Near-Optimal Linear Regression under Distribution Shift

Qi Lei [1] **Wei Hu** [1] **Jason D. Lee** [1]

Abstract

Transfer learning is essential when sufficient data comes from the source domain, with scarce labeled data from the target domain. We develop estimators that achieve minimax linear risk for linear regression problems under distribution shift. Our algorithms cover different transfer learning settings including covariate shift and model shift. We also consider when data are generated from either linear or general nonlinear models. We show that linear minimax estimators are within an absolute constant of the minimax risk even among nonlinear estimators for various source/target distributions.

1. Introduction

The success of machine learning crucially relies on the availability of labeled data. The data labeling process usually requires extensive human labor and can be very expensive and time-consuming, especially for large datasets like ImageNet (Deng et al., 2009). On the other hand, models trained on one dataset, despite performing well on test data from the same distribution they are trained on, are often sensitive to *distribution shifts*, i.e., they do not adapt well to related but different distributions. Even small distributional shift can result in substantial performance degradation (Recht et al., 2018; Lu et al., 2020).

Transfer learning has been an essential paradigm to tackle the challenges associated with insufficient labeled data (Pan & Yang, 2009; Weiss et al., 2016; Long et al., 2017). The main idea is to make use of a *source domain* with plentiful labeled data (e.g., ImageNet) and to learn a model that performs well on the *target domain* (e.g., medical images) where few or no labels are available. Despite the lack of labeled data, we may still use unlabeled data from the target domain, which are usually much easier to obtain and can provide helpful marginal distribution information about the target domain. Although this approach is integral to many applications, many fundamental questions are left open even in very basic settings.

In this work, we focus on the setting of *linear regression under distribution shift* and ask the fundamental question of how to optimally learn a linear model for the target domain, using labeled data from a source domain and unlabeled data (and possibly limited unlabeled data) from the target domain. We design a two-stage meta-algorithm that addresses this problem in various settings, including covariate shift (i.e., when $p(\boldsymbol{x})$ changes) and model shift (i.e., when $p(y|\boldsymbol{x})$ changes). Following the meta-algorithm, we develop estimators that achieve *near minimax risk* (up to universal constant factors) among all linear estimation rules under some standard data concentration properties. Here linear estimators refer to all estimators that depend linearly on the label vector; these include almost all popular estimators known in linear regression, such as ridge regression and its variants. When the second moment matrix of input variables in source and target domains commute, we prove that our estimators achieve near minimax risk among all possible estimators. We also provide a separation result demonstrating our algorithm can be better than ridge regression by a multiplicative factor of $\tilde{O}(d^{-1/4})$.

A crucial insight from our results is that when covariate shift is present, we need to apply data-dependent regularization that adapts to changes in the input distribution. For linear regression, this is characterized by the input covariances of source and target tasks, estimated using unlabeled data. Our experiments verify that our estimator has significant improvement over ridge regression and similar heuristics.

1.1. Related work

Different types of distribution shift are introduced in Storkey (2009); Quionero-Candela et al. (2009). Specifically, covariate shift occurs when the marginal distribution on $P(\boldsymbol{x})$ changes from source to target domain (Heckman, 1979; Shimodaira, 2000; Huang et al., 2007). Wang et al. (2014); Wang & Schneider (2015) tackle model shift ($P(y|\boldsymbol{x})$) provided the change is smooth as a function of $\boldsymbol{x}$. Sun et al. (2011) design a two-stage reweighting method based on both covariate shift and model shift. Other meth-

[1]Princeton University. Correspondence to: Qi Lei <qilei@princeton.edu>, Wei Hu <huwei@cs.princeton.edu>, Jason D. Lee <jasonlee@princeton.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

ods like the change of representation, adaptation through instance pruning are proposed in Jiang & Zhai (2007). In this work, we focus on the above two kinds of distribution shifts. Other distribution shift settings involving label/target shift ($P(y)$) and conditional shift ($P(\boldsymbol{x}|y)$) are beyond the scope of this paper. Some prior work also focuses on these settings (See reference therein (Saerens et al., 2002; Zhang et al., 2013; Lipton et al., 2018)). For instance, Zhang et al. (2013) exploits the benefit of multi-layer adaptation by a location-scale transformation on $\boldsymbol{x}$.

Transfer learning/domain adaptation are sub-fields within machine learning to cope with distribution shift. A variety of prior work falls into the following categories. 1) Importance-reweighting is mostly used in the covariate shift (Shimodaira, 2000; Huang et al., 2007; Cortes et al., 2010); 2) One fruitful line of work focuses on exploring robust/causal features or domain-invariant representations (Wu et al., 2019) through invariant risk minimization (Arjovsky et al., 2019), distributional robust minimization (Sagawa et al., 2019), human annotation (Srivastava et al., 2020), adversarial training (Long et al., 2017; Ganin et al., 2016), or by minimizing domain discrepancy measured by some distance metric (Pan et al., 2010; Long et al., 2013; Baktashmotlagh et al., 2013; Gong et al., 2013; Zhang et al., 2013; Wang & Schneider, 2014) ; 3) Several approaches seek gradual domain adaptation (Gopalan et al., 2011; Gong et al., 2012; Glorot et al., 2011; Kumar et al., 2020) through self-training or a gradual change in the training distribution.

Near minimax estimations are introduced in Donoho (1994) for linear regression problems with Gaussian noise. For a more general setting, Juditsky et al. (2009) estimate the linear functional using convex programming. Blaker (2000) compares ridge regression with a minimax linear estimator using weighted squared error. Kalan et al. (2020) considers a setting similar to this work of minimax estimator under distribution shift but focuses on computing the lower bound for the linear and one-hidden-layer neural network under distribution shift. A few more interesting results are derived by minimizing the generalization error bounds for distribution shift under various settings (David et al., 2010; Hanneke & Kpotufe, 2019; Ben-David et al., 2010; Zhao et al., 2019).

2. Preliminary

We formalize the setting considered in this paper for transfer learning under the distribution shift.

Notation and setup. Let $p_S(\boldsymbol{x})$ and $p_T(\boldsymbol{x})$ be the marginal distribution for $\boldsymbol{x}$ in source and target domain. The associated second-moment matrices are $\Sigma_S := \mathbb{E}_{p_S}[\boldsymbol{x}\boldsymbol{x}^\top]$, and $\Sigma_T := \mathbb{E}_{p_S}[\boldsymbol{x}\boldsymbol{x}^\top]$. Labeled data $(\boldsymbol{x}, y)$ satisfies $\mathbb{E}_{p_S}[y|\boldsymbol{x}] = \mathbb{E}_{p_T}[y|\boldsymbol{x}] = f^*(\boldsymbol{x})$ and $y = f^*(\boldsymbol{x}) + z$ with Gaussian noise $z \sim \mathcal{N}(0, \sigma^2)$. We consider both linear ($f^*(\boldsymbol{x}) := \mathbb{E}[y|\boldsymbol{x}] = \boldsymbol{x}^\top\boldsymbol{\beta}^*$) and general nonlinear data generation model. When the optimal linear model changes from source to domain we add a subscript for distinction, i.e., $\boldsymbol{\beta}_S^*$ and $\boldsymbol{\beta}_T^*$. We use bold ($\boldsymbol{x}$) symbols for vectors, lower case letter (x) for scalars and capital letter (A) for matrices.

We observe n_S, n_T labeled samples from source and target domain, and n_U unlabeled target samples. Labeled data is scarce in target domain: $n_S \gg n_T$ and n_T can be 0. Specifically, data is collected as $X_S = [\boldsymbol{x}_1^\top | \boldsymbol{x}_2^\top | \cdots | \boldsymbol{x}_{n_S}^\top]^\top \in \mathbb{R}^{n_S \times d}$, with $\boldsymbol{x}_i, i \in [n_S]$ drawn from p_S, noise $\boldsymbol{z} = [z_1, z_2, \cdots z_{n_S}]^\top, z_i \sim \mathcal{N}(0, \sigma^2)$. $\boldsymbol{y}_S = [y_1, y_2, \cdots, y_{n_S}]^\top \in \mathbb{R}^{n_S}$, $\boldsymbol{y}_T \in \mathbb{R}^{n_T}$ and $X_U \in \mathbb{R}^{n_U \times d}$ are similarly defined). Denote by $\hat{\Sigma}_S = X_S^\top X_S / n_S$ the empirical second-moment matrix. The positive part of a number is denoted by $(x)_+$.

