A Class of Geometric Structures in Transfer Learning: Minimax Bounds and Optimality

Xuhui Zhang
Stanford University

Jose Blanchet
Stanford University

Soumyadip Ghosh
IBM Research

Mark S. Squillante
IBM Research

Abstract

We study the problem of transfer learning, observing that previous efforts to understand its information-theoretic limits do not fully exploit the geometric structure of the source and target domains. In contrast, our study first illustrates the benefits of incorporating a natural geometric structure within a linear regression model, which corresponds to the generalized eigenvalue problem formed by the Gram matrices of both domains. We next establish a finite-sample minimax lower bound, propose a refined model interpolation estimator that enjoys a matching upper bound, and then extend our framework to multiple source domains and generalized linear models. Surprisingly, as long as information is available on the distance between the source and target parameters, negative-transfer does not occur. Simulation studies show that our proposed interpolation estimator outperforms state-of-the-art transfer learning methods in both moderate- and high-dimensional settings.

1 INTRODUCTION

The task of transferring knowledge from one domain (source) to another related domain (target) is known as transfer learning. This task arises naturally in a wide range of applications where data is scarce in the target domain but substantial in a source domain believed to be somewhat similar to the target. For example, in the context of marketing and demand prediction for products in a new market, it is natural to use source information involving well-tested markets. Similarly, demand prediction for new products can be estimated using source information from current market products (Afrin et al., 2018). Rigorous statistical formulations of transfer learning introduce non-trivial challenges. This includes balancing the tension between tractability in the training procedure and flexibility in order to reflect the differences between the source and target environments. In addition to this tension, a useful modeling framework should provide statistical insights on the efficiency gain induced by introducing source information into target inference.

Our goal in this paper is to introduce a transfer learning formulation under linear and generalized linear models that addresses the above tractability-flexibility tension and produces effective efficiency insights. In particular, the contributions of this paper include:

(i) Our formulation provides an easy-to-compute transfer learning estimator that optimally (in a precise sense) interpolates the target and source parameters subject to an uncertainty region which controls the differences between source and target models.

(ii) Our modelling framework exposes a natural geometric structure that is built on using the Fisher information metric (also known as information geometry) which we exploit in order to understand the main drivers of transfer learning from source to target.

(iii) We are able to provide a finite-sample minimax lower bound and show that the worst-case risk of our estimator in (i) achieves (up to a constant) the minimax lower bound uniformly over the magnitude of the difference between the target and source models.

One of the insights from our formulation, for example, is that as long as information is available on the distance between the source and target parameters, negative-transfer does not occur. Namely, the worst-case risk of our estimator in (i) is always smaller than the minimax risk of using the target dataset alone.

Although there is a significant amount of literature on applied transfer learning procedures (Pan and Yang, 2009; Torrey and Shavlik, 2010; Weiss et al., 2016; Taskesen et al., 2021), the literature on rigorous math-

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

ematical formulations that lead to minimax optimal estimators is limited. Cai and Wei (2021) consider transfer learning in the context of a stylized non-parametric classification setting under a different set of assumptions including a model that only allows posterior drift. Kpotufe and Martinet (2021) study classification settings similar to that of Cai and Wei (2021), albeit under covariate-shift assumptions on the difference in source and target environments. In strong contrast, while our methodology is parametric, our analysis is not limited to classification problems and it further supports more general environments including more general drift conditions between the target and source models, where indeed our analysis of the underlying geometry explicitly accounts for these differences.

Bastani (2021) considers linear and non-linear regression models similar to our work, and proposes a two-step joint estimator for transfer learning. However, their focus is on high-dimensional settings under sparsity assumptions and no minimax optimality result is established; moreover, since they use an l_1 norm for the analysis, their results are not directly comparable to our results. Li et al. (2020) extend the work of Bastani (2021) to allow multiple source domains and establish a minimax lower bound for high-dimensional linear regression models (LRMs). However, their bound is asymptotic in the number of target samples, and they further constrain the difference between the source and target parameters relative to the size of the target sample. Tian and Feng (2021) further extend this work of Li et al. (2020) to high-dimensional generalized linear models (GLMs). In strong contrast, we study a fixed-difference environment without any such constraint and the performance of our optimal estimator is measured in terms of finite-sample bounds which match (up to a computable constant factor) the minimax lower bound uniformly over the magnitude of the difference between the models.

Kalan et al. (2020) consider the LRM setting and involve the spectral gap of the generalized eigenvalue problem we consider, with definitions for the population distribution of their random-design setting analogous to ours in the fixed-design setting. However, in strong contrast, our analysis handles the entire spectrum of the generalized eigenvalues and our results essentially provide a tighter lower bound in comparison with the results of Kalan et al. (2020), thus illustrating the significance of our geometric perspective.

