Statistical Minimax Lower Bounds for Transfer Learning in Linear Binary Classification

Seyed Mohammadreza Mousavi Kalan, Mahdi Soltanolkotabi, and A. Salman Avestimehr
Ming Hsieh Department of Electrical Engineering, University of Southern California, CA, USA
Email: mmousavi@usc.edu, soltanol@usc.edu, avestimehr@ee.usc.edu

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834760

***Abstract*—Modern machine learning models require a large amount of labeled data for training to perform well. A recently emerging paradigm for reducing the reliance of large model training on massive labeled data is to take advantage of abundantly available labeled data from a related *source* task to boost the performance of the model in a desired *target* task where there may not be a lot of data available. This approach, which is called transfer learning, has been applied successfully in many application domains. However, despite the fact that many transfer learning algorithms have been developed, the fundamental understanding of "when" and "to what extent" transfer learning can reduce sample complexity is still limited. In this work, we take a step towards foundational understanding of transfer learning by focusing on binary classification with linear models and Gaussian features and develop statistical minimax lower bounds in terms of the number of source and target samples and an appropriate notion of similarity between source and target tasks. To derive this bound, we reduce the transfer learning problem to hypothesis testing via constructing a packing set of source and target parameters by exploiting Gilbert-Varshamov bound, which in turn leads to a lower bound on sample complexity. We also evaluate our theoretical results by experiments on real data sets.**

I. Introduction

Modern machine learning models have achieved unprecedented success in numerous applications spanning computer vision to natural language processing. Most of these models consist of millions of parameters which require an abundance of labeled data for training. In many applications however due to scarcity of data training models that also generalize well is challenging. Yet another challenge is that these models do not adapt well to new environments. In particular, their performance degrades with modest changes in the data set and they may require as much data as training from scratch in the new environment.

Transfer learning is a recent promising approach to tackle the aforementioned challenges by effectively utilizing the samples of a different but related *source* task, where there are typically many labeled samples, in order to improve the performance of the model on a *target* task with only a few available labeled samples for training. Indeed, in modern deep learning literature such transfer learning approaches that use pretrained models and fining tuning have enjoyed wide empirical success [1]. Nevertheless, fundamental limits and benefits of transfer learning have not been well understood and many key questions remain unanswered. How can we measure the similarity of two tasks to decide whether they are appropriate for transfer learning? Given access to a limited number of samples what would be the best possible accuracy we can achieve using any algorithms regardless of the computational complexity? How can we characterize the generalization error of the target task as a function of the number of source and target samples as well as a measure of similarity between them?

In this paper, we focus on answering these questions for linear models. This serves as a stepping stone for more general models and provides guidelines for development of more effective transfer learning algorithms. To this aim, we first define a measure to quantitatively capture the similarity distance of different tasks. We then derive statistical minimax lower bounds for binary classification with Gaussian features as a function of source and target samples as well as the measure of similarity. Our lower bounds consist of different regimes. When the distance of source and target tasks is high the corresponding lower bound is only a function of the target samples indicating that the target error is determined by the number of target samples and source samples are useful only up to a point. On the other hand, in the regime where the distance of source and target is low the target error depends on both the number of source and target samples which demonstrate that source samples are useful when source is similar to target. Finally, we perform various experiments on real data sets to corroborate our theoretical findings and investigate the utility of the measure defined in this paper in practical scenarios.

II. Prior Works

Related to transfer learning, [2]–[8] study domain adaptation problem where the goal is to adapt the hypothesis learned on the source domain for the target domain to achieve small target generalization error. The common assumption in domain adaptation is that the source and target share the same labeling function which is denoted as covariate shift, and the marginal distributions have a small shift under an appropriate notion of similarity measure. There are numerous results in this literature that provide sufficiency results by finding upper bounds on the target generalization error. These upper bounds guarantee that some types of algorithms achieve a small target generalization error not exceeding a threshold. [9], [10] first introduce a similarity measure of source and target which can be estimated by a finite number of unlabeled source and target samples. Then, they provide an upper bound for the target generalization error of a hypothesis in terms of the error of the hypothesis in the source domain as well as the distance of source and target using the introduced measure. [11] introduces a new discrepancy distance utilizing Radamacher complexity which extends the results of [9], [10] for a broad family of loss functions. More recently, [12] develops novel algorithms that achieve near optimal minimax risk in linear regression for two scenarios: 1- source and target share the same conditional distribution which is also denoted as covariate shift 2- Marginal distributions of the source and target are the same which is denoted as model shift.

