that if the columns of $\mathbf{B}_T^*$ can be constructed by the vectors in $\mathbf{V}^*$, the above is satisfied for $\epsilon = 0$. Assumption 1 can also be re-written in terms of the target covariance matrix as:

$$\left\|\Sigma_T^{1/2}(\mathbf{V}^*\widetilde{\mathbf{w}}_T^* - \mathbf{B}_T^*\mathbf{w}_T^*)\right\|_2^2 \le \epsilon^2 \tag{4}$$

We are given access to n_T samples for the target machine given by $(\mathbf{X}_T, \mathbf{y}_T)$ and pre-trained models representations from the sources $\{\widehat{\mathbf{B}}_i\}_{i=1}^m$ (c.f. Equation (3)). Our proposed training scheme consists of two phases, which we will now describe independently in the following. We split the available n_T target samples into n_{T_1}, n_{T_2} for the two respective phases. At a high level, in *Phase 1*, we make use of the available source representations to construct a target representation and adapt it to the target task using n_{T_1} samples. The obtained model is then used as an *initialization* for *Phase 2* where we train the entire (over-parameterized) model, including the representation matrix, using n_{T_2} samples. We provide the resulting excess risk bounds for the model obtained after *Phase 1* and the final target model after *Phase 2* in Section IV.

A. Phase 1: Transferring source representation to target

In the context of utilizing pre-trained models, we will make use of the empirical source representations $\{\widehat{\mathbf{B}}_i\}_{i=1}^m$ to learn the target model. We first construct a matrix $\widehat{\mathbf{V}} \in \mathbb{R}^{d\times q}$ whose columns are the orthonormal basis of the columns of $\{\widehat{\mathbf{B}}_i\}_{i=1}^m$ which denotes a dictionary of the learned source representation matrices[1]. Note that we have $q \le mk$. Having constructed the representation, we train a head vector $\widehat{\mathbf{w}}_{T_1} \in \mathbb{R}^q$ minimizing the empirical risk on n_{T_1} samples:

$$\widehat{\mathbf{w}}_{T_1} \leftarrow \min_{\mathbf{w}_T\in\mathbb{R}^q} \frac{1}{n_{T_1}}\left\|\mathbf{y}_{T_1} - \mathbf{X}_{T_1}\widehat{\mathbf{V}}\mathbf{w}_T\right\|_2^2 \tag{5}$$

Here, $\mathbf{y}_{T_1} \in \mathbb{R}^{n_{T_1}}$ denotes the first n_{T_1} values of $\mathbf{y}_T$ and $\mathbf{X}_{T_1} \in \mathbb{R}^{n_{T_1}\times d}$ the first n_{T_1} rows of $\mathbf{X}_T$. We denote the resulting model at the end of this phase by $\boldsymbol{\theta}_{\text{Phase}_1} := \widehat{\mathbf{V}}\widehat{\mathbf{w}}_{T_1}$. Since we only have to learn the head vector using the available representation $\widehat{\mathbf{V}}$, the sample complexity requirement is greatly reduced, which is also evident from our bound for $\text{EER}(\boldsymbol{\theta}_{\text{Phase}_1}, \boldsymbol{\theta}_T^*)$ provided in Theorem 1.

B. Phase 2: Fine-tuning with initialization

The obtained model $\boldsymbol{\theta}_{\text{Phase}_1}$ from the previous phase utilizes empirical source representation for its construction. However, the true target model $\boldsymbol{\theta}_T^*$ may not lie in the space spanned by

[1]The construction of $\widehat{\mathbf{V}}$ from $\{\widehat{\mathbf{B}}_i\}$ can be done by the Gram-Schmidt process. This can be done in the pre-deployment phase after training the source models, and $\widehat{\mathbf{V}}$ can be made available directly to the target task.

the source representation and thus $\boldsymbol{\theta}_{\text{Phase}_1}$ lies in a ball centered $\boldsymbol{\theta}_T^*$ whose radius scales with ϵ (c.f. Assumption 5). To move towards the true model $\boldsymbol{\theta}_T^*$, we utilize n_{T_2} number of target samples (independent from the n_{T_1} samples in the previous phase) to train the entire linear model using Gradient Descent (GD) with $\boldsymbol{\theta}_{\text{Phase}_1}$ as the initialization. In particular, the GD procedure minimizes the following starting from $\boldsymbol{\theta}_{\text{Phase}_1}$:

$$f(\boldsymbol{\theta}) = \frac{1}{n_{T_2}} \|\mathbf{y}_{T_2} - \mathbf{X}_{T_2}\boldsymbol{\theta}\|_2^2 \tag{6}$$