A simulation study compares our estimator to the work of Bastani (2021), Li et al. (2020) and more direct methods commonly used in practice (e.g., pooling all available data). As our estimator is designed in a fixed-dimension setting, the simulation results confirm the strong performance of our approach in moderate dimensions. Moreover, a heuristic modification of our methods using an l_1 regularization, as in Li et al. (2020) and Bastani (2021), seems to provide further improvements over these methods in the sparse high-dimensional case. We plan to investigate these types of modifications as part of our future research. Finally, results based on a real-world dataset further confirm the strong performance of our approach.

To briefly summarize our approach, let us consider the linear regression transfer learning problem between a source and a target with Gaussian errors. We start the construction of our estimator by reparameterizing the linear regression estimators. Instead of expressing them in terms of the canonical bases, as is customary, we express them in terms of a generalized eigenvalue problem that arises when computing the Fisher-Rao distance using the design matrices of the source and target. This distance induces a Riemmanian geometric structure between the models. Specifically, convex combinations in the reparameterized (Riemmanian) space correspond to general interpolations in the (original space of) LRM parameters.

Our estimator is obtained from the class generated by the convex combination of estimators for the source and target in the reparameterized space. Then, we obtain the minimax estimator by maximizing over models within a given distance while minimizing over our chosen class of estimators. The minimization is carried out over convex combination parameters which are different for each coefficient. The final estimator is transformed back to the canonical basis.

For our minimax lower bound, we rely once again on the Fisher information metric and the reparameterization used in the design of our estimator. We then apply Le Cam's two-point minimax method which in our case reduces to studying two carefully designed hypotheses for each coefficient.

Section 2 presents our mathematical framework and theoretical results for transfer learning under LRMs. Then, Section 3 extends our framework to support a class of GLMs as well as multiple sources. Section 4 presents simulation results showing that our algorithm outperforms various transfer learning methods. The supplement contains additional theoretical and simulation results, related technical details, and all proofs.

2 LINEAR REGRESSION MODELS

We consider our mathematical framework for transfer learning within the context of LRMs. We introduce our framework and establish our theoretical results on minimax bounds in Section 2.1, where we also propose a refined model interpolation estimator that is

minimax optimal. Then, in Section 2.2, we compare the minimax bounds to the basic approaches discussed in Daumé (2007), showing the latter to be suboptimal.

2.1 Mathematical Framework

Let $\mathcal{X}$ denote a d-dimensional feature space, $\mathcal{Y}$ a response space, and $\mathcal{N}(\mu, \sigma^2)$ the normal distribution with mean μ and variance σ^2. Our base source LRM can then be formally written as

$$y_i = x_i^\top \theta_S + \epsilon_i, \qquad i \in [n_S], \tag{1}$$

where $\theta_S \in \mathbb{R}^d$ is the regression coefficient for the source model, n_S is the number of samples from the source model, $x_i \in \mathcal{X}$ are (fixed) designs and $y_i \in \mathcal{Y}$ are independent response samples, $\epsilon_i \sim \mathcal{N}(0, \sigma_S^2)$ are independent noise random variables for $i \in [n_S]$, and $[n] := \{1, \ldots, n\}$. Similarly, our base target LRM can be formally written as

$$v_i = w_i^\top \theta_T + \eta_i, \qquad i \in [n_T], \tag{2}$$

where $\theta_T \in \mathbb{R}^d$ is the regression coefficient for the target model, n_T is the number of samples from the target model, $w_i \in \mathcal{X}$ are (fixed) designs and $v_i \in \mathcal{Y}$ are independent response samples, and $\eta_i \sim \mathcal{N}(0, \sigma_T^2)$ are independent noise random variables for $i \in [n_T]$. We denote the distribution of the models (1) and (2) by P_S and P_T, respectively.

For ease of exposition, we collect the designs and responses in (1) from the source domain into the design matrix $\mathbf{X}$ and the response vector $\mathbf{Y}$, respectively; i.e., $x_i^\top$ is the i-th row of $\mathbf{X}$ and y_i is the i-th element of $\mathbf{Y}$. Similarly, we collect the target designs and responses in (2) into the design matrix $\mathbf{W}$ and the response vector $\mathbf{V}$, respectively. With $\epsilon := (\epsilon_i)_{i \in [n_S]}$ and $\eta := (\eta_i)_{i \in [n_T]}$, we can then write the LRM as

$$\mathbf{Y} = \mathbf{X}\theta_S + \epsilon, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_S^2), \tag{3}$$

$$\mathbf{V} = \mathbf{W}\theta_T + \eta, \qquad \eta_i \sim \mathcal{N}(0, \sigma_T^2). \tag{4}$$

We consider fixed-design matrices to be (arbitrarily) different for the source and target domains, thus focusing on a combination of concept drift and a version of covariate shift best suited to the fixed-design setting, which is similar to Bastani (2021).