There are also a few results providing lower bounds for the target generalization error and necessary results for successful learning. [13] studies the assumptions of covariate shift, existence of a joint optimal hypothesis, as well as similarity of distributions in the domain adaptation problem and provides scenarios where one cannot guarantee successful learning in a PAC-style learning model. [14] by defining a new discrepancy measure, called *transfer exponent*, derives a minimax lower bound for target generalization error. [14] makes a relaxed covariate-shift assumption as well as Bernstein Class Condition assumption on label noise. More closely related to this work, [15] derives minimax lower bounds for transfer learning with one-hidden layer neural networks for regression. [15] unlike the most of results in this literature does not make the covariate shift assumption. Our paper without making the covariate shift assumption derives minimax lower bounds for classification with linear models.

III. Problem Formulation

In this section we formalize the transfer learning in binary classification. First we introduce the model considered in this paper and then describe the minimax framework to derive the desired lower bounds.

A. Transfer Learning model

We consider a problem where there are n_S and n_T number of labeled samples from a source and a target domain. Specifically, we denote the labeled source and target samples by $(\boldsymbol{x}_S, y_S) \sim \mathbb{P}$ and $(\boldsymbol{x}_T, y_T) \sim \mathbb{Q}$ where $\boldsymbol{x}_S, \boldsymbol{x}_T \in \mathbb{R}^d$ as well as $y_S, y_T \in \{-1, 1\}$ denote the features/inputs and labels/outputs. Moreover, $\mathbb{P}$ and $\mathbb{Q}$ denote the underlying joint distributions of the source and target samples. We also assume that features are generated by normal distributions, $\boldsymbol{x}_S, \boldsymbol{x}_T \sim \mathcal{N}(0, I_d)$, and the distribution of the labels are as follows

$$\text{Prob}(y_S = 1|\boldsymbol{x}_S) = \frac{1}{1 + \exp(-\boldsymbol{x}_S^T \boldsymbol{\theta}_S)} \tag{1}$$

$$\text{Prob}(y_T = 1|\boldsymbol{x}_T) = \frac{1}{1 + \exp(-\boldsymbol{x}_T^T \boldsymbol{\theta}_T)} \tag{2}$$

where $\boldsymbol{\theta}_S, \boldsymbol{\theta}_T \in \mathbb{R}^d$ are the ground truth parameters of the source and target tasks. Then, the optimal Bayes classifier of the target is as follows

$$C_{\theta_T}(\boldsymbol{x}_T) = \begin{cases} 1 & \text{if } \boldsymbol{x}_T^T \boldsymbol{\theta}_T \geq 0 \\ -1 & \text{o.w.} \end{cases} \tag{3}$$

In a transfer learning problem we aim at finding $\hat{\theta}_T$ by exploiting both the source and target samples such that the corresponding classifier, i.e. $C_{\hat{\boldsymbol{\theta}}_T}$, is close to the optimal Bayes classifier by the following risk

$$\text{Prob}\big\{C_{\hat{\boldsymbol{\theta}}_T}(\boldsymbol{x}_T) \neq C_{\boldsymbol{\theta}_T}(\boldsymbol{x}_T)\big\}$$

Note that the Bayes and estimated classifiers do not depend on the magnitude of $\boldsymbol{\theta}_T$ and $\hat{\boldsymbol{\theta}}_T$. Therefore, without loss of generality we can assume that they lie on the unit sphere in $\mathbb{R}^d$.