Minimax (linear) risk. In this work, we focus on designing linear estimators $\hat{\boldsymbol{\beta}} : \mathbb{R}^n \to \mathbb{R}^d, \boldsymbol{y}_S \to A\boldsymbol{y}_S$[1] for $\boldsymbol{\beta}_T^* \in \mathcal{B}$. Here $\boldsymbol{\beta}_T^*$ is the optimal linear model in target domain ($:= \arg\min_{\boldsymbol{\beta}} \mathbb{E}_{\boldsymbol{x}\sim p_T, z\sim\mathcal{N}(0,\sigma^2)}[(f^*(\boldsymbol{x}) + z - \boldsymbol{x}^\top\boldsymbol{\beta})^2]$). [2]

Our estimator is evaluated by the excess risk on target domain, with the worst case $\boldsymbol{\beta}_T^*$ in some set $\mathcal{B}$: $L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta}^* \in \mathcal{B}} \mathbb{E}_{\boldsymbol{y}_S} \mathbb{E}_{\boldsymbol{x}\sim p_T} \left(\boldsymbol{x}^\top(\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S) - \boldsymbol{\beta}_T^*)\right)^2$. Minimax linear risk and minimax risk among all estimators are respectively defined as:

$$R_L(\mathcal{B}) \equiv \min_{\hat{\boldsymbol{\beta}} \text{ linear in } \boldsymbol{y}_S} L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}); \quad R_N(\mathcal{B}) \equiv \min_{\hat{\boldsymbol{\beta}}} L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}).$$

The subscript "N" or "L" is a mnemonic for "non-linear" or "linear" estimators. R_N is the optimal risk with no restriction placed on the class of estimators. R_L only considers the linear function class for $\hat{\boldsymbol{\beta}}$. Minimax linear estimator and minimax estimator are the estimators that respectively attain R_L and R_N within universal multiplicative constants. Normally we only consider $\mathcal{B} = \{\boldsymbol{\beta} | \|\boldsymbol{\beta}\|_2 \le r\}$. When there is no ambiguity, we simplify $\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S)$ by $\hat{\boldsymbol{\beta}}$.

Our meta-algorithm. Our paper considers different settings with distribution shift. Our methods are unified under the following meta-algorithm:

Step 1: Construct an unbiased sufficient statistic $\hat{\boldsymbol{\beta}}_{\text{SS}}$[3] for the unknown parameter.
Step 2: Construct $\hat{\boldsymbol{\beta}}_{\text{MM}}$, a linear function of the sufficient

[1] $A \in \mathbb{R}^{d\times n}$ may depend in an arbitrary way on X_S, n_S, or Σ_T. The estimator is linear in the observation $\boldsymbol{y}_S$.

[2] We do not distinguish linear and affine regression since one could simply add another constant coordinate to $\boldsymbol{x}$ to take into consideration the intercept part.

[3] With samples $\boldsymbol{y}_S$, a statistic $t = T(\boldsymbol{y}_S)$ is sufficient for the underlying parameter $\boldsymbol{\beta}^*$ if the conditional probability distribution of the data $\boldsymbol{y}_S$, given the statistic $t = T(\boldsymbol{y}_S)$, does not depend on the parameter $\boldsymbol{\beta}^*$.

statistic $\hat{\boldsymbol{\beta}}_{\text{SS}}$ that minimizes $L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}_{\text{MM}})$.

For each setting, we will show that $\hat{\boldsymbol{\beta}}_{\text{MM}}$ achieves linear minimax risk R_L. Furthermore, under some conditions, the minimax risk R_N is uniformly lower bounded by a universal constant times $L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}_{\text{MM}})$.

Outline. In the sections below, we tackle the problem in several different settings. In Section 3, we design algorithms with only covariate shift and linear data-generation models (f^* is linear) for unsupervised domain adaptation ($n_T = 0$) in Section 3.1, and supervised domain adaptation ($n_T > 0$) in Section 3.4. Section 4 is about linear regression with approximation error ($n_T = 0$ and $f^*(\boldsymbol{x})$ is a general nonlinear function). Finally we consider model shift for linear models ($\boldsymbol{\beta}_S^* \neq \boldsymbol{\beta}_T^*$) in Section 5.

3. Covariate shift with linear models

In this section, we consider the setting with only covariate shift and f^* is linear. That is, only Σ_S (marginal distribution $p_S(\boldsymbol{x})$) changes to Σ_T (marginal distribution $p_T(\boldsymbol{x})$), but $f^* = \mathbb{E}[y|\boldsymbol{x}] = \boldsymbol{x}^\top \boldsymbol{\beta}^*$ (conditional distribution $p(y|\boldsymbol{x})$) is shared.

3.1. Unsupervised domain adaptation with linear models

We observe n_S samples from source domain: $\boldsymbol{y}_S = X_S\boldsymbol{\beta}^* + \boldsymbol{z}, \boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 I)$ and only some unlabeled samples X_U from the target domain. Our goal is to find the minimax linear estimator $\hat{\boldsymbol{\beta}}_{\text{MM}}(\boldsymbol{y}_S) = A\boldsymbol{y}_S$ with some linear mapping A that attains $R_L(\mathcal{B})$ [4].

Following our meta-algorithm, let $\hat{\boldsymbol{\beta}}_{\text{SS}} = \frac{1}{n_S}\hat{\Sigma}_S^{-1}X_S^\top \boldsymbol{y}_S$[5] be an unbiased sufficient statistic for $\boldsymbol{\beta}^*$:

$$\hat{\boldsymbol{\beta}}_{\text{SS}} = \frac{1}{n_S}\hat{\Sigma}_S^{-1}X_S^\top \boldsymbol{y}_S = \frac{1}{n_S}\hat{\Sigma}_S^{-1}X_S^\top X_S\boldsymbol{\beta}^* + \frac{1}{n_S}\hat{\Sigma}_S^{-1}X_S^\top \boldsymbol{z}.$$
$$= \boldsymbol{\beta}^* + \frac{1}{n_S}\hat{\Sigma}_S^{-1}X_S^\top \boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{\beta}^*, \frac{\sigma^2}{n_S}\hat{\Sigma}_S^{-1}\right). \quad (1)$$

The fact that $\hat{\boldsymbol{\beta}}_{\text{SS}}(\boldsymbol{y}_S)$ is a sufficient statistic for $\boldsymbol{\beta}^*$ is proven in Claim 3.9 for a more general case, using the Fisher-Neyman factorization theorem. We prove that the minimax linear estimator is of the form $\hat{\boldsymbol{\beta}}_{\text{MM}} = C\hat{\boldsymbol{\beta}}_{\text{SS}}$ and then design algorithms that calculate the optimal C.

Claim 3.1. *The minimax linear estimator is of the form $\hat{\boldsymbol{\beta}}_{MM} = C\hat{\boldsymbol{\beta}}_{SS}$ for some $C \in \mathbb{R}^{d\times d}$.*

Warm-up: commutative second-moment matrices. In order to derive the minimax linear estimator, we first consider the simple case when Σ_T and $\hat{\Sigma}_S$ are simultaneously diagonalizable. We note that under this setting, minimax estimation under covariate shift reduces to the well-studied problem of finding a minimax linear estimator under weighted square loss (see e.g., (Blaker, 2000)). One could apply Pinsker's Theorem (Johnstone, 2011) and get an estimator function and the minimax risk with a closed form:

Theorem 3.2 (Linear Minimax Risk with Covariate Shift)**.** *Suppose the observations follow sequence model $\boldsymbol{y}_S = X_S\boldsymbol{\beta}^* + \boldsymbol{z}, \boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 I_n)$. If $\Sigma_T = U\text{diag}(\boldsymbol{t})U^\top$ and $\hat{\Sigma}_S \equiv X_S^\top X_S/n_S = U\text{diag}(\boldsymbol{s})U^\top$, then the minimax linear risk*

$$R_L(\mathcal{B}) \equiv \min_{\hat{\boldsymbol{\beta}}=A\boldsymbol{y}_S}\max_{\boldsymbol{\beta}^*\in\mathcal{B}} \mathbb{E}\,\|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2$$
$$= \sum_i \frac{\sigma^2}{n_S}\frac{t_i}{s_i}\left(1 - \frac{\lambda}{\sqrt{t_i}}\right)_+,$$

where $\mathcal{B} = \{\boldsymbol{\beta} | \|\boldsymbol{\beta}\| \le r\}$, and $\lambda = \lambda(r)$ is determined by $\frac{\sigma^2}{n_S}\sum_{i=1}^d \frac{1}{s_i}(\sqrt{t_i}/\lambda - 1)_+ = r^2$. The linear minimax estimator is given by:

$$\hat{\boldsymbol{\beta}}_{MM} = \Sigma_T^{-1/2}U(I - \text{diag}(\lambda/\sqrt{\boldsymbol{t}}))_+U^\top\Sigma_T^{1/2}\hat{\boldsymbol{\beta}}_{SS}, \quad (2)$$
$$\text{where } \hat{\boldsymbol{\beta}}_{SS} = \frac{1}{n_S}\hat{\Sigma}_S^{-1}X_S^\top \boldsymbol{y}_S.$$