Our interests lie in estimating the regression coefficient θ_T for the target domain. Hence, we assume that the noise variance σ_T^2 is known and then we consider the family of distributions (4) parameterized by θ_T as a manifold. Various geometries can be defined on a manifold of statistical models (Nielsen, 2020), among which the Fisher-Riemannian manifold given by the Fisher information metric tensor is particularly useful, representing the unique invariant metric tensor under Markov embeddings up to a scaling constant (Campbell, 1986). For the Gaussian location model (4), we can compute the Fisher information matrix with respect to θ_T, up to a scaling constant, as $\mathbf{W}^\top \mathbf{W}$.

Given an estimate $\widehat{\theta}$ of θ_T, we consider the Riemannian geodesic metric distance (or the Fisher-Rao distance) as a principled way to measure the dissimilarity of $\widehat{\theta}$ to the (unknown) ground truth θ_T. We therefore define the loss function $\ell(\widehat{\theta}, \theta_T)$ as

$$\ell(\widehat{\theta}, \theta_T) = (\widehat{\theta} \quad \theta_T)^\top (\mathbf{W}^\top \mathbf{W})(\widehat{\theta} \quad \theta_T). \tag{5}$$

Such a loss function was used and termed "prediction loss" in Lee and Courtade (2020) without the above geometrical motivation. However, as we will show, the geometric structure of the source and target models can play an important role in transfer learning, in particular, the discrepancy between the Fisher-Rao distances induced by the models (3) and (4).

We aim to employ minimax theory to establish the optimality of statistical learning procedures. Given an estimator $\widehat{\theta}$ arising from any learning procedure, we consider its worst-case risk over an uncertainty set of plausible distributions; refer to (Tsybakov, 2008, Chapter 2). In the transfer learning setting, one natural uncertainty set is given by all source and target distributions whose dissimilarity is upper-bounded. However, instead of the l_0 or l_1-norm typically used in the high-dimensional setting (Tian and Feng, 2021; Bastani, 2021; Li et al., 2020), we characterize the dissimilarity between θ_S and θ_T in terms of the Fisher-Rao distance induced by (4), and thus the transfer learning uncertainty set is given by $\{D(\theta_S, \theta_T) \quad U^2\}$ where

$$D(\theta_S, \theta_T) = (\theta_S \quad \theta_T)^\top (\mathbf{W}^\top \mathbf{W})(\theta_S \quad \theta_T). \tag{6}$$

While our approach is based on knowledge of U, we note that requirements of knowing certain population quantities to rigorously prove optimality are also prevalent in recent studies such as Bastani (2021), Cai and Wei (2021), Li et al. (2020), Tian and Feng (2021).

We further remark that the form of $D(\theta_S, \theta_T)$ in (6) is chosen for the convenience of our analysis, though more general forms are available to us. In fact, for our analysis to go through, it suffices to choose

$$D(\theta_S, \theta_T) = (\theta_S \quad \theta_T)^\top \mathbf{O}(\theta_S \quad \theta_T),$$

for $\mathbf{O}$ positive-definite, and where $(\mathbf{W}^\top \mathbf{W})^{-1}\mathbf{O}$ commutes with $(\mathbf{W}^\top \mathbf{W})^{-1}\mathbf{X}^\top \mathbf{X}$.

These ingredients lead to the minimax risk formulation

$$R = \inf_{\widehat{\theta}} \sup_{D(P_S, P_T) \le U^2} \mathbb{E}_{P_S, P_T}[\ell(\widehat{\theta}, \theta_T)]. \tag{7}$$

Our minimax risk takes the infimum over all estimators $\widehat{\theta}$ given the data and takes the supremum over all pairs

(θ_S, θ_T) whose distance (6) is bounded by U^2. Fixing some θ_S implies that samples from the source domain are not useful, whereas we establish minimax bounds with the finiteness of both source and target samples playing a role, consistent with the existing literature.

In order to determine R, we obtain an upper bound B and a lower bound L on R. Note that the maximum risk of any estimator provides an upper bound. Hence, we first construct a novel estimator that utilizes the geometric structure of the source and target models, and then derive a matching lower bound (up to a constant factor) using Le Cam's method (Tsybakov, 2008, Chapter 2). We make the following assumption throughout the rest of Section 2.1.