Here, $\mathbf{y}_{T_2} \in \mathbb{R}^{n_{T_2}}$ and $\mathbf{X}_{T_2} \in \mathbb{R}^{n_{T_2} \times d}$ are the remaining sample values from Phase 1. Since $n_{T_2} \ll d$, we are in an over-parameterized regime, for which it is known that GD procedure optimizing the objective in (6) converges, under appropriate choice of learning rate, to a solution closest in norm to the initialization [17], [18], [31]; mathematically:

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{Phase}_1}\|_2 \\ &\text{s.t. } \|\mathbf{y}_{T_2} - \mathbf{X}_{T_2}\boldsymbol{\theta}\|_2 = \min_{\mathbf{b}} \|\mathbf{y}_{T_2} - \mathbf{X}_{T_2}\mathbf{b}\|_2 \end{aligned} \tag{7}$$

We denote the solution of the above optimization problem as $\boldsymbol{\theta}_{\text{Phase}_2}$, which forms our final target task model. We provide bounds for $\text{EER}(\boldsymbol{\theta}_{\text{Phase}_2}, \boldsymbol{\theta}_T^*)$ in Theorem 2.

IV. Main Results

We now provide theoretical bounds on the excess risk for the target (c.f. Equation (2)) when leveraging pre-trained source models. In Section IV-A, we first state excess risk bounds for the model obtained after Phase 1 (see Section III-A), denoted by $\boldsymbol{\theta}_{\text{Phase}_1} := \widehat{\mathbf{V}}\widehat{\mathbf{w}}_T$, where target representation $\widehat{\mathbf{V}}$ is constructed as a combination of source representations and adapted to the target data using n_{T_1} amount of target samples by training a target-specific head vector $\widehat{\mathbf{w}}_T$. In Section IV-B, we provide our overall excess risk for the model $\boldsymbol{\theta}_{\text{Phase}_2}$ (c.f. Equation (7)) obtained by re-training the entire (over-parameterized) model via Gradient Descent with n_{T_2} number of target samples (independent form the previously utilized n_{T_1} samples) using $\boldsymbol{\theta}_{\text{Phase}_1}$ as the initialization.

A. Theoretical results for representation transfer: Phase 1

We work with the following assumptions:

Assumption 2 (Subgaussian features): We assume that $\mathbb{E}_{\mathbf{x}\sim p_j}[\mathbf{x}] = 0$ for all $j \in [m] \cup \{T\}$. We consider $\bar{p}_j$ to be the *whitening* of p_j (for $j \in [m] \cup \{T\}$) such that $\mathbb{E}_{\bar{\mathbf{x}}\sim \bar{p}_j}[\mathbf{x}] = 0$ and $\mathbb{E}_{\bar{\mathbf{x}}\sim \bar{p}_j}[\bar{\mathbf{x}}\bar{\mathbf{x}}^\top] = \mathbf{I}$. We assume there exists $\rho > 0$ such that the random vector $\bar{\mathbf{x}} \sim \bar{p}_j$ is ρ^2-subgaussian.

Assumption 3 (Covariance Dominance): There exists $r > 0$ such that $\Sigma_i \succeq r\Sigma_T$ for all $i \in [m]$.

Assumption 4 (Diverse source tasks): Consider the source models $\boldsymbol{\theta}_i^* = \mathbf{V}^*\widetilde{\mathbf{w}}_i^*$ for $i \in [m]$. We assume that the matrix $\widetilde{\mathbf{W}}^* := [\widetilde{\mathbf{w}}_1^*, \ldots, \widetilde{\mathbf{w}}_m^*] \in \mathbb{R}^{l\times m}$ satisfies $\sigma_l^2(\widetilde{\mathbf{W}}^*) \geq \Omega\big(\frac{m}{l}\big)$

Assumption 5 (Distribution of target task): We assume that $\widetilde{\mathbf{w}}_T^*$ follows a distribution ν such that $\big\|\mathbb{E}_{\widetilde{\mathbf{w}}\sim\nu}[\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top]\big\|_2$ is $\mathcal{O}\big(\frac{1}{l}\big)$. We denote $\Sigma_{\widetilde{\mathbf{w}}_T^*} = \mathbb{E}_{\widetilde{\mathbf{w}}\sim\nu}[\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top]$.