Since r is unknown in practice, we could simply view either r or directly λ as the tuning parameter. We compare the functionality of λ with that of ridge regression: $\hat{\boldsymbol{\beta}}_{\text{RR}}^\lambda = \arg\min_{\hat{\boldsymbol{\beta}}} \mathbb{E}\,\frac{1}{2n}\|X_S\hat{\boldsymbol{\beta}} - \boldsymbol{y}_S\|^2 + \frac{\lambda}{2}\|\hat{\boldsymbol{\beta}}\|^2 = (\hat{\Sigma}_S + \lambda I)^{-1}X_S^\top \boldsymbol{y}_S/n_S$. For both algorithms, λ balances the bias and variance: $\lambda = 0$ gives an unbiased estimator, and a big λ gives a (near) zero estimator with no variance. The difference is, the minimax linear estimator shrinks some signal directions based on the value of t_i, since the risk in those directions is downweighted in the target loss. The estimator tends to sacrifice the directions of signal where t_i is smaller. Ridge regression, however, respects the value of s_i. A natural counterpart is for ridge to also regularize based on $\boldsymbol{t}$: let $\hat{\boldsymbol{\beta}}_{\text{RR},T}^\lambda = \arg\min \frac{1}{n}\|\Sigma_T^{1/2}(\boldsymbol{\beta} - \hat{\Sigma}_S^{-1}X_S^\top \boldsymbol{y}_S)\|^2 + \lambda\|\boldsymbol{\beta}\|^2 = (\Sigma_T + \lambda I)^{-1}\Sigma_T\hat{\boldsymbol{\beta}}_{\text{SS}}$. We will compare their performances in the experimental section.

Non-commutative second-moment matrices. For non-commutative second-moment shift, we follow the same procedure. Our estimator is achieved by optimizing over C: $\hat{\boldsymbol{\beta}}_{\text{MM}} = C\hat{\boldsymbol{\beta}}_{\text{SS}}$:

$$R_L(\mathcal{B}) \equiv \min_{\hat{\boldsymbol{\beta}}=A\boldsymbol{y}_S}\max_{\boldsymbol{\beta}^*\in\mathcal{B}} \mathbb{E}\,\|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2$$
$$= \min_{\hat{\boldsymbol{\beta}}=C\hat{\boldsymbol{\beta}}_{\text{SS}}}\max_{\|\boldsymbol{\beta}^*\|\le r}\left\{\|\Sigma_T^{1/2}(C - I)\boldsymbol{\beta}^*\|_2^2\right.$$
$$\left. + \frac{\sigma^2}{n_S}\text{Tr}(\Sigma_T^{1/2}C\hat{\Sigma}_S^{-1}C^\top\Sigma_T^{1/2})\right\} \quad \text{(Claim 3.1)}$$

[4] For linear estimator $\hat{\boldsymbol{\beta}} = A\boldsymbol{y}_S$, $\boldsymbol{y}_S$ is the only source of randomness and A depends on X_S, n_S, which are considered fixed.

[5] Throughout the paper $\hat{\Sigma}_S^{-1}$ could be replaced by pseudo-inverse and our algorithm also applies when $n < d$.

$$= \min_{\tau, C} \left\{ r^2\tau + \frac{\sigma^2}{n_S} \mathrm{Tr}(\Sigma_T^{1/2} C \hat{\Sigma}_S^{-1} C^\top \Sigma_T^{1/2}) \right\}, \quad (3)$$
$$\text{s.t. } (C-I)^\top \Sigma_T (C-I) \preceq \tau I.$$

Unlike the commutative case, this problem does not have a closed form solution, but is still computable:

Proposition 3.3. *Problem (3) is a convex program and computable in polynomial-time.*

We achieve near-optimal minimax risk among all estimators under some conditions:

Theorem 3.4 (Near minimaxity of linear estimators)**.** *When Σ_S, Σ_T commute, or Σ_T is rank 1, the best linear estimator from (2) or (3) achieves near-optimal minimax risk: $L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}_{MM}) = R_L(\mathcal{B}) \le 1.25 R_N(\mathcal{B})$.*

Note that $R_N \le R_L$ by definition. Therefore 1) our estimator $\hat{\boldsymbol{\beta}}_{\text{MM}}$ is near-optimal, and 2) our lower bound for R_N is tight. Lower bounds (without matching upper bounds) for general non-commutative problem is presented in Kalan et al. (2020) and we improve their result for the commutative case and provide a matching algorithm. Their lower bound scales with $\frac{d}{n_S} \min_i \frac{t_i}{s_i}$ for large r, while ours becomes $\frac{1}{n_S} \sum_i \frac{t_i}{s_i}$. Our lower bound is always larger and thus tighter, and potentially arbitrarily larger when $\max_i \frac{t_i}{s_i}$ and $\min_i \frac{t_i}{s_i}$ are very different. We defer our proof to the appendix.

3.2. Connection to ridge regression

From a probabilistic perspective, ridge regression is equivalent to maximum a posteriori (MAP) inference with a Gaussian prior: $\boldsymbol{\beta}^* \sim \mathcal{N}(0, r^2 I)$ (see e.g. Murphy (2012)). Similarly, instead of considering a worst-case risk that minimizes $L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) := \max_{\boldsymbol{\beta}^* \in \mathcal{B}} \mathbb{E}_{\boldsymbol{y}_S} \|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S) - \boldsymbol{\beta}^*)\|^2$, one could also study the average setting that minimizes $\bar{L}_{\mathcal{B}} := \mathbb{E}_{\boldsymbol{\beta}^* \sim \mathcal{N}(0, r^2 I)} \mathbb{E}_{\boldsymbol{y}_S} \|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S) - \boldsymbol{\beta}^*)\|^2$ instead. With distribution shift, the performance is evaluated on Σ_T instead of Σ_S. Interestingly with Gaussian prior, this does not give us a different algorithm other than the original ridge regression.

Proposition 3.5. *The optimal estimator under Gaussian prior $\boldsymbol{\beta}^* \sim \mathcal{N}(0, r^2 I)$ evaluated on p_T is:*

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &\leftarrow \arg\min_{\boldsymbol{\beta} = A\boldsymbol{y}_S} \mathbb{E}_{\boldsymbol{\beta}^* \sim \mathcal{N}(0, r^2 I)} \mathbb{E}_{\boldsymbol{y}_S} \mathbb{E}_{\boldsymbol{x} \sim p_T} \left(\boldsymbol{x}^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right)^2 \\
&= \frac{1}{n_S} \left(\frac{\sigma^2}{r^2 n_S} I + \hat{\Sigma}_S \right)^{-1} X_S^\top \boldsymbol{y}_S \\
&\equiv \arg\min_{\hat{\boldsymbol{\beta}}} \mathbb{E}\, \frac{1}{2n} \|X_S \hat{\boldsymbol{\beta}} - \boldsymbol{y}_S\|^2 + \frac{\lambda}{2} \|\hat{\boldsymbol{\beta}}\|^2 \\
&= (\hat{\Sigma}_S + \lambda I)^{-1} \hat{\Sigma}_S \hat{\boldsymbol{\beta}}_{SS} =: \hat{\boldsymbol{\beta}}_{RR}^{\lambda},
\end{aligned}$$

when $\lambda = \sigma^2/(n_S r^2)$. Namely, the average-case best linear estimator with Gaussian prior is equivalent to ridge regression with regularization strength $\lambda = \frac{\sigma^2/n_S}{r^2}$: the variance ratio between the noise distribution and prior distribution.

Even though ridge regression achieves the optimal risk in the average sense, it could be much worse than the minimax linear estimator in the worst case. We prove a separation result on a specific example (that is deferred to the appendix).