Assumption 2.1. *The Gram matrix $\mathbf{W}^\top\mathbf{W}$ corresponding to the target model is positive-definite.*

2.1.1 Derivation of the Upper Bound

The key to obtain a tight upper bound B is to note that the geometric structural difference between the Fisher-Riemannian manifolds induced by the models (3) and (4) is fully described by the generalized eigenvalue problem of the pencil $(\mathbf{X}^\top\mathbf{X}, \mathbf{W}^\top\mathbf{W})$; refer to Golub and Van Loan (2013). Specifically, we have

$$\mathbf{X}^\top\mathbf{X}e_i = \lambda_i \mathbf{W}^\top\mathbf{W}e_i,$$

with eigenvalues λ_i arranged in descending order and eigenvectors $E = (e_1, \ldots, e_d)$ normalized so that

$$E^\top(\mathbf{W}^\top\mathbf{W})E = I, \quad E^\top(\mathbf{X}^\top\mathbf{X})E = \mathrm{diag}(\lambda_1, \ldots, \lambda_d).$$

These eigenvectors represent the directions of the spread of designs with the corresponding eigenvalues representing the relative magnitude of the spread in these directions. The generalized eigenvalue problem has been previously used to define suitable loss functions for positive definite matrices, such as the Förstner metric (Förstner and Moonen, 2003) and Stein's loss (James and Stein, 1992).

Next, let us write θ_S and θ_T in the eigenbasis E as

$$\theta_S = E\beta_S, \qquad \theta_T = E\beta_T, \tag{8}$$

respectively, and thus the problem is given by

$$\mathbf{Y} = (\mathbf{X}E)\beta_S + \epsilon, \qquad \epsilon_i \quad N(0, \sigma_S^2), \tag{9}$$

$$\mathbf{V} = (\mathbf{W}E)\beta_T + \eta, \qquad \eta_i \quad N(0, \sigma_T^2), \tag{10}$$

$$\ell(\widehat{\theta}, \theta_T) = \tilde{\ell}(\widehat{\beta}, \beta_T) = k\widehat{\beta} \quad \beta_T k_2^2,$$
$$D(\theta_S, \theta_T) = \tilde{D}(\beta_S, \beta_T) = k\beta_S \quad \beta_T k_2^2,$$

where $\widehat{\theta} = E\widehat{\beta}$. Hence, it is more convenient to work with the following reparameterization of the original formulation in (7):

$$\inf_{\text{b}} \sup_{D(\ S;\ T)\ \ U^2} \mathsf{E}_{P_S;P_T}[\tilde{\ell}(\widehat{\beta}, \beta_T)]. \tag{11}$$

Denote by $\widehat{\beta}_S$ and $\widehat{\beta}_T$ the ordinary least squares estimate for problems (9) and (10), respectively; namely,

$$\widehat{\beta}_S = (E^\top\mathbf{X}^\top\mathbf{X}E)\ \ {}^1E^\top\mathbf{X}^\top\mathbf{Y},$$
$$\widehat{\beta}_T = (E^\top\mathbf{W}^\top\mathbf{W}E)\ \ {}^1E^\top\mathbf{W}^\top\mathbf{V}.$$

Our proposed model averaging estimator then interpolates $\widehat{\beta}_S$ and $\widehat{\beta}_T$ coordinate-wise, i.e., we have

$$\begin{aligned}\widehat{\theta}_{t_1;\cdots;t_d} &= E\widehat{\beta}_{t_1;\cdots;t_d},\\ \widehat{\beta}_{t_1;\cdots;t_d} &= \mathrm{diag}(t_1, \ldots, t_d)\widehat{\beta}_S \\ &\quad + \mathrm{diag}(1 \quad t_1, \ldots, 1 \quad t_d)\widehat{\beta}_T, \qquad t_i\ 2\ [0,1].\end{aligned} \tag{12}$$

We have the following main result for an upper bound B on problem (7), or equivalently on problem (11).

Theorem 2.1. *Under Assumption 2.1, an upper bound B is given by*

$$\inf_{t_1;\cdots;t_d} \sup_{D(\ S;\ T)\ \ U^2} \mathsf{E}_{P_S;P_T}[\ell(\widehat{\theta}_{t_1;\cdots;t_d}, \theta_T)] \tag{13}$$

$$= \sum_{i=1}^{d} \frac{1}{\frac{1}{\frac{2}{T}} + \frac{1}{\frac{?}{i}U^2 + \frac{2}{\frac{S}{i}}}}, \tag{14}$$

where

$$\alpha_i^? = \begin{cases} \sum_{j=i}^{K^?}\kappa_j + \frac{1}{K^?+1}(1 \quad \sum_{j=1}^{K^?} j\kappa_j) & \text{if } i \quad K^?+1,\\ 0 & \text{if } i > K^?+1,\end{cases}$$

$$K^? = \mathsf{P}_{\ \ i=1}^{\ K}\ i\ \ \max_{\ i\ \ 1;0\ \ K\ \ d\ \ 1} K,$$

$$\kappa_i = \frac{\sigma_S^2}{U^2}\left(\frac{1}{\lambda_{i+1}} \quad \frac{1}{\lambda_i}\right), \qquad i = 1, \ldots, d \quad 1.$$