B. Minimax Framework

In order to develop a minimax framework for the transfer learning problem, we need to define a class of transfer learning problems consisting of pairs of source and target tasks. Here we denote each pair of source and target tasks by $(\mathbb{P}_{\boldsymbol{\theta}_S}, \mathbb{Q}_{\boldsymbol{\theta}_T})$ parametrized by $\boldsymbol{\theta}_S$ as well as $\boldsymbol{\theta}_T$, and the distributions $\mathbb{P}_{\boldsymbol{\theta}_S}$ and $\mathbb{Q}_{\boldsymbol{\theta}_T}$ denote the joint distributions of features and labels over the source and target, i.e. $(\boldsymbol{x}_S, y_S) \sim \mathbb{P}_{\boldsymbol{\theta}_S}$ and $(\boldsymbol{x}_T, y_T) \sim \mathbb{P}_{\boldsymbol{\theta}_T}$. In a transfer learning problem we use the source and target samples to find an estimate $\hat{\boldsymbol{\theta}}_T$ of $\boldsymbol{\theta}_T$. In other words, $\hat{\boldsymbol{\theta}}_T$ is a function of source and target samples. In a minimax

framework, the target parameter, $\boldsymbol{\theta}_T$, is chosen in an adversarial manner, and we are interested in minimizing the risk $\sup \mathbb{E}_{\substack{\text{source and target}\\ \text{samples}}}\Big[\text{Prob}\big\{C_{\hat{\boldsymbol{\theta}}_T}(\boldsymbol{x}_T) \neq C_{\boldsymbol{\theta}_T}(\boldsymbol{x}_T)\big\}\Big]$ where the supremum is taken over an appropriate class of source and target tasks within a distance to reflect the difficulty of transfer learning. In order to define the classes of source and target tasks, we need to define an appropriate notion of transfer distance between them. First we state the following proposition regarding the considered risk for measuring the performance of the estimate of the target parameter.

Proposition 1. *Let $\boldsymbol{\theta}_T$ be the target parameter in Equation* (2) *and $C_{\boldsymbol{\theta}_T}$ be the optimal Bayes classifier defined in* (3)*. Furthermore, let $C_{\hat{\boldsymbol{\theta}}_T}$ be an estimate of the Bayes classifier using an estimate $\hat{\boldsymbol{\theta}}_T$ of $\boldsymbol{\theta}_T$. Then the risk measuring the performance of the estimation is given by*

$$Prob\big\{C_{\hat{\boldsymbol{\theta}}_T}(\boldsymbol{x}_T) \neq C_{\boldsymbol{\theta}_T}(\boldsymbol{x}_T)\big\} = \frac{1}{\pi}\arccos(\hat{\boldsymbol{\theta}}_T^T\boldsymbol{\theta}_T)$$

Proposition 1 motivates us to define the transfer distance between a source and target as follows.

Definition 1. *(Transfer distance) For a source and target with parameters $\boldsymbol{\theta}_S$ and $\boldsymbol{\theta}_T$, we define the transfer distance between them as*

$$\rho(\boldsymbol{\theta}_S, \boldsymbol{\theta}_T) := \frac{1}{\pi}\arccos(\boldsymbol{\theta}_S^T\boldsymbol{\theta}_T)$$

Remark 1. *Definition 1 is inspired from Proposition 1 which is based on the assumption of Gaussian features. However, the Transfer distance defined here works for any source and target tasks with arbitrary distributions.*

Equipped with the notion of transfer distance, we can now state the transfer learning minimax risk.

$$\mathcal{R}_T^{\Delta} := \inf_{\hat{\boldsymbol{\theta}}_T} \sup_{\rho(\boldsymbol{\theta}_S,\boldsymbol{\theta}_T)\leq\Delta} \mathbb{E}_{\substack{S_{\mathbb{P}_{\boldsymbol{\theta}_S}}\sim\mathbb{P}_{\boldsymbol{\theta}_S}^{1:n_S}\\ S_{\mathbb{Q}_{\boldsymbol{\theta}_T}}\sim\mathbb{Q}_{\boldsymbol{\theta}_T}^{1:n_T}}}\Big[\text{Prob}\big\{C_{\hat{\boldsymbol{\theta}}_T}(\boldsymbol{x}_T) \neq C_{\boldsymbol{\theta}_T}(\boldsymbol{x}_T)\big\}\Big] \tag{4}$$