Remark 3.1 (Benefit of minimax linear estimator)**.** *There is an example that $R_L(\mathcal{B}) \le \mathcal{O}(d^{-1/4} L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}_{RR}^{\lambda}))$ even with the optimal hyperparameter λ.* [6]

Adaptation on the prior distribution. With specific problems, one should adjust the prior distribution instead of simply assume $\boldsymbol{\beta}^* \sim \mathcal{N}(0, r^2)$. If one replaces the prior by $\boldsymbol{\beta}^* \sim \mathcal{N}(\hat{\boldsymbol{\beta}}_{\text{SS}}, r^2)$, one could get another heuristic method:

Proposition 3.6. *Let $\hat{\boldsymbol{\beta}}_{SS}$ be the estimator from ordinary least square: $\hat{\boldsymbol{\beta}}_{SS} = \hat{\Sigma}_S^{-1} X_S^\top \boldsymbol{y}_S / n_S$. The optimal estimator under Gaussian prior $\boldsymbol{\beta}^* \sim \mathcal{N}(\hat{\boldsymbol{\beta}}_{SS}, r^2 I)$ evaluated on p_T is:*

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &\leftarrow \arg\min_{\boldsymbol{\beta} = A\boldsymbol{y}_S} \mathbb{E}_{\boldsymbol{\beta}^* \sim \mathcal{N}(\hat{\boldsymbol{\beta}}_{SS}, r^2 I)} \mathbb{E}_{\boldsymbol{y}_S} \mathbb{E}_{\boldsymbol{x} \sim p_T} \left(\boldsymbol{x}^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right)^2 \\
&= \frac{1}{n_S} \left(\frac{\sigma^2}{r^2 n_S} I + \Sigma_T \right)^{-1} \Sigma_T \hat{\Sigma}_S^{-1} X_S^\top \boldsymbol{y}_S \\
&\equiv \arg\min_{\boldsymbol{\beta}} \|\Sigma_T^{1/2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{SS})\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \\
&= (\Sigma_T + \lambda I)^{-1} \Sigma_T \hat{\boldsymbol{\beta}}_{SS} =: \hat{\boldsymbol{\beta}}_{RR,T}^{\lambda},
\end{aligned}$$

when $\lambda = \sigma^2/(n_S r^2)$.

Comparing the closed-form estimator $\hat{\boldsymbol{\beta}}_{\text{RR},T}^{\lambda} := (\Sigma_T + \lambda I)^{-1} \Sigma_T \hat{\boldsymbol{\beta}}_{\text{SS}}$ to the original ridge regression $\hat{\boldsymbol{\beta}}_{\text{RR}}^{\lambda} := (\hat{\Sigma}_S + \lambda I)^{-1} \hat{\Sigma}_S \hat{\boldsymbol{\beta}}_{\text{SS}}$, we could see that this algorithm regularizes $\hat{\boldsymbol{\beta}}$ based on the signal strength from the target distribution, and it is equivalent to ridge regression by adjusting the prior distribution to center at $\hat{\boldsymbol{\beta}}_{\text{SS}}$, the unbiased estimator for the ground truth $\boldsymbol{\beta}^*$. We will compare both methods with our minimax estimator in the experimental section.

3.3. Minimax linear estimator with finite unlabeled samples from target domain

In practice, we have finite unlabeled samples $X_U \in \mathbb{R}^{n_U \times d}$, where we denote the empirical second-moment matrix as $\hat{\Sigma}_U = X_U^\top X_U / n_U$. Let $\hat{L}_{\mathcal{B}}$ to denote the worst case excess risk measured on the observed target samples: $\hat{L}_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta}^* \in \mathcal{B}} \mathbb{E}_{\boldsymbol{y}_S} \frac{1}{n_U} \|X_U(\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S) - \boldsymbol{\beta}^*)\|^2$. To find the best linear estimator that minimizes $\hat{L}_{\mathcal{B}}$, our proposed algorithm becomes:

$$\hat{C} \leftarrow \min_{\tau, C} \left\{ r^2\tau + \frac{\sigma^2}{n_S} \mathrm{Tr}(C \hat{\Sigma}_S^{-1} C^\top \hat{\Sigma}_U) \right\}, \quad (4)$$

[6] Note this goes without saying that our method can also be orderwise better than ordinary least square, which is a special case of ridge regression by setting $\lambda = 0$.

$$\text{s.t. } (C-I)^\top \hat{\Sigma}_U (C-I) \preceq \tau I.$$

Let $\hat{\boldsymbol{\beta}} = \hat{C}\hat{\Sigma}_S^{-1} X_S^\top \boldsymbol{y}_S / n_S$. We want to show that in spite of the existence of estimation error due to the replacement of Σ_T with $\hat{\Sigma}_T$, our generated $\hat{\boldsymbol{\beta}}$ still achieves minimax linear risk (up to constant multiplicative error).

For simplicity, in this section we assume input samples are centered: $\mathbb{E}_{p_S}[\boldsymbol{x}] = \mathbb{E}_{p_T}[\boldsymbol{x}] = 0$. This assumption results in no loss of generality. Since the sample mean is more sample-efficient to estimate than covariance matrix, one will be able to first estimate the mean and center the data. We assume some standard light-tail property on the target samples:

Definition 3.7 (ρ^2-subgaussian distribution)**.** *We call a distribution $p, \mathbb{E}[p] = 0$ to be ρ^2-subgaussianwhen there exists $\rho > 0$ such that the random vector $\bar{\boldsymbol{x}} \sim \bar{p}$ is ρ^2-subgaussian. $\bar{p}$ is the whitening of p such that $\bar{\boldsymbol{x}} \sim \bar{p}$ is equivalent to $\boldsymbol{x} = \Sigma^{1/2}\bar{\boldsymbol{x}} \sim p$, where $\Sigma = \mathbb{E}_p[\boldsymbol{x}\boldsymbol{x}^\top]$.* [7]

Note that ρ is defined on the whitening of the data. It doesn't scale with $\|\Sigma\|_{op}$ and should be viewed as universal constant.

Theorem 3.8. *Fix a failure probability $\delta \in (0,1)$. Suppose target distribution p_T is ρ^2-subgaussian, and the sample size in target domain satisfies $n_U \gg \rho^4(d + \log\frac{1}{\delta})$. Let $\hat{\boldsymbol{\beta}} : \boldsymbol{y}_S \to \hat{C}\hat{\Sigma}_S^{-1} X_S^\top \boldsymbol{y}_S$ where $\hat{C}$ is defined from Eqn. (4). Then with probability at least $1-\delta$ over the unlabeled samples from target domain, and for each fixed X_S from source domain, our learned estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S)$ satisfies:*

$$L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) \le (1 + O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n}}))R_L(\mathcal{B}). \quad (5)$$

When Σ_T commutes with $\hat{\Sigma}_S$ or is rank 1, we have:

$$L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}) \le (1.25 + O(\sqrt{\frac{\rho^4(d+\log(1/\delta))}{n}}))R_N(\mathcal{B}). \quad (6)$$

Similarly all other results in the paper could be extended to $\hat{\boldsymbol{\beta}} \leftarrow \arg\min \hat{L}_{\mathcal{B}}(\cdot)$, the estimator obtained with finite target samples X_U.

Remark 3.2 (Incorporating the randomness from source data)**.** *For linear estimators, it naturally considers X_S as fixed and Theorem 3.4 is comparing our estimator with the optimal nonlinear estimator using the same data X_S from the source domain. In Appendix D, we compare our estimator with an even stronger linear estimator with infinite access to p_S and show that our estimator is still within multiplicative factor of it.*

[7] A random vector $\boldsymbol{x}$ is called ρ^2-subgaussian if for any fixed unit vector $\boldsymbol{v}$ of the same dimension, the random variable $\boldsymbol{v}^\top \boldsymbol{x}$ is ρ^2-subgaussian, i.e., $\mathbb{E}[e^{s\cdot \boldsymbol{v}^\top(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])}] \le e^{s^2\rho^2/2}$ ($\forall s \in \mathbb{R}$).

3.4. Utilize source and target labeled data jointly

In some scenarios, we have moderate amount of labeled data from target domain as well. In such cases, it is important to utilize the source and target labeled data jointly. Let $\boldsymbol{y}_S = X_S\boldsymbol{\beta}^* + \boldsymbol{z}_S$, $\boldsymbol{y}_T = X_T\boldsymbol{\beta}^* + \boldsymbol{z}_T$. We consider X_S, X_T as deterministic variables, $\hat{\Sigma}_S^{-1}X_S^\top \boldsymbol{y}_S/n_S \sim \mathcal{N}(\boldsymbol{\beta}^*, \frac{\sigma^2}{n_S}\hat{\Sigma}_S^{-1})$ and $\hat{\Sigma}_T^{-1}X_T^\top \boldsymbol{y}_T/n_T \sim \mathcal{N}(\boldsymbol{\beta}^*, \frac{\sigma^2}{n_T}\hat{\Sigma}_T^{-1})$. Therefore conditioned on the observations $\boldsymbol{y}_S, \boldsymbol{y}_T$, a sufficient statistic for $\boldsymbol{\beta}^*$ is $\hat{\boldsymbol{\beta}}_{\text{SS}} := (n_S\hat{\Sigma}_S + n_T\hat{\Sigma}_T)^{-1}(X_S^\top \boldsymbol{y}_S + X_T^\top \boldsymbol{y}_T)$.