Moreover, the optimal estimator $\widehat{\theta}_{t_1^?;\cdots;t_d^?}$ satisfies

$$t_i^? = \frac{\sigma_T^2}{\sigma_T^2 + \alpha_i^? U^2 + \frac{2}{\frac{S}{i}}}. \tag{15}$$

Sketch of Proof: Problem (13) can be reformulated into the following finite-dimensional optimization:

$$\inf_{t_i\ 2[0;1]}\ \ {}_i\ \ 0;\ \ \mathsf{P}^{\sup}_{\ \ d\ \ i=1}\ \ {}_{i=1}\sum_{i=1}^{d} t_i^2\left(\frac{\sigma_S^2}{\lambda_i} + \alpha_i U^2\right) + (1 \quad t_i)^2\sigma_T^2.$$

Since the objective function is convex in t_i and concave in α_i, by Sion's minimax theorem (Sion, 1958) we can swap the infimum and supremum to obtain (and, moreover, a pair of Nash equilibrium exists for)

$$\sup_{\ \ i\ \ 0;\ \ \mathsf{P}^{d}_{\ \ i=1}\ \ {}_{i=1}}\ \inf_{t_i\ 2[0;1]} \sum_{i=1}^{d} t_i^2\left(\frac{\sigma_S^2}{\lambda_i} + \alpha_i U^2\right) + (1 \quad t_i)^2\sigma_T^2.$$

The inner problem is quadratic and easy to solve, and thus we arrive at

$$\sup_{\ i\ \ 0;\ \overset{d}{\ i=1}\ \ i=1} \sum_{i=1}^{d} \frac{1}{\frac{1}{\frac{2}{T}} + \frac{1}{{}_i U^2 + \frac{2}{S_i}}},$$

which again has an explicit solution. □

2.1.2 Derivation of the Lower Bound

Utilizing the coordinate transformation (8) and results from Theorem 2.1, we next have the following main result for a lower bound L on problem (7), or equivalently on problem (11).

Theorem 2.2. *Under Assumption 2.1, a lower bound L is given by*

$$\inf_{\mathsf{b}} \sup_{D(\ s;\ T)\ \ U^2} \mathsf{E}_{P_S;P_T}[\ell(\widehat{\theta}, \theta_T)] \qquad \frac{\exp\left(\ \frac{1}{2}\right)}{16} \sum_{i=1}^{d} \frac{1}{\frac{1}{\frac{2}{T}} + \frac{1}{{}_i^? U^2 + \frac{2}{S_i}}}. \quad (16)$$

Sketch of Proof: It is more convenient to work with the reparametrization (11), which is lower bounded by

$$\inf_{\mathsf{b}\ ((\ s)_i\ \ (\ T)_i)^2} \sup_{\ \ {}_i^? U^2 8 i} \mathsf{E}_{P_S;P_T}[\tilde{\ell}(\widehat{\beta}, \beta_T)] \qquad \sum_{i=1}^{d} \inf_{\mathsf{b}_i\ ((\ s)_i\ \ (\ T)_i)^2} \sup_{\ \ {}_i^? U^2} \mathsf{E}_{P_S;P_T}[(\widehat{\beta}_i \quad (\beta_T)_i)^2]. \quad (17)$$

We note that $(E^{\top}\mathbf{X}^{\top}\mathbf{Y})_i$ and $(E^{\top}\mathbf{W}^{\top}\mathbf{V})_i$ are sufficient statistics for $(\beta_S)_i$ and $(\beta_T)_i$, respectively, and

$$(E^{\top}\mathbf{X}^{\top}\mathbf{Y})_i = \lambda_i(\beta_S)_i + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \quad N(0, \lambda_i\sigma_S^2),$$
$$(E^{\top}\mathbf{W}^{\top}\mathbf{V})_i = (\beta_T)_i + \tilde{\eta}_i, \quad \tilde{\eta}_i \quad N(0, \sigma_T^2),$$

where the noise $\tilde{\epsilon}_i$ and $\tilde{\eta}_i$ are independent. For each of the d one-dimensional minimax problems in (17), we reduce the problem to the testing of two carefully constructed hypotheses via Le Cam's method. □

From Theorems 2.1 and 2.2, we observe that the upper bound B and the lower bound L differ by only a constant factor (i.e., $\exp\left(\ \frac{1}{2}\right)/16$). We therefore have established that the minimax risk R obeys the rate

$$R \quad \sum_{i=1}^{d} \frac{1}{\frac{1}{\frac{2}{T}} + \frac{1}{{}_i^? U^2 + \frac{2}{S_i}}}.$$

Under mild conditions, we obtain that the Gram matrix $\mathbf{W}^{\top}\mathbf{W}$ grows on the order $O_p(n_T)$, and then the minimax risk for the usual l_2 loss has the rate

$$\inf_{\mathsf{b}} \sup_{D(\ s;\ T)\ \ U^2} \mathsf{E}_{P_S;P_T}[k\widehat{\theta} \quad \theta_T k_2^2] \quad \sum_{i=1}^{d} \frac{\frac{1}{n_T}}{\frac{1}{\frac{2}{T}} + \frac{1}{{}_i^? U^2 + \frac{2}{S_i}}}$$

with probability one on the realization of designs.