Here, $S_{\mathbb{P}_{\boldsymbol{\theta}_S}}$ and $S_{\mathbb{Q}_{\boldsymbol{\theta}_T}}$ denote i.i.d. samples $\{(\boldsymbol{x}_S^{(i)}, y_S^{(i)})\}_{i=1}^{n_S}$ and $\{(\boldsymbol{x}_T^{(i)}, y_T^{(i)})\}_{i=1}^{n_T}$ generated from the source and target distributions. Moreover, the parameter $0 \leq \Delta \leq 1$ captures the class of pairs of source and target tasks within a distance over which the supremum is taken. Furthermore, we would like to highlight that $\hat{\boldsymbol{\theta}}_T$ is a function of samples $S_{\mathbb{P}_{\boldsymbol{\theta}_S}}$ and $S_{\mathbb{Q}_{\boldsymbol{\theta}_T}}$.

IV. Main Results

We now present our lower bounds for the transfer learning minimax risk (4)

Theorem 1. *Consider the transfer learning model defined in Section III-A consisting of n_S and n_T source and target training data generated i.i.d. according to a class of source/target tasks with transfer distance at most Δ per Definition 1. Furthermore, assume the dimension d obeys $d \geq 300$ and $n_T > \frac{d}{800}$. Then, the transfer learning minimax risk* (4) *obeys the following lower bounds.*

$$\mathcal{R}_T^{\Delta} \geq \begin{cases} c\frac{d}{n_T}, & \text{if } \Delta \geq B1 \\ (1-\cos^2(\Delta))\big[1 - \frac{n_T(1-\cos^2(\Delta))+\log 2}{.04d}\big], & \text{if } B2 \leq \Delta < B1 \\ c\frac{d}{n_S+n_T}, & \text{if } \Delta < B2 \end{cases} \tag{5}$$

where

$$B1 = \frac{1}{\pi}\arccos\left(\sqrt{1 - \frac{d}{200n_T}\big[\frac{1}{4} - \frac{100\log 2}{d}\big]}\right)$$

$$B2 = \frac{.04d - \log 2}{16\pi(n_S+n_T)}$$

and c is a numerical constant.

Theorem 1 consists of three regimes:

- **Large transfer distance** ($\Delta \geq B1$). In this regime, the lower bound is independent of number of source samples, n_S, which indicates that source samples are helpful until the target error for estimating the Bayes classifier reaches $c\frac{d}{n_T}$. Beyond this point, increasing n_S is no longer helpful in reducing the target error, since in this regime source is far from the target and the similarity between them is low.
- **Moderate distance** ($B2 \leq \Delta < B1$). The distance between source and target in this regime is lower than that in the previous regime. In this regime, the lower bound also does not depend on n_S which shows that even in the case that the distance is not high but it is strictly positive, i.e. $\Delta > 0$, number of source samples cannot compensate for the target samples. Because even if there are infinitely many source samples, the error does not go to zero.
- **Small distance** ($\Delta < B2$). In this regime, since the source and target are similar to each other, source samples are as useful as target samples in reducing the target error. Furthermore, in a non-transfer learning setting, the minimax risk is proportional to the dimension and reciprocal of the number of samples. Similarly in this regime the risk is proportional to the dimension and reciprocal of combination of the source and target samples

as if the source samples are as effective as target samples.

V. Experimental Results

In this section we evaluate our theoretical formulations on a subset of DomainNet data set [16]. By plotting the theoretical lower bounds as well as upper bounds obtained by weighted empirical risk minimization we investigate the sharpness of the lower bounds. Furthermore, we investigate that the semantic transfer distance defined in 1 conforms with practical settings.

Experimental setup. We use DomainNet to perform image classification task. We first pick three pairs of source and target tasks as described in Table I. Then we extract features of dimension 2048 by passing the raw images through a ResNet50 network pretrained on Imagenet.