Claim 3.9. *$\hat{\boldsymbol{\beta}}_{SS}$ is an unbiased sufficient statistic of $\boldsymbol{\beta}^*$ with samples $\boldsymbol{y}_S, \boldsymbol{y}_T$. $\hat{\boldsymbol{\beta}}_{SS} \sim \mathcal{N}(\boldsymbol{\beta}^*, \sigma^2(n_S\hat{\Sigma}_S + n_T\hat{\Sigma}_T)^{-1})$.*

Algorithm: First consider the estimator $\hat{\boldsymbol{\beta}}_{\text{SS}} = (n_S\hat{\Sigma}_S + n_T\hat{\Sigma}_T)^{-1}(X_S^\top \boldsymbol{y}_S + X_T^\top \boldsymbol{y}_T)$. Next find the best linear function of $\hat{\boldsymbol{\beta}}_{\text{SS}}$:

$$\hat{\boldsymbol{\beta}}_{\text{MM}} = \arg\min_{C,\tau} r^2\tau + \sigma^2 \text{Tr}((n_S\hat{\Sigma}_S + n_T\hat{\Sigma}_T)^{-1}C^\top\Sigma_T C),$$
$$\text{s.t. } (C-I)^\top\Sigma_T(C-I) \preceq \tau.$$

Proposition 3.10. *The minimax estimator $\hat{\boldsymbol{\beta}}_{MM}$ is of the form $C\hat{\boldsymbol{\beta}}_{SS}$ for some C. When choosing C with our proposed algorithm and when $\hat{\Sigma}_S$ commutes with $\hat{\Sigma}_T$ and Σ_T, we achieve the minimax risk $R_L(\mathcal{B}) \le 1.25R_N(\mathcal{B})$.*

4. Covariate shift with approximation error

Now we consider observations coming from nonlinear models: $\boldsymbol{y}_S = f^*(X_S) + \boldsymbol{z}$. Let $\boldsymbol{\beta}_S^* = \arg\min_{\boldsymbol{\beta}} \mathbb{E}_{\boldsymbol{x}\sim p_S, z\sim\mathcal{N}(0,\sigma^2)}[(f^*(\boldsymbol{x}) + z - \boldsymbol{\beta}^\top\boldsymbol{x})^2]$, and similarly for $\boldsymbol{\beta}_T^*$. Notice now even with f^* unchanged across domains, the input distribution affects the best linear model. Approximation error on source domain is $a_S(\boldsymbol{x}) := f^*(\boldsymbol{x}) - \boldsymbol{x}^\top\boldsymbol{\beta}_S^*$ and vice versa for a_T.

Define the reweighting vector $\boldsymbol{w} \in \mathbb{R}^n$ as $w_i = p_T(\boldsymbol{x}_i)/p_S(\boldsymbol{x}_i)$. We form an unbiased estimator via

$$\hat{\boldsymbol{\beta}}_{\text{LS}} = \arg\min_{\boldsymbol{\beta}}\{\sum_i \frac{p_T(\boldsymbol{x}_i)}{p_S(\boldsymbol{x}_i)}(\boldsymbol{\beta}^\top\boldsymbol{x}_i - y_i)^2\}$$
$$= (X_S^\top \text{diag}(\boldsymbol{w})X_S)^{-1}(X_S^\top \text{diag}(\boldsymbol{w})\boldsymbol{y}_S).$$

Claim 4.1. *$\hat{\boldsymbol{\beta}}_{LS}$ is asymptotically unbiased and normally distributed with covariance matrix $M := \Sigma_T^{-1}\,\mathbb{E}_{\boldsymbol{x}\sim p_T}[\frac{p_T(\boldsymbol{x})}{p_S(\boldsymbol{x})}(a_T(\boldsymbol{x})^2 + \sigma^2)\boldsymbol{x}\boldsymbol{x}^\top]\Sigma_T^{-1}$:*

$$\sqrt{n_S}(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}_T^*) \xrightarrow{d} \mathcal{N}(0, M).$$

Note that large importance weights greatly inflates the variance of the estimator, especially when p_T/p_S blows up somewhere. Therefore here we design the an algorithm to cope with the inflated variance. Again we want to minimize

the worst case risk:

$$\min_{\hat{\boldsymbol{\beta}}=C\hat{\boldsymbol{\beta}}_{\text{LS}}} \max_{\boldsymbol{\beta}_T^*\in\mathcal{B}} \mathbb{E}\,\|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_T^*)\|^2$$
$$\xrightarrow{d} \min_C \max_{\|\boldsymbol{\beta}_T^*\|\le r} \left\{\|\Sigma_T^{1/2}(C-I)\boldsymbol{\beta}_T^*\|_2^2 + \frac{1}{n_S}\text{Tr}(CMC^\top\Sigma_T)\right\}$$
$$= \min_C \left\{\|(C-I)^\top\Sigma_T(C-I)\|_2 r^2 + \frac{1}{n_S}\text{Tr}(CMC^\top\Sigma_T)\right\}$$

With $\hat{\beta}_{\text{LS}}$ computed beforehand, one could first estimate M by let $\hat{M} := \frac{1}{n_S}\sum_i \Sigma_T^{-1}\frac{p_T^2(\boldsymbol{x})}{p_S^2(\boldsymbol{x})}(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{\text{LS}})^2\boldsymbol{x}_i\boldsymbol{x}_i^\top\Sigma_T^{-1}$. Therefore our estimator is $\hat{\boldsymbol{\beta}}_{\text{MM}} \leftarrow \hat{C}\hat{\boldsymbol{\beta}}_{\text{LS}}$, where $\hat{C}$ finds

$$\hat{C} \leftarrow \arg\min_{\tau,C}\left\{r^2\tau + \frac{1}{n_S}\text{Tr}(C\hat{M}C^\top\Sigma_T)\right\} \quad (7)$$
$$\text{s.t. } (C-I)^\top\Sigma_T(C-I) \preceq \tau I.$$

Claim 4.2. *Let $\mathcal{B} = \{\boldsymbol{\beta}|\|\boldsymbol{\beta}\| \le r\}$, and $f^* \in \mathcal{F}$ is some compact symmetric function class: $f \in \mathcal{F} \Leftrightarrow -f \in \mathcal{F}$. Then linear minimax estimator is of the form $C\hat{\boldsymbol{\beta}}_{LS}$ for some C. When $\hat{C}$ solves Eqn. (7), $L_{\mathcal{B}}(\hat{\boldsymbol{\beta}}_{MM})$ asymptotically matches $R_L(\mathcal{B})$, the linear minimax risk.*

By reducing from $\boldsymbol{y}_S$ to $\hat{\boldsymbol{\beta}}_{\text{LS}}$ we eliminate $n-d$ dimensions, and this claim says that $X_S^\top\boldsymbol{y}_S$ is sufficient to predict $\boldsymbol{\beta}_T^*$. We note that f^* is more general than a linear function and therefore the lower bound could only be larger than $R_N(\mathcal{B})$ defined in the previous section.

4.1. Estimating $p_T(\boldsymbol{x})/p_S(\boldsymbol{x})$

Even though estimating $p_T(\boldsymbol{x})/p_S(\boldsymbol{x})$ might be sample inefficient, it only involves unlabeled data and therefore instance weighting related algorithms still attract prior studies as demonstrated in the related work section. Practical ways to estimate the density ratio involve respectively estimating p_T and p_S (Lin et al., 2002; Zadrozny, 2004), kernel mean matching (KMM) (Huang et al., 2006)), and some common divergence minimization between weighted source distribution and target distribution (Sugiyama et al., 2008; 2012; Uehara et al., 2016; Menon & Ong, 2016; Kanamori et al., 2011).We propose another simple algorithm that is very convenient to use.

We conduct regression on the data samples $(\boldsymbol{x}, y) \sim q(\boldsymbol{x}, y)$ where $q_Y(y)$ is Bernouli($\frac{1}{2}$)[8] and $q_{X|Y}(\boldsymbol{x}|y=1) = p_T$, $q_{X|Y}(\boldsymbol{x}|y=0) = p_S$. Empirically, we will concatenate X_S and X_U to form input data and stack $\mathbf{0} \in \mathbb{R}^{n_S}$ and $\mathbf{1} \in \mathbb{R}^{n_U}$ as the target vector $\boldsymbol{y}$.

Proposition 4.3. *The optimal function that solves $\alpha \leftarrow \arg\min_f \mathbb{E}_{\boldsymbol{x},y\sim q}(f(\boldsymbol{x})-y)^2$ satisfies: $\alpha(\boldsymbol{x}) = \frac{p_T(\boldsymbol{x})}{p_S(\boldsymbol{x})+p_T(\boldsymbol{x})}$.*

[8]The scalar $1/2$ should be adjusted based on the number of unlabeled samples from source and target domain.