Note on Assumptions: Assumption 2 on sub-Gaussian features is commonly used in literature to obtain probabilistic tail bounds [8], [9], [14], [30]. Following [8], Assumption 3 states the target data covariance matrix is covered by the covariance matrices of the source data distributions. We remark that this assumption allows the covariance matrices to be different, in contrast to works [9], [10] that assume a common covariance matrix for all the distributions. Assumption 4 (also made in related works [8], [9], [32]) says that the head vectors corresponding to the matrix $\mathbf{V}^*$ for each source model should span $\mathbb{R}^l$. This effectively allows us to recover the representation $\mathbf{V}^*$ provided enough source machines ($m > l$) that individually capture one or more features of $\mathbf{V}^*$. This assumption is also central to proving our result in Lemma 1 provided below which show that the matrices $\widehat{\mathbf{V}}$ and $\mathbf{V}^*$, whose columns form an orthonormal basis for the span of $\{\widehat{\mathbf{B}}_i\}$ and $\{\mathbf{B}_i^*\}$, respectively, span the same subspace for constructing the target model.

Lemma 1: Let matrix $\widehat{\mathbf{V}} \in \mathbb{R}^{d\times q}$ be formed by empirical source representations $\{\widehat{\mathbf{B}}_i\}$ obtained from solving (3) and the matrix $\mathbf{V}^* \in \mathbb{R}^{d\times l}$ formed from the true representations $\{\mathbf{B}_i^*\}$. Under Assumption 2-4, for any $\mathbf{b} \in \mathbb{R}^l$ such that $\|\mathbf{b}\|_2 = 1$, $n_s \gg \rho^4(d + \log(m/\delta))$ and $n_{T_1} \gg \rho^4(\max\{l, q\} + \log(1/\delta))$, with probability at-least $1 - \delta_1$, we have:

$$\begin{aligned} &\min_{\mathbf{u}\in\mathbb{R}^q} \|\mathbf{X}_{T_1}\widehat{\mathbf{V}}\mathbf{u} - \mathbf{X}_{T_1}\mathbf{V}^*\mathbf{b}\|_2 \\ &\qquad \leq \frac{\sigma^2 n_{T_1}}{r n_S}\left(km + kdm\log(\kappa n_s) + \log\left(\frac{1}{\delta_1}\right)\right) \end{aligned}$$

We now state our main result for the excess risk on after Phase 1.

Theorem 1 (Phase 1 training result): Fix a failure probability $\delta \in (0, 1)$ and further assume $2k \leq \min\{d, m\}$ and the number of samples in the sources and target satisfy $n_s \gg \rho^4(d + \log(m/\delta))$ and $n_{T_1} \gg \rho^4(\max\{l, q\} + \log(1/\delta))$, respectively. Define $\kappa = \frac{\max_{i\in[m]}\lambda_{\max}(\Sigma_i)}{\min_{i\in[m]}\lambda_{\min}(\Sigma_i)}$ where $\lambda_{\max}(\Sigma_i)$ denotes the maximum eigenvalue of Σ_i. Then with probability at least $1 - \delta$ over the samples, under Assumptions 1 - 5, the expected excess risk of $\boldsymbol{\theta}_{\text{Phase}_1} := \widehat{\mathbf{V}}\widehat{\mathbf{w}}$ satisfies:

$$\begin{aligned} \mathbb{E}[\text{EER}(\boldsymbol{\theta}_{\text{Phase}_1}, \boldsymbol{\theta}_T^*)] &\lesssim \frac{\sigma^2}{n_{T_1}}(q + \log(1/\delta)) + \epsilon^2 \\ &\quad + \sigma^2\left[\frac{1}{r n_s m}\log\left(\frac{1}{\delta}\right) + \left(\frac{kd\log(\kappa n_s) + k}{r n_s}\right)\right] \end{aligned}$$

where expectation is taken over $\widetilde{\mathbf{w}}_T^*$ for the target task (c.f. Assumption 5).