Training. For each pair, We train linear networks separately for the source and target tasks. Using the estimated parameters appearing in Definition 1 we calculate the semantic distance for each pair as shown in Table I. For finding the corresponding upper bounds, we run weighted empirical risk minimization using the following formulation

$$\min_{\boldsymbol{\theta}} \frac{1-\lambda}{n_T}\sum_{i=1}^{n_T}\text{Cost}(C_{\boldsymbol{\theta}}(\boldsymbol{x}_T), y_T^{(i)}) + \frac{\lambda}{n_S}\sum_{i=1}^{n_S}\text{Cost}(C_{\boldsymbol{\theta}}(\boldsymbol{x}_S), y_S^{(i)}) \tag{6}$$

where $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and the cost function is logistic regression. We run each experiment for 5 times and report the average of the results.

Results. As Table I shows, the pair (Source1, Target) has the lowest transfer distance among other pairs since both the source and target share the same objects, namely Clock and Ambulance. The semantic distance of pair 2 is less than that of pair 3, because in pair 2 the source and target share at least one common object which is Clock. For plotting the theoretical lower bounds one needs to know the numerical constants appearing in Theorem 1. We will discuss how to estimate the corresponding numerical constant in the proof section of the long version [17].

Fig 1 demonstrates that pairs with small semantic distance have lower target generalization error when the number of target samples is small. Because source samples would be more useful and compensate for the target samples. Furthermore, Fig 2 shows that pairs with lower semantic distance have higher λ used in (6) which suggests the effectiveness of source samples when the distance is small.

Tasks	Transfer distance
Target: Clock vs. Ambulance (Clipart)	-
Source1: Clock vs. Ambulance (Sketch)	0.35
Source2: Clock vs. apple (Sketch)	0.41
Source3:apple vs. animal-migration (Sketch)	0.48

TABLE I: Three pairs of source and target tasks along with corresponding semantic distance

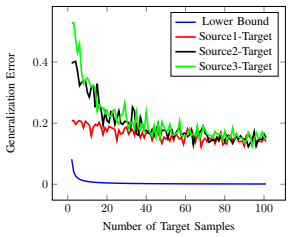

Fig. 1: Theoretical lower bound along with upper bounds for three pairs of source and target obtained by weighted empirical risk minimization.

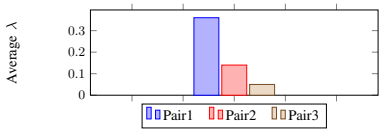

Fig. 2: Average λ used in weighted ERM (6) for three different pairs of source and target tasks shown in Table I.

VI. Proof Outline

A. Proof of Proposition 1

$$\begin{aligned}&\text{Prob}\{C_{\hat{\boldsymbol{\theta}}_T}(\boldsymbol{x}_T) \neq C_{\theta_T}(\boldsymbol{x}_T)\}\\ &=\text{Prob}\{\boldsymbol{x}_T^T\hat{\boldsymbol{\theta}}_T > 0, \boldsymbol{x}_T^T\boldsymbol{\theta}_T < 0\}\\ &\quad+\text{Prob}\{\boldsymbol{x}_T^T\hat{\boldsymbol{\theta}}_T < 0, \boldsymbol{x}_T^T\boldsymbol{\theta}_T > 0\}\end{aligned}$$

Let $w_1 = \boldsymbol{x}_T^T\hat{\boldsymbol{\theta}}_T$ and $w_2 = \boldsymbol{x}_T^T\boldsymbol{\theta}_T$. Since $\boldsymbol{x}_T \sim \mathcal{N}(0, I_d)$, we have

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \|\hat{\boldsymbol{\theta}}\|_{\ell_2}^2 & \hat{\boldsymbol{\theta}}^T\boldsymbol{\theta} \\ \hat{\boldsymbol{\theta}}^T\boldsymbol{\theta} & \|\boldsymbol{\theta}\|_{\ell_2}^2 \end{bmatrix}\right)$$