Remark 2.1. *Using the channel capacity of a non-Gaussian additive noise channel (Ihara, 1978), we can improve the uniform constant* $\exp\left(\ \frac{1}{2}\right)/16$ *to* $\max\{\exp\left(\ \frac{1}{2}\right)/16, \left((\sigma_S^2/\lambda_i)/({}_i^? U^2 + \sigma_S^2/\lambda_i)\right)^2\}$ *for the i-th summand in* (16)*. Note that the second term is* 1 *if* $\alpha_i^? = 0$*, and it is arbitrarily close to* 1 *if U is su ciently small. Details are given in the supplement.*

Remark 2.2. *The analysis in Kalan et al. (2020) (in random design) involves the spectral gap of the generalized eigenvalue problem, while our analysis (in xed design) takes care of the entire spectrum of the generalized eigenvalues. Details are given in the supplement.*

2.2 Comparison with Basic Approaches

For an analytical comparison with our theoretical results above, we now consider a corresponding mathematical analysis of three basic approaches for transfer learning often deployed in practice: use only the source dataset; use only the target dataset; and pool both the source and target datasets to train a single model as discussed in Daumé (2007). We then compare and discuss the results for each of these basic transfer learning approaches with those above in Theorems 2.1 and 2.2.

Our theoretical results for the three basic approaches are summarized in the following proposition.

Proposition 2.1. *1. For the LRM based solely on the source dataset, the estimator* $\widehat{\theta}_S = E\widehat{\beta}_S$ *satis es*

$$\sup_{D(\ s;\ T)\ \ U^2} \mathsf{E}_{P_S;P_T}[\ell(\widehat{\theta}_S, \theta_T)] = U^2 + \sigma_S^2 \sum_{i=1}^{d} \lambda_i\ {}^1.$$

2. For the LRM based solely on the target dataset, the estimator $\widehat{\theta}_T = E\widehat{\beta}_T$ *satis es*

$$\sup_{D(\ s;\ T)\ \ U^2} \mathsf{E}_{P_S;P_T}[\ell(\widehat{\theta}_T, \theta_T)] = d\sigma_T^2.$$

3. Finally, for the LRM based on pooling the source and target datasets, the estimator

$$\widehat{\theta}_P = \left((\mathbf{X}^{\top} \ \mathbf{W}^{\top})\begin{pmatrix}\mathbf{X}\\ \mathbf{W}\end{pmatrix}\right)^{\ 1} (\mathbf{X}^{\top} \ \mathbf{W}^{\top})\begin{pmatrix}\mathbf{Y}\\ \mathbf{V}\end{pmatrix} \quad (18)$$

satis es

$$\sup_{D(\ s;\ T)\ \ U^2} \mathsf{E}_{P_S;P_T}[\ell(\widehat{\theta}_P, \theta_T)] = U^2 \max_{1\ \ i\ \ d}\left\{\left(\frac{\lambda_i}{1+\lambda_i}\right)^2\right\} + \sigma_T^2 \sum_{i=1}^{d}\left(\frac{1}{1+\lambda_i}\right)^2 + \sigma_S^2 \sum_{i=1}^{d} \frac{\lambda_i}{(1+\lambda_i)^2}.$$

We observe that the worst-case risk (14) attained by our proposed estimator is smaller than that of the basic approaches using only the source or target dataset. Moreover, the ordinary least squares estimate $\widehat{\theta}_T$ is minimax optimal when using only the target dataset (Hodges and Lehmann, 1950, Theorem 6.5). Thus we show, surprisingly, that negative transfer cannot occur if we have access to the bound U on the distance between the source and target parameters. In particular, negative transfer does not occur if there exists an estimator such that the worst-case risk of the estimator is smaller than the minimax risk of only using the samples from the target domain, noting that our model interpolation estimator satisfies this condition. We also show in the supplement that the worst-case risk (14) is smaller than that of the pooling method.

3 GENERALIZED LINEAR MODEL

Our mathematical framework and theoretical results above have been limited to the case of transfer learning with respect to LRMs. We next turn to consider our corresponding mathematical framework and theoretical results for the case of transfer learning within the context of a class of GLMs. We also extend our framework to allow multiple source domains. The class of GLMs of interest is presented first, followed by our mathematical analysis that leads to theoretical results for GLMs analogous to Theorems 2.2 and 2.1.