Discussion: The bound in Theorem 1 shows the population risk of the learned model $\boldsymbol{\theta}_{\text{Phase}_1}$ lies in a ball centered at the true target model risk $R(\boldsymbol{\theta}_T^*)$ with radius ϵ^2, which represents the *approximation error* for using source representations for the target task (see Assumption 1). Note that the expected excess risk scales as $\mathcal{O}(q/n_{T_1})$ with respect to the number of target samples when the representation is learned from the source representations. This demonstrates a sample gain compared to the baseline of $\mathcal{O}(d/n_{T_1})$ for learning the entire model (including representation) with the target data when $q \ll d$, that is, when the empirical source representations together span a subspace of dimension much smaller than d. For the case when source and target representations are all the

same, $\mathbf{B}_T^* = \mathbf{B}_i^* = \mathbf{B}^* \in \mathbb{R}^{d\times k}$ for all $i \in [m]$, the excess risk scales as $\mathcal{O}(k/n_{T_1})$, which recovers the result of [8].

B. Theoretical results for overall scheme: Phase1 + Phase2

We require the following additional assumptions:

Assumption 6: The rows of the target data matrix $\mathbf{X}_T$ are linearly independent.

Assumption 7: The Gradient Descent procedure to optimize (6) converges to $\boldsymbol{\theta}_{\text{Phase}_2}$ with $f(\boldsymbol{\theta}_{\text{Phase}_2}) = 0$.

Assumption 6 is typically made in literature for analysis in the over-parameterized regime for linear regression, see [14], and can also be relaxed to hold with high probability instead and incorporated in the analysis [30]. Assumption 7 holds in our setting as the objective in (6) is strongly convex and smooth for which GD can converge to the optimum [33].

Theorem 2 (Phase 1 + Phase 2 training result): Consider obtaining the final target model by using n_{T_1} samples during Phase 1 for representation transfer and then fine-tuning in Phase 2 with n_{T_2} samples (independently drawn from Phase 1). Denote the eigenvalues of the covariance matrix of the underlying data Σ_T by $\{\lambda_i\}_{i=1}^d$. Then under Assumptions 1-7, the excess risk of the final parameter $\hat{\boldsymbol{\theta}}_T := \boldsymbol{\theta}_{\text{Phase}_2}$ is bounded as follows with probability at least $1-\delta$:

$$\begin{aligned}\mathbb{E}[\text{EER}(\hat{\boldsymbol{\theta}}_T, \boldsymbol{\theta}_T^*)] \lesssim\ & \frac{\lambda_1}{\lambda_d}\frac{r_0(\Sigma_T)}{n_{T_2}}\left(\frac{\sigma^2}{n_{T_1}}(q+\log(1/\delta)) + \epsilon^2\right) \\ & + r\sigma^2\log\left(\frac{1}{\delta}\right)\left(\frac{k^*}{n_{T_2}} + \frac{n_{T_2}}{R_{k^*}(\Sigma_T)}\right) \\ & + \frac{\lambda_1\sigma^2}{\lambda_d}\frac{r_0(\Sigma_T)}{n_{T_2}}\left(\frac{1}{rn_s m}\log\left(\frac{1}{\delta}\right) + \left(\frac{kd\log(\kappa n_s)+k}{rn_s}\right)\right)\end{aligned}$$

where $r_k(\Sigma_T) = \frac{\Sigma_{i>k}\lambda_i}{\lambda_{k+1}}$, $R_{k^*}(\Sigma_T) = \frac{(\Sigma_{i>k}\lambda_i)^2}{\Sigma_{i>k}\lambda_i^2}$. Here, constant $b > 1$ and $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn_{T_2}\}$ with $k^* \leq \frac{n_{T_2}}{c_1}$ for some universal constant $c_1 > 1$.