Hence,

$$\begin{aligned}&\text{Prob}\{C_{\hat{\boldsymbol{\theta}}_T}(\boldsymbol{x}_T) \neq C_{\theta_T}(\boldsymbol{x}_T)\}\\ &= \text{Prob}\{w_1 > 0, w_2 < 0\} + \text{Prob}\{w_1 < 0, w_2 > 0\}\\ &= 1 - 2\left(\frac{1}{4} + \frac{1}{2\pi}\arcsin\frac{\hat{\boldsymbol{\theta}}^T\boldsymbol{\theta}}{\|\hat{\boldsymbol{\theta}}\|_{\ell_2}\|\boldsymbol{\theta}\|_{\ell_2}}\right)\\ &= \frac{1}{\pi}\arccos\frac{\hat{\boldsymbol{\theta}}^T\boldsymbol{\theta}}{\|\hat{\boldsymbol{\theta}}\|_{\ell_2}\|\boldsymbol{\theta}\|_{\ell_2}} = \frac{1}{\pi}\arccos(\hat{\boldsymbol{\theta}}^T\boldsymbol{\theta})\end{aligned}$$

Where in the last equation we use the fact that $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ lie on the unit sphere.

B. Sketch of Proof of Theorem 1

By using Proposition 1 we can write the minimax risk as follows

$$\mathcal{R}_T^{\Delta} = \inf_{\widehat{\boldsymbol{\theta}}_T} \sup_{\rho(\theta_S,\theta_T)\le\Delta} \mathbb{E}_{\substack{S_{\mathbb{P}_{\boldsymbol{\theta}_S}}\sim\mathbb{P}_{\boldsymbol{\theta}_S}^{1:n_S} \\ S_{\mathbb{Q}_{\boldsymbol{\theta}_T}}\sim\mathbb{Q}_{\boldsymbol{\theta}_T}^{1:n_T}}} \Big[\text{Prob}\{C_{\widehat{\boldsymbol{\theta}}_T}(\boldsymbol{x}_T)\neq C_{\theta_T}(\boldsymbol{x}_T)\}\Big]$$

$$= \inf_{\widehat{\boldsymbol{\theta}}_T} \sup_{\rho(\theta_S,\theta_T)\le\Delta} \mathbb{E}_{\substack{S_{\mathbb{P}_{\boldsymbol{\theta}_S}}\sim\mathbb{P}_{\boldsymbol{\theta}_S}^{1:n_S} \\ S_{\mathbb{Q}_{\boldsymbol{\theta}_T}}\sim\mathbb{Q}_{\boldsymbol{\theta}_T}^{1:n_T}}} \Big[\rho\Big(\widehat{\boldsymbol{\theta}}_T(S_{\mathbb{P}_{\boldsymbol{\theta}_S}}, S_{\mathbb{P}_{\boldsymbol{\theta}_T}}), \boldsymbol{\theta}_T\Big)\Big]$$

where $\widehat{\boldsymbol{\theta}}_T = \widehat{\boldsymbol{\theta}}_T(S_{\mathbb{P}_{\boldsymbol{\theta}_S}}, S_{\mathbb{P}_{\boldsymbol{\theta}_T}})$ is a function of samples $(S_{\mathbb{P}_{\boldsymbol{\theta}_S}}, S_{\mathbb{P}_{\boldsymbol{\theta}_T}})$. Note that the distance ρ, that is the Geodesic distance on the unit sphere, is a metric, which would be necessary in the sequel. Then we follow the usual technique of reducing the minimax risk to hypothesis testing inspired by the proof [15] (See also [18, Chapter 15] for non-transfer learning minimax risk). [15] provides lower bounds for minimax risk in regression problems. In this paper we use some ideas of [15] to find minimax lower bounds for classification.