Suppose we have access to M different source distributions. For each source $m \in [M]$, assume the n_{S_m} samples $y_1^{(m)}, \ldots, y_{n_{S_m}}^{(m)}$ come from the GLM density (with respect to some dominating measure μ_{S_m})

$$f^{(m)}(y_i^{(m)}; \theta_{S_m}) = t^{(m)}(y_i^{(m)}) \exp\left(\frac{y_i^{(m)} \langle x_i^{(m)}, \theta_{S_m} \rangle - \Psi^{(m)}(\langle x_i^{(m)}, \theta_{S_m} \rangle)}{a^{(m)}(\sigma_{S_m})}\right),$$

where $(x_i^{(m)})^\top$ is the i-th row of the design matrix $\mathbf{X}^{(m)}$, $t^{(m)}(\cdot)$ is a nonnegative-valued function defined on the response space, $a^{(m)}(\cdot)$ is a positive function of σ_{S_m}, and $\Psi^{(m)}(\cdot)$ is the log-partition function. In a similar manner for the target, assume the n_T samples $v_1, \ldots, v_{n_T}$ come from the GLM density (with respect to a possibly different dominating measure μ_T)

$$g(v_i; \theta_T) = l(v_i) \exp\left(\frac{v_i \langle w_i, \theta_T \rangle - \Gamma(\langle w_i, \theta_T \rangle)}{b(\sigma_T)}\right),$$

where $w_i^\top$ is the i-th row of the design matrix $\mathbf{W}$, $l(\cdot)$ is analogous to $t^{(m)}(\cdot)$, $b(\cdot)$ is analogous to $a^{(m)}(\cdot)$, and $\Gamma(\cdot)$ is analogous to $\Psi^{(m)}(\cdot)$.

Note that the above GLM, under which samples are generated according to an exponential family with natural parameter equal to a linear transformation of the underlying parameter θ, was used in Lee and Courtade (2020) though not in the context of transfer learning. We further assume that

$$\sup_z (\Psi^{(m)})''(z) \le C_{S_m} \ \forall m, \qquad \sup_z \Gamma''(z) \le C_T.$$

Following in a manner similar to Section 2.1, we next can respectively define the loss function $\ell(\widehat{\theta}, \theta_T)$ and the uncertainty set $\{D(\theta_{S_m}, \theta_T) \le U_m^2\}$ as

$$\ell(\widehat{\theta}, \theta_T) = (\widehat{\theta} - \theta_T)^\top (\mathbf{W}^\top \mathbf{W})(\widehat{\theta} - \theta_T),$$
$$D(\theta_{S_m}, \theta_T) = (\theta_{S_m} - \theta_T)^\top (\mathbf{W}^\top \mathbf{W})(\theta_{S_m} - \theta_T).$$

This leads to the problem formulation

$$\inf_{\hat{\theta}} \sup_{D(\theta_{S_m}, \theta_T) \le U_m^2, \forall m} \mathsf{E}_{P_T, P_{S_m}, m \in [M]}[\ell(\widehat{\theta}, \theta_T)]. \quad (19)$$

To simplify the notation, when no confusion arises, we shall henceforth abbreviate the expectation $\mathsf{E}_{P_T, P_{S_m}, m \in [M]}[\ell(\widehat{\theta}, \theta_T)]$ by $\mathsf{E}[\ell(\widehat{\theta}, \theta_T)]$.

Towards solving this problem formulation, consider the generalized eigenvalue problem of the matrix pencil $((\mathbf{X}^{(m)})^\top \mathbf{X}^{(m)}, \mathbf{W}^\top \mathbf{W})$; refer to Golub and Van Loan (2013). More specifically, we have

$$(\mathbf{X}^{(m)})^\top \mathbf{X}^{(m)} e_i^{(m)} = \lambda_i^{(m)} \mathbf{W}^\top \mathbf{W} e_i^{(m)},$$

where the eigenvalues are arranged such that $\lambda_1^{(m)} \ge \cdots \ge \lambda_d^{(m)}$. Observe that, for any $\theta \in \mathsf{R}^d$,

$$\theta^\top (\mathbf{X}^{(m)})^\top \mathbf{X}^{(m)} \theta \le \lambda_1^{(m)} \theta^\top \mathbf{W}^\top \mathbf{W} \theta.$$

We then have the following main result for a lower bound on the solution of (19) using Le Cam's and Fano's methods (Tsybakov, 2008, Chapter 2).

Theorem 3.1. *A lower bound of the minimax risk corresponding to the GLMs is given by*

$$\inf_{\hat{\theta}} \sup_{D(\theta_{S_m}, \theta_T) \le U_m^2, \forall m} \mathsf{E}[\ell(\widehat{\theta}, \theta_T)] \ge \frac{e^{-1}}{800} \frac{d}{\sum_{m=1}^M \frac{1}{\frac{U_m^2}{d} + \frac{a^{(m)}(\sigma_{S_m})}{C_{S_m} \lambda_1^{(m)}}} + \frac{C_T}{b(\sigma_T)}}.$$

Details are given in the supplement.