Discussion: Theorem 2 shows the excess risk of our overall target model ($\hat{\boldsymbol{\theta}}_T = \boldsymbol{\theta}_{\text{Phase}_2}$) as a function of the number of samples n_S, n_{T_1}, n_{T_2} and parameters depending on the target data covariance matrix, Σ_T. Since we re-train the entire model (including the representation) with n_{T_2} target samples, the population risk of the learned model $R(\boldsymbol{\theta}_{\text{Phase}_2})$ can be made closer to the true risk $R(\boldsymbol{\theta}_T^*)$ by increasing n_{T_2}, which is in contrast to the result of Theorem 1 which shows closeness only in an ϵ^2 radius ball due to using source representation directly to construct the target model.

We now provide a baseline comparison to the standard linear regression scenario where we do no utilize any source models and instead learn the target task model from scratch using the available $n_T = n_{T_1} + n_{T_2}$ samples. The excess risk in this setting is $\mathcal{O}\left(\frac{\sigma^2 d}{n_T}\right)$. If the number of source samples are large enough ($n_S \gg d$) to get a good empirical performance on the source models (c.f. Equation (3)), the bound from Theorem 2 demonstrates a sample gain compared to the baseline when:

$$\begin{aligned}&\frac{\lambda_1}{\lambda_d}\frac{r_0(\Sigma_T)}{n_{T_2}}\left(\frac{\sigma^2}{n_{T_1}}(q+\log(1/\delta)) + \epsilon^2\right) \\ &+ c\sigma^2\log\left(\frac{1}{\delta}\right)\left(\frac{k^*}{n_{T_2}} + \frac{n_{T_2}}{R_{k^*}(\Sigma_T)}\right) \ll \frac{\sigma^2 d}{n_{T_1}+n_{T_2}} \qquad (8)\end{aligned}$$

It can be seen that for the above relation to hold, we require:

- The target data covariance matrix Σ_T should be such that the term $R_{k^*}(\Sigma_T)$ is much larger than n_{T_2}, and $k^* \ll n_{T_2}$. This is satisfied, for e.g., in the case when eigenvalues of Σ_T decay slowly from largest to smallest, and are all larger than a small constant [14].
- Using the definition of $r_0(\Sigma_T) = \sum_{i=1}^d \lambda_i/\lambda_1$, and noting that $\sum_{i=1}^d \lambda_i/\lambda_d < d$, the following provides a sufficient condition the first term on the L.H.S. of (8):
$$\frac{qd}{n_{T_1}n_{T_2}} \ll \frac{d}{n_{T_1}+n_{T_2}}$$
This, is turn, imposes the following restriction on q, which is the dimension of the subspace formed by the source representations $\{\widehat{\mathbf{B}}_i\}$:
$$q \ll \frac{n_{T_1}n_{T_2}}{n_{T_1}+n_{T_2}} \qquad (9)$$
Since $n_{T_1}+n_{T_2} = n_T$, it is easy to check that the R.H.S. of (9) is maximized when $n_{T_1} = n_{T_2} = n/2$. With this optimal splitting of the target samples for each of the phases, we require $q \ll n_T/2$ for the inequality in (8).

V. Conclusion

In this work, we proposed a method for training linear regression models via representation transfer learning in the limited sample regime, when given access to multiple pre-trained linear models trained on data domains (sources) different form the target of interest. We established excess risk bounds for the learned target model when (i) source representations are used directly to construct a target representation and adapted to the target task, and (ii) when the entire resulting model is fine-tuned in the over-parameterized regime using target task samples. Our bounds show a gain in target sample complexity compared to the baseline case of learning without access to the pre-trained models, thus demonstrating the benefit of transfer learning for better generalization in the limited sample regime.

As future extensions to this work, it is of interest to see how non-linear activation functions can be introduced in the model to analyze more complicated architectures like Neural Networks (NNs). Analyzing representation transfer learning with multiple NNs and utilizing recently developed results in benign over-fitting for this setting [34] is an interesting next step. In many scenarios of interest, for training the source task models, *unlabeled* data from the target distribution might be available. While there are empirical works utilizing unlabeled samples in the context of semi-supervised adaptation ([26], [35], [36]), theoretical results on understanding generalization of representation transfer learning methods (with pre-training/fine-tuning) and their sample complexity requirements are missing and would be an interesting direction to pursue.

Acknowledgments

This work was supported in part by NSF grants 2139304, 2146838 and 2007714, and Army Research Laboratory grant under Cooperative Agreement W911NF-17-2-0196.