Since our goal is to estimate the the target parameter using the source and target data, we need to pick N pairs of distributions $(\mathbb{P}_{\boldsymbol{\theta}_S}^{(1)}, \mathbb{Q}_{\boldsymbol{\theta}_T}^{(1)}), ..., (\mathbb{P}_{\boldsymbol{\theta}_S}^{(N)}, \mathbb{Q}_{\boldsymbol{\theta}_T}^{(N)})$ such that

$$\rho(\boldsymbol{\theta}_T^{(i)}, \boldsymbol{\theta}_T^{(j)}) \ge 2\delta \quad \text{for each } i \neq j \in [N]\times[N], \tag{7}$$

$$\rho(\boldsymbol{\theta}_S^{(i)}, \boldsymbol{\theta}_T^{(i)}) \le \Delta \quad \text{for each } i \in [N] \tag{8}$$

(7) assures that the target parameters are 2δ-separated uing the ρ distance defined in 1, and (8) assures that the source and target distributions belong to the class of transfer learning problem over which the supremum of the minimax risk is taken.

Now by considering the following hypothesis testing:

- Let J be a random sample from the uniform distribution over $[N] := \{1, 2, ..., N\}$.
- Given $J = j$, sample $S_{\mathbb{P}_{\boldsymbol{\theta}_S^{(j)}}} \sim \mathbb{P}_{\boldsymbol{\theta}_S^{(j)}}^{1:n_S}$ and $S_{\mathbb{Q}_{\boldsymbol{\theta}_T^{(j)}}} \sim \mathbb{P}_{\boldsymbol{\theta}_T^{(j)}}^{1:n_T}$.

We aim at finding the true index using the $n_S + n_T$ samples using a testing function.

Using the proof [15], one can find the following lower bound for the minimax risk:

$$\mathcal{R}_T^{\Delta} \ge \delta\left(1 - \frac{n_S I(J;E) + n_T I(J;F) + \log 2}{\log N}\right) \tag{9}$$

where E and F are random variables such that $E|\{J = j\} \sim \mathbb{P}_{\boldsymbol{\theta}_S^{(j)}}$ and $F|\{J = j\} \sim \mathbb{Q}_{\boldsymbol{\theta}_T^{(j)}}$, and I denotes the mutual information.

Then by using the convexity of KL-divergence and mixture representation $I(J;E) = \frac{1}{N}\sum_{j=1}^N D_{KL}(\mathbb{P}_{\boldsymbol{\theta}_S^{(j)}} \| \frac{1}{N}\sum_{j=1}^N \mathbb{P}_{\boldsymbol{\theta}_S^{(j)}})$ and $I(J;F) = \frac{1}{N}\sum_{j=1}^N D_{KL}(\mathbb{Q}_{\boldsymbol{\theta}_T^{(j)}} \| \frac{1}{N}\sum_{j=1}^N \mathbb{Q}_{\boldsymbol{\theta}_T^{(j)}})$, we can bound the mutual information appearing in (9) as follows

$$I(J;E) \le \frac{1}{N^2}\sum_{i,j} D_{KL}(\mathbb{P}_{\boldsymbol{\theta}_S^{(i)}} \| \mathbb{P}_{\boldsymbol{\theta}_S^{(j)}})$$

$$I(J;F) \le \frac{1}{N^2}\sum_{i,t} D_{KL}(\mathbb{Q}_{\boldsymbol{\theta}_T^{(i)}} \| \mathbb{Q}_{\boldsymbol{\theta}_T^{(j)}}). \tag{10}$$

Following lemma bounds the KL-divergence in (10).

Lemma 1. *Let $\mathbb{P}_{\boldsymbol{\theta}_S}$, $\mathbb{P}_{\boldsymbol{\theta}'_S}$ be two joint distributions of the features and labels in the source task and $\mathbb{Q}_{\boldsymbol{\theta}_T}$, $\mathbb{Q}_{\boldsymbol{\theta}'_T}$ be those in the target task according to the model defined in Section III-A. Then*

$$D_{KL}(\mathbb{P}_{\boldsymbol{\theta}_S} \| \mathbb{P}_{\boldsymbol{\theta}'_S}) + D_{KL}(\mathbb{P}_{\boldsymbol{\theta}'_S} \| \mathbb{P}_{\boldsymbol{\theta}_S}) \le \|\boldsymbol{\theta}_S - \boldsymbol{\theta}'_S\|_{\ell_2}^2$$

$$D_{KL}(\mathbb{Q}_{\boldsymbol{\theta}_T} \| \mathbb{Q}_{\boldsymbol{\theta}'_T}) + D_{KL}(\mathbb{Q}_{\boldsymbol{\theta}'_T} \| \mathbb{Q}_{\boldsymbol{\theta}_T}) \le \|\boldsymbol{\theta}_T - \boldsymbol{\theta}'_T\|_{\ell_2}^2$$