Therefore with proper transformation[9] on α one could get the importance weights. In practice, one might be flexible on choosing the function class $\mathcal{F}$ for estimating α and sample complexity will be bounded by some standard measure of $\mathcal{F}$'s complexity, e.g., Rademacher or Gaussian complexity (Bartlett & Mendelson, 2002). Unlike KMM, this parametrized estimation applies to unseen data $\boldsymbol{x}$ which makes cross-validation possible.

5. Near minimax estimator with model shift

The general setting of transfer learning in linear regression involves both model shift and covariate shift. Namely, the generative model of the labels might be different: $\boldsymbol{y}_S = X_S\boldsymbol{\beta}_S^* + \boldsymbol{z}_S$, and $\boldsymbol{y}_T = X_T\boldsymbol{\beta}_T^* + \boldsymbol{z}_T$. Denote by $\boldsymbol{\delta} := \boldsymbol{\beta}_S^* - \boldsymbol{\beta}_T^*$ as the model shift. We are interested in the minimax linear estimator when $\|\boldsymbol{\delta}\| \le \gamma$ and $\|\boldsymbol{\beta}_T^*\| \le r$. Thus our problem becomes to find minimax estimator for $\boldsymbol{\beta}_T^* \in \mathcal{B} = \{\boldsymbol{\beta}|\|\boldsymbol{\beta}\| \le r\}$ from $\boldsymbol{y}_S, \boldsymbol{y}_T$.

Algorithm: First consider a sufficient statistic $(\bar{\boldsymbol{\beta}}_S, \bar{\boldsymbol{\beta}}_T)$ for $(\boldsymbol{\beta}_T^*, \boldsymbol{\delta})$. Here $\bar{\boldsymbol{\beta}}_S = \hat{\Sigma}_S^{-1}X_S^\top\boldsymbol{y}_S/n_S \sim \mathcal{N}(\boldsymbol{\beta}_T^* + \boldsymbol{\delta}, \frac{\sigma^2}{n_S}\hat{\Sigma}_S^{-1})$, and $\bar{\boldsymbol{\beta}}_T = \hat{\Sigma}_T^{-1}X_T^\top\boldsymbol{y}_T/n_T \sim \mathcal{N}(\boldsymbol{\beta}_T^*, \frac{\sigma^2}{n_T}\hat{\Sigma}_T^{-1})$. Then consider the best linear estimator on top of it: $\hat{\boldsymbol{\beta}} = A_1\bar{\boldsymbol{\beta}}_S + A_2\bar{\boldsymbol{\beta}}_T$. Write $\Delta = \{\boldsymbol{\delta}|\|\boldsymbol{\delta}\| \le \gamma\}$ and $L_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}}) := \max_{\boldsymbol{\beta}_T^*\in\mathcal{B},\boldsymbol{\delta}\in\Delta}\|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_T^*)\|^2$.

$$R_L(\mathcal{B},\Delta) := \min_{\hat{\boldsymbol{\beta}}=A_1\bar{\boldsymbol{\beta}}_S+A_2\bar{\boldsymbol{\beta}}_T} L_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}})$$
$$\le \min_{A_1,A_2}\max_{\|\boldsymbol{\beta}_T^*\|\le r,\|\boldsymbol{\delta}\|\le\gamma}\Big\{2\|\Sigma_T^{1/2}((A_1+A_2-I)\boldsymbol{\beta}_T^*\|^2$$
$$+2\|\Sigma_T^{1/2}A_1\boldsymbol{\delta}\|^2 + \frac{\sigma^2}{n_S}\text{Tr}(A_1\hat{\Sigma}_S^{-1}A_1^\top) \quad (8)$$
$$+\frac{\sigma^2}{n_T}\text{Tr}(A_2\hat{\Sigma}_T^{-1}A_2^\top)\Big\} \quad \text{(AM-GM)}$$
$$= \min_{A_1,A_2}\Big\{2\|\Sigma_T^{1/2}((A_1+A_2-I)\|_2^2 r^2 + 2\|\Sigma_T^{1/2}A_1\|_2^2\gamma^2$$
$$+\frac{\sigma^2}{n_S}\text{Tr}(A_1\hat{\Sigma}_S^{-1}A_1^\top) + \frac{\sigma^2}{n_T}\text{Tr}(A_2\hat{\Sigma}_T^{-1}A_2^\top)$$
$$=: r_{\mathcal{B},\Delta}(A_1,A_2)\Big\}. \quad (9)$$

Therefore we optimize over this upper bound and reformulate the problem as a convex program:

$$(\hat{A}_1,\hat{A}_2) \leftarrow \arg\min_{A_1,A_2,a,b}\Big\{2ar^2 + 2b\gamma^2$$
$$+\frac{\sigma^2}{n_S}\text{Tr}(A_1\hat{\Sigma}_S^{-1}A_1^\top) + \frac{\sigma^2}{n_T}\text{Tr}(A_2\hat{\Sigma}_T^{-1}A_2^\top)\Big\}$$
$$\text{s.t.}(A_1+A_2-I)^\top\Sigma_T(A_1+A_2-I) \preceq aI,$$
$$A_1^\top\Sigma_T A_1 \preceq bI. \quad (10)$$

[9]Apply $f(x) \to \frac{1}{1/f(x)-1}$

Our estimator is given by: $\hat{\boldsymbol{\beta}}_{\text{MM}} = \hat{A}_1\bar{\boldsymbol{\beta}}_S + \hat{A}_2\bar{\boldsymbol{\beta}}_T$. Since $\hat{\boldsymbol{\beta}}_{\text{MM}}$ is a relaxation of the linear minimax estimator, it is important to understand how well $\hat{\boldsymbol{\beta}}_{\text{MM}}$ performs on the original objective:

Claim 5.1. $R_L(\mathcal{B}, \Delta) \le L_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}}_{MM}) \le 2R_L(\mathcal{B}, \Delta)$.

Finally we show with the relaxation we still achieve a near-optimal estimator even among all nonlinear rules.

Theorem 5.2. *When Σ_T commutes with $\hat{\Sigma}_S$, it satisfies:*

$$L_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}}_{MM}) := \max_{\boldsymbol{\beta}_T^* \in \mathcal{B}, \boldsymbol{\delta} \in \Delta} \|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}}_{MM} - \boldsymbol{\beta}_T^*)\|^2 \le 27 R_N(\mathcal{B}, \Delta).$$

Here $R_N(\mathcal{B}, \Delta) := \min_{\hat{\boldsymbol{\beta}}(\boldsymbol{y}_S, \boldsymbol{y}_T)} \max_{\boldsymbol{\beta}_T^* \in \mathcal{B}, \boldsymbol{\delta} \in \Delta} \|\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_T^*)\|$ is the minimax risk.

We defer the complete proof to the appendix. The main proof technique is to decompose the problem to 2-d subproblems with closed-form solutions and are solvable with Le Cam's two point lemma. We include the proof sketch here:

Proof sketch of Theorem 5.2. For the ease of understanding, we provide a simple proof sketch when $\Sigma_S = \Sigma_T$ are diagonal. We first define the hardest hyperrectangular subproblem. Let $\mathcal{B}(\boldsymbol{\tau}) = \{\boldsymbol{b} : |\beta_i| \le \tau_i\}$ be a subset of $\mathcal{B}$ and similarly for $\Delta(\boldsymbol{\zeta})$. We show that $R_L(\mathcal{B}, \Delta) = \max_{\boldsymbol{\tau} \in \mathcal{B}, \boldsymbol{\zeta} \in \Delta} R_L(\mathcal{B}(\boldsymbol{\tau}), \Delta(\boldsymbol{\zeta}))$, and clearly $R_N(\mathcal{B}, \Delta) \ge \max_{\boldsymbol{\tau} \in \mathcal{B}, \boldsymbol{\zeta} \in \Delta} R_N(\mathcal{B}(\boldsymbol{\tau}), \Delta(\boldsymbol{\zeta}))$. Meanwhile we show when the sets are hyperrectangles the minimax (linear) risk could be decomposed to 2-d problems: $R_L(\mathcal{B}(\boldsymbol{\tau}), \Delta(\boldsymbol{\zeta})) = \sum_i R_L(\tau_i, \zeta_i)$. Each $R_L(\tau_i, \zeta_i)$ is the linear minimax risk to estimate β_i from $x \sim \mathcal{N}(\beta_i + \delta_i, 1)$ and $y \sim \mathcal{N}(\beta_i, 1)$ where $|\beta_i| \le \tau_i$ and $|\delta_i| \le \zeta_i$. This 2-d problem for linear risk has a closed form solution, and the minimax risk can be lower bounded using Le Cam's two point lemma. We show $R_L(\tau_i, \zeta_i) \le 13.5 R_N(\tau_i, \zeta_i)$ and therefore:

$$\begin{aligned}
\frac{1}{2} L_{\mathcal{B},\Delta}(\hat{\boldsymbol{\beta}}_{\text{MM}}) &\overset{\text{Claim } 5.1}{\le} R_L(\mathcal{B}, \Delta) \\
&\overset{\text{Lemma } C.2}{=} \max_{\boldsymbol{\tau} \in \mathcal{B}, \boldsymbol{\zeta} \in \Delta} R_L(\mathcal{B}(\boldsymbol{\tau}), \Delta(\boldsymbol{\zeta})) \\
&\overset{\text{Prop } C.4.a}{=} \max_{\boldsymbol{\tau} \in \mathcal{B}, \boldsymbol{\zeta} \in \Delta} \sum_i R_L(\tau_i, \zeta_i) \\
&\overset{\text{Lemma } C.6}{\le} \max_{\boldsymbol{\tau} \in \mathcal{B}, \boldsymbol{\zeta} \in \Delta} 13.5 \sum_i R_N(\tau_i, \zeta_i) \\
&\overset{\text{Prop } C.4.b}{=} 13.5 \max_{\boldsymbol{\tau} \in \mathcal{B}, \boldsymbol{\zeta} \in \Delta} R_N(\mathcal{B}(\boldsymbol{\tau}), \Delta(\boldsymbol{\zeta})) \\
&\le 13.5 R_N(\mathcal{B}, \Delta).
\end{aligned}$$

□

6. Experiments

Our estimators are provably near optimal for the worst case $\boldsymbol{\beta}^*$. However, it remains unknown whether on average they outperform other baselines. With synthetic data we explore the performances with random $\boldsymbol{\beta}^*$. We are also interested to investigate the conditions when we win more.

Setup. We set $n_S = 2000, d = 50, \sigma = 1, r = \sqrt{d}$. For each setting, we sample $\boldsymbol{\beta}_T^*$ from standard normal distribution and rescale it to be norm r. We estimate Σ_T by $n_U = 2000$ unlabeled samples. We compare our estimator with ridge regression (S-ridge) and a variant of ridge regression transformed to target domain (T-ridge): $\hat{\boldsymbol{\beta}}^{\lambda}_{\text{RR},T} = \arg\min \frac{1}{n}\|\Sigma_T^{1/2}(\boldsymbol{\beta} - \hat{\Sigma}_S^{-1} X_S^\top \boldsymbol{y}_S)\|^2 + \lambda\|\boldsymbol{\beta}\|^2 = (\Sigma_T + \lambda I)^{-1}\Sigma_T \hat{\boldsymbol{\beta}}_{\text{SS}}$.

Covariate shift. In order to understand the effect of covariate shift on our algorithm, we consider three types of settings, each with a unique varying factor that influences the performance: 1) covariate eigenvalue shift with shared eigenspace; 2) covariate eigenspace shift with fixed eigenvalues[10]; 3) signal strength change. We also have an additional 200 labeled data from target domain as validation set only for hyper-parameter tuning.

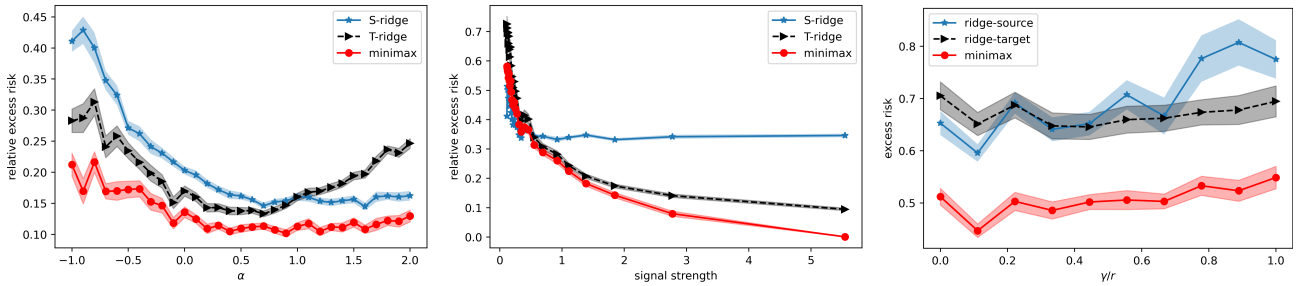
Model shift. Next we consider the problem with model shift. We sample a random $\boldsymbol{\delta}$ with norm γ varying from 0 to $r = \sqrt{d}$ and observe data generated by $\boldsymbol{y}_S = X_S(\boldsymbol{\beta}_T^* + \boldsymbol{\delta}) + \boldsymbol{z}_S \in \mathbb{R}^{2000}, \boldsymbol{z}_S \sim \mathcal{N}(0, I)$ and $\boldsymbol{y}_T = X_T\boldsymbol{\beta}_T^* + \boldsymbol{z}_T \in \mathbb{R}^{500}, \boldsymbol{z}_T \sim \mathcal{N}(0, I)$. We compare our estimator with two baselines: "ridge-source" denotes ridge regression using only source data, and "ridge-target" is from ridge regression with target data.

Figure 1 demonstrates the better performance of our estimator in all circumstances. From (a) we see that with more discrepancy between Σ_S and Σ_T, our estimator tends to perform better. (b) shows our estimator is better when the signal is relatively stronger. From (c) we can see that with the increasing model shift measured by γ/r, S-ridge becomes worse and is outperformed by T-ridge that remains unchanged. Our estimator becomes slightly worse as well due to the less utility from source data, but remains the best among others. When $\gamma/r \approx 0.2$, our method has the most improvement in percentage compared to the best result among ridge-source and ridge-target.

6.1. Experiments with approximation error

Finally, we conduct empirical studies with nonlinear models. We maintain the same setting as before. We also generate a small validation dataset from target domain: $X_{\text{CV}} \in \mathbb{R}^{500\times 50}$, sampled from $\mathcal{N}(0, \Sigma_T)$, $\boldsymbol{y}_{\text{CV}} = f^*(X_{\text{CV}}) + \boldsymbol{z}_{\text{CV}}$,

[10] We leave this result in appendix since performance appears invariant to this factor.



(a) covariate eigen-spectrum (b) signal strength (c) model shift

Figure 1: *Performance comparisons.* (a): The x-axis α defines the spread of eigen-spectrum of Σ_S: $s_i \propto 1/i^\alpha, t_i \propto 1/i$. (b) x-axis is the normalized value of signal strength: $\|\Sigma_T \boldsymbol{\beta}^*\|/r$. (c) X-axis is the model shift measured by γ/r. Performance with standard error bar is from 40 runs.

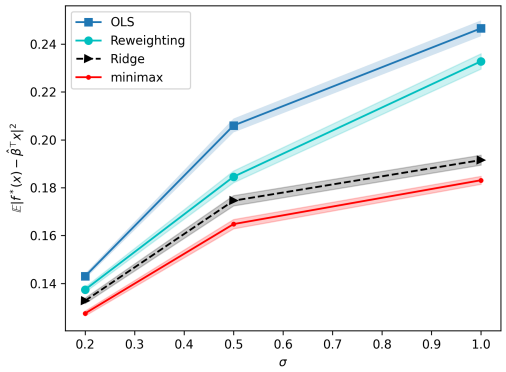


Figure 2: The x-axis is noise level σ and y-axis is the excess risk (with approximation error). Performance with standard error bar is from 40 runs.

with $\boldsymbol{z}_{\text{CV}} \sim \mathcal{N}(0, \sigma^2 I)$. We choose $\lambda_i(\Sigma_S) \propto i, \lambda_i(\Sigma_T) \propto 1/i$, and the eigenspace for both Σ_S and Σ_T are random orthonormal matrices. ($\|\Sigma_S\|_F^2 = \|\Sigma_T\|_F^2 = d$.) The ground truth model is a one-hidden-layer ReLU network: $f^*(\boldsymbol{x}) = 1/d\boldsymbol{a}^\top (W\boldsymbol{x})_+$, where W and $\boldsymbol{a}$ are randomly generated from standard Gaussian distribution. We observe noisy labels: $\boldsymbol{y}_S = f^*(\boldsymbol{x}) + \boldsymbol{z}$, where $z_i \sim \mathcal{N}(0, \sigma^2)$.