REFERENCES

[1] Q. S. Y. Liu, T. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings, 2012, pp. 17–36.

[4] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[5] Z. Liu, Y. Lin, and M. Sun, *Representation learning for natural language processing*. Springer Nature, 2020.

[6] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning bounds for domain adaptation," *Advances in neural information processing systems*, vol. 20, 2007.

[7] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," in *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. [Online]. Available: http://www.cs.mcgill.ca/\%7Ecolt2009/papers/003.pdf\#page=1

[8] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei, "Few-shot learning via learning the representation, provably," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=pW2Q2xLwIMD

[9] K. Chua, Q. Lei, and J. D. Lee, "How fine-tuning allows for effective meta-learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8871–8884, 2021.

[10] N. Tripuraneni, C. Jin, and M. Jordan, "Provable meta-learning of linear representations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 434–10 443.

[11] A. Maurer, M. Pontil, and B. Romera-Paredes, "The benefit of multitask representation learning," *Journal of Machine Learning Research*, vol. 17, no. 81, pp. 1–32, 2016.

[12] M. Belkin, D. J. Hsu, and P. Mitra, "Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate," *Advances in neural information processing systems*, vol. 31, 2018.

[13] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 322–332.

[14] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 063–30 070, 2020.

[15] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *The Annals of Statistics*, vol. 50, no. 2, pp. 949–986, 2022.

[16] V. Shah, S. Basu, A. Kyrillidis, and S. Sanghavi, "On generalization of adaptive methods for over-parameterized linear regression," *arXiv preprint arXiv:2011.14066*, 2020.

[17] J. Wu, D. Zou, V. Braverman, and Q. Gu, "Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=3X64RLgzY6O

[18] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Characterizing implicit bias in terms of optimization geometry," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1832–1841.

[19] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[20] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[21] D. McNamara and M.-F. Balcan, "Risk bounds for transferring representations with and without fine-tuning," in *International conference on machine learning*. PMLR, 2017, pp. 2373–2381.

[22] T. Galanti, L. Wolf, and T. Hazan, "A theoretical framework for deep transfer learning," *Information and Inference: A Journal of the IMA*, vol. 5, no. 2, pp. 159–209, 2016.

[23] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon, "On learning invariant representations for domain adaptation," in *International conference on machine learning*. PMLR, 2019, pp. 7523–7532.

[24] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, vol. 20, 2007.

[25] P. Stojanov, Z. Li, M. Gong, R. Cai, J. Carbonell, and K. Zhang, "Domain adaptation with invariant representation learning: What transformations to learn?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 791–24 803, 2021.

[26] Y. Zhang, T. Liu, M. Long, and M. Jordan, "Bridging theory and algorithm for domain adaptation," in *International conference on machine learning*. PMLR, 2019, pp. 7404–7413.

[27] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation with multiple sources," *Advances in neural information processing systems*, vol. 21, 2008.

[28] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6028–6039.

[29] S. M. Ahmed, D. S. Raychaudhuri, S. Paul, S. Oymak, and A. K. Roy-Chowdhury, "Unsupervised multi-source domain adaptation without access to source data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 103–10 112.

[30] G. Shachaf, A. Brutzkus, and A. Globerson, "A theoretical analysis of fine-tuning with linear teachers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 382–15 394, 2021.

[31] M. Belkin, D. Hsu, and J. Xu, "Two models of double descent for weak features," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 4, pp. 1167–1180, 2020.

[32] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2089–2099.

[33] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[34] N. S. Chatterji, P. M. Long, and P. L. Bartlett, "The interplay between implicit bias and benign overfitting in two-layer linear networks," *Journal of machine learning research*, vol. 23, no. 263, pp. 1–48, 2022.

[35] A. Singh, "Clda: Contrastive learning for semi-supervised domain adaptation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5089–5101, 2021.

[36] S. Mishra, K. Saenko, and V. Saligrama, "Surprisingly simple semi-supervised domain adaptation with pretraining and consistency," in *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

[37] V. Koltchinskii and K. Lounici, "Concentration inequalities and moment bounds for sample covariance operators," *Bernoulli*, vol. 23, no. 1, pp. 110–133, 2017.