See the long version [17] for the proof. Based on the distance of source and target, Δ, we divide the proof of Theorem 1 into two parts and one can conclude the proof using the following two lemmas.

Lemma 2. *Assume that*

$$\Delta \ge \frac{1}{\pi}\arccos\left(\sqrt{1 - \frac{d}{200 n_T}\big[\frac{1}{4} - \frac{100\log 2}{d}\big]}\right)$$

where n_T and d are the number of target samples and the dimension. Then we would have $\mathcal{R}_T^{\Delta} \ge c\frac{d}{n_T}$. Furthermore, if $\Delta < \frac{1}{\pi}\arccos\left(\sqrt{1 - \frac{d}{200 n_T}\big[\frac{1}{4} - \frac{100\log 2}{d}\big]}\right)$ then $\mathcal{R}_T^{\Delta} \ge (1 - \cos^2(\Delta))\big[1 - \frac{n_T(1-\cos^2(\Delta))+\log 2}{.04d}\big]$

Lemma 3. *Suppose that there are n_S and n_T number of source and target samples and $\Delta < \frac{.04d - \log 2}{16\pi(n_S+n_T)}$. Then*

$$\mathcal{R}_T^{\Delta} \ge c\frac{d}{n_S + n_T}$$

See [17] for the proof of Lemma 2 and 3.

VII. Acknowledgments

This material is based upon work supported by Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-19-2-1005. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. M. Soltanolkotabi is also supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship in Mathematics, an NSF-CAREER under award #1846369, DARPA FastNICS program, and NSF-CIF awards #1813877 and #2008443.

REFERENCES

[1] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[2] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning bounds for domain adaptation," in *Advances in neural information processing systems*, 2008, pp. 129–136.

[3] K. You, X. Wang, M. Long, and M. Jordan, "Towards accurate model selection in deep unsupervised domain adaptation," in *International Conference on Machine Learning*, 2019, pp. 7124–7133.

[4] X. Chen, S. Wang, M. Long, and J. Wang, "Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation," in *International Conference on Machine Learning*, 2019, pp. 1081–1090.

[5] Y. Wu, E. Winston, D. Kaushik, and Z. Lipton, "Domain adaptation with asymmetrically-relaxed distribution alignment," *arXiv preprint arXiv:1903.01689*, 2019.

[6] K. Azizzadenesheli, A. Liu, F. Yang, and A. Anandkumar, "Regularized learning for domain adaptation under label shifts," *arXiv preprint arXiv:1903.09734*, 2019.

[7] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Advances in neural information processing systems*, 2016, pp. 136–144.

[9] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[10] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Advances in neural information processing systems*, 2007, pp. 137–144.

[11] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," *arXiv preprint arXiv:0902.3430*, 2009.

[12] Q. Lei, W. Hu, and J. Lee, "Near-optimal linear regression under distribution shift," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6164–6174.

[13] S. Ben-David, T. Lu, T. Luu, and D. Pál, "Impossibility theorems for domain adaptation," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 129–136.

[14] S. Hanneke and S. Kpotufe, "On the value of target data in transfer learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 9867–9877.

[15] M. Kalan and Z. Fabian, "Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks," *Neural Information Processing Systems (NeuRIPS 2020)*, 2020.

[16] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1406–1415.

[17] [Online]. Available: https://drive.google.com/file/d/1l_x9Q2RRXgZwzJTBYBVEEGoF1X7SyKzw/view?usp=sharing

[18] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.