Estimating weights $p_T(\boldsymbol{x})/p_S(\boldsymbol{x})$. Since the generated data samples are Gaussian, the absolute weights for $p_T(\boldsymbol{x})/p_S(\boldsymbol{x}) = \sqrt{\frac{|\Sigma_S|}{|\Sigma_T|}} \exp(\frac{1}{2}\boldsymbol{x}^\top(\Sigma_S^{-1} - \Sigma_T^{-1})\boldsymbol{x})$. However, this absolute value scales exponentially with the norm of $\boldsymbol{x}$ and can amplify the variance. Meanwhile, when one multiplies both $X_S, \boldsymbol{y}_S$ by 10, the ground truth β doesn't change but the absolute value for $p_T(\boldsymbol{x})/p_S(\boldsymbol{x})$ will change drastically. This discrepancy highlights the importance of relative magnitudes (among samples) instead of the absolute value, as noted by Kanamori et al. (2009).

To obtain a relative score, we first estimate the absolute density ratio $\alpha(\boldsymbol{x}) \approx p_T(\boldsymbol{x})/(p_S(\boldsymbol{x}) + p_T(\boldsymbol{x}))$ following our algorithm in Section 4.1 with linear regression. We then uniformly assign the weight w_i for each sample by 10 discrete values $1, 2, 3 \cdots 10$ based on the absolute value of $\alpha(\boldsymbol{x})$ and then rescale the reweighting vector properly. We use the conventional way to adjust the reweighting strength by using $w_i^c, c \in [0, 1]$ and choose c by cross validation.

We implement our method (Eqn. 7) using the estimated weights as above, and plot the excess risk comparisons in Figure 2. The baselines we choose are ordinary least square ("OLS" in Figure (2)), ridge regression (Legend is "Ridge") and weighted least square (Kanamori et al., 2009) (Legend is "Reweighting"; $\hat{\boldsymbol{\beta}}_{\text{LS}}$ in our main text). For ridge regression, reweighting and our methods, we tune hyperparameters through cross-validation. All results are presented from 40 runs where the randomness comes from f^* and the eigenspaces of Σ_S, Σ_T. From Figure 2 we could see that reweighting algorithm improves over ordinary least square but is outperformed by ridge regression due to large variance. Our algorithm achieves the best performance among others by appropriately reweighting then reducing the variance.

Experiments on Berkeley Yearbook Dataset To verify the performance of our algorithm on real-world data, we conduct an experiment on the Berkeley Yearbook dataset (Ginosar et al., 2015). We randomly split the data to form source and target tasks, where the source has 63.2% male photos and 43.4% male images for the target task. Input X is gray-scale portraits, and Y is the year the photo is taken (ranging from 1905 to 2013). We implement our algorithms together with the baselines and estimate the density ratio from the data. We demonstrate the performance improvement in Figure 3. The x-axis is the scalar that adjusts reweighting strength c defined in the previous paragraph.

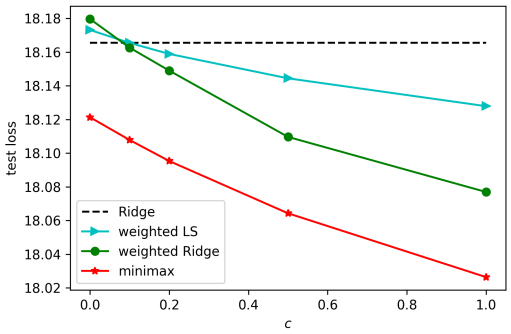


Figure 3: Comparisons on Yearbook Dataset (Ginosar et al., 2015).

7. Conclusion

We study in depth the minimax linear estimator for linear regression under various distribution shift settings. We investigated the optimal linear estimators with covariate shift for linear models in unsupervised and supervised domain adaptation settings, with no or scarce labeled data from the target distribution. For nonlinear models with approximation error, we also introduce the minimax linear estimator together with an easy-to-use density ratio estimation method. We further explore some moderate model shift in the linear setting. Our estimators achieve near-optimal worst-case excess risk measured on the target domain and, in some circumstances, are within constant of the minimax risk among all nonlinear rules. The significant improvement of our estimators over ridge regression is demonstrated by a theoretical separation result and by empirical validations even for average case with random parameters.

In future work, we will extend our algorithm to classification problems under distribution shift and apply the algorithms to fine-tuning the last-layer of a deep network.

Acknowledgements

QL was supported by NSF #2030859 and the Computing Research Association for the CIFellows Project. WH was supported by NSF, ONR, Simons Foundation, Schmidt Foundation, Amazon Research, DARPA and SRC. JDL was supported by ARO under MURI Award W911NF-11-1-0303, the Sloan Research Fellowship, and NSF CCF 2002272.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Baktashmotlagh, M., Harandi, M. T., Lovell, B. C., and Salzmann, M. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 769–776, 2013.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Blaker, H. Minimax estimation in linear regression under restrictions. *Journal of statistical planning and inference*, 90(1):35–55, 2000.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pp. 442–450, 2010.

David, S. B., Lu, T., Luu, T., and Pál, D. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 129–136, 2010.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Donoho, D. L. Statistical estimation and optimal recovery. *The Annals of Statistics*, pp. 238–270, 1994.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Ginosar, S., Rakelly, K., Sachs, S., Yin, B., and Efros, A. A. A century of portraits: A visual historical record of american high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1–7, 2015.

Glorot, X., Bordes, A., and Bengio, Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011.

Gong, B., Shi, Y., Sha, F., and Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073. IEEE, 2012.

Gong, B., Grauman, K., and Sha, F. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 222–230, 2013.

Gopalan, R., Li, R., and Chellappa, R. Domain adaptation for object recognition: An unsupervised approach. In *2011 international conference on computer vision*, pp. 999–1006. IEEE, 2011.

Hanneke, S. and Kpotufe, S. On the value of target data in transfer learning. In *Advances in Neural Information Processing Systems*, pp. 9871–9881, 2019.

Heckman, J. J. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pp. 153–161, 1979.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19:601–608, 2006.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. J. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pp. 601–608, 2007.

Jiang, J. and Zhai, C. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 264–271, 2007.

Johnstone, I. M. Gaussian estimation: Sequence and wavelet models. *Unpublished manuscript*, 2011.

Juditsky, A. B., Nemirovski, A. S., et al. Nonparametric estimation by convex programming. *The Annals of Statistics*, 37(5A):2278–2300, 2009.

Kalan, S. M. M., Fabian, Z., Avestimehr, A. S., and Soltanolkotabi, M. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *arXiv preprint arXiv:2006.10581*, 2020.

Kanamori, T., Hido, S., and Sugiyama, M. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

Kanamori, T., Suzuki, T., and Sugiyama, M. f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE transactions on information theory*, 58(2):708–720, 2011.

Kumar, A., Ma, T., and Liang, P. Understanding self-training for gradual domain adaptation. *arXiv preprint arXiv:2002.11361*, 2020.

Lin, Y., Lee, Y., and Wahba, G. Support vector machines for classification in nonstandard situations. *Machine learning*, 46(1):191–202, 2002.

Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pp. 3122–3130. PMLR, 2018.

Long, M., Wang, J., Ding, G., Sun, J., and Yu, P. S. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pp. 2200–2207, 2013.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pp. 2208–2217. PMLR, 2017.

Lu, S., Nott, B., Olson, A., Todeschini, A., Vahabi, H., Carmon, Y., and Schmidt, L. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

Menon, A. and Ong, C. S. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pp. 304–313. PMLR, 2016.

Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

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

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. 2009.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Saerens, M., Latinne, P., and Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Srivastava, M., Hashimoto, T., and Liang, P. Robustness to spurious correlations via human annotations. *arXiv preprint arXiv:2007.06661*, 2020.

Storkey, A. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pp. 3–28, 2009.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Sugiyama, M., Suzuki, T., and Kanamori, T. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

Sun, Q., Chattopadhyay, R., Panchanathan, S., and Ye, J. A two-stage weighting framework for multi-source domain adaptation. In *Advances in neural information processing systems*, pp. 505–513, 2011.

Uehara, M., Sato, I., Suzuki, M., Nakayama, K., and Matsuo, Y. Generative adversarial nets from a density ratio estimation perspective. *arXiv preprint arXiv:1610.02920*, 2016.

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

Wang, X. and Schneider, J. Flexible transfer learning under support and model shift. In *Advances in Neural Information Processing Systems*, pp. 1898–1906, 2014.

Wang, X. and Schneider, J. G. Generalization bounds for transfer learning under model shift. In *UAI*, pp. 922–931, 2015.

Wang, X., Huang, T.-K., and Schneider, J. Active transfer learning under model shift. In *International Conference on Machine Learning*, pp. 1305–1313, 2014.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.

Wu, Y., Winston, E., Kaushik, D., and Lipton, Z. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International Conference on Machine Learning*, pp. 6872–6881. PMLR, 2019.

Zadrozny, B. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 114, 2004.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pp. 819–827, 2013.

Zhao, H., Combes, R. T. d., Zhang, K., and Gordon, G. J. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*, 2019.