Remark 3.1. *For the non-transfer learning setting considered in Lee and Courtade (2020), our proof method gives rise to a lower bound of $d \cdot b(\sigma_T)/C_T$ which is sharper than their lower bound of*

$$\max\left\{\frac{k \Lambda_{\mathbf{W}} k_1^2}{k \Lambda_{\mathbf{W}} k_2^2}, \lambda_{min}(\mathbf{W}^\top \mathbf{W}) k \Lambda_{\mathbf{W}}^{-1} k_1\right\} \cdot b(\sigma_T)/C_T,$$

where $\Lambda_{\mathbf{W}}$ is the vector of eigenvalues of the positive-definite matrix $\mathbf{W}^\top \mathbf{W}$, and $\Lambda_{\mathbf{W}}^{-1}$ denotes its coordinate-wise inverse.

Specializing Theorem 3.1 to the LRM case, we derive a multiple sources analog to Theorem 2.2.

Corollary 3.1. *When the GLMs considered are Gaussian LRMs, then a lower bound of the minimax risk is*

$$\inf_{\hat{\theta}} \sup_{D(\Sigma_m;\Sigma_T)\ U_m^2;\forall m} \mathsf{E}[\ell(\hat{\theta},\theta_T)] \geq \frac{e^{-1}}{800} \frac{d}{\sum_{m=1}^{M} \frac{1}{\frac{U_m^2}{d}+\frac{\lambda_1^{(m)}}{\lambda_d^{(m)}}\frac{\Sigma_m^2}{n_m}} + \frac{1}{n_T}}.$$

In comparison to Theorem 2.2, the rate in Corollary 3.1 involves the spectral gap $\lambda_1^{(m)}/\lambda_d^{(m)}$.

Turning to consider an upper bound within the context of GLMs, we assume that $\inf_z(\Psi^{(m)})''(z) \geq L_{\Sigma_m}$, $\forall m$, and $\inf_z \Gamma''(z) \geq L_T$. Then, using the sub-Gaussian concentration bound for GLM noise and results of Bastani (2021), we obtain the following upper bound of the minimax risk corresponding to the GLMs (ignoring the leading constant):

$$\frac{d}{\sum_{m=1}^{M} \frac{1}{\frac{U_m^2}{d}+\frac{2C_{\Sigma_m}a^{(m)}(\Sigma_m)}{L_{\Sigma_m}^2\lambda_d^{(m)}}} + \frac{1}{\frac{2C_T b(\Sigma_T)}{L_T^2}}}. \tag{20}$$

The details are provided in the supplement. Comparing (20) with Theorem 3.1, we observe that our upper and lower bounds match up to the ratios $C_{\Sigma_m}/L_{\Sigma_m}$ and C_T/L_T and the spectral gap $\lambda_1^{(m)}/\lambda_d^{(m)}$. We plan to consider the tight analysis of upper and lower bounds in the GLM setting as part of future research.

4 SIMULATION RESULTS

Our focus in this paper is on a mathematical framework of transfer learning and corresponding theoretical results related to geometric structures, minimax bounds, and minimax optimality. To provide further insights and understanding with respect to our framework and results, we now present a collection of simulation results that investigate the quantitative performance of our model interpolation estimator under various conditions, environments, and parameter settings. These simulation results showcase the ability of our proposed estimator to outperform the basic transfer learning approach discussed in Daumé (2007) and the recent state-of-the-art transfer learning methods of Bastani (2021) and Li et al. (2020).

We consider the LRM in the transfer learning setting of a single source domain and a single target domain. The optimal interpolation scheme (15) requires that we specify the parameters U, σ_S and σ_T, which are typically unknown a priori. In Section 4.1, we first qualitatively explore the behavior of our proposed method with respect to the setting of U relative to its true value, while assuming perfect knowledge of the remaining parameters. Then, in Section 4.2, we treat all three parameters as unknown and develop heuristic estimation procedures with which we compare this full-fledged version of our proposed method against other competing methods in the research literature.[1]

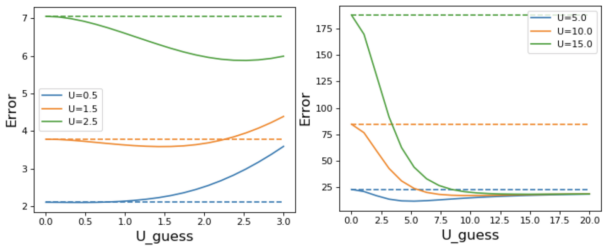

Figure 1: Simulation Results with Different Ground Truth U. Solid Lines Represent the Proposed Method. Dashed Lines Represent the Basic Pooling Method.

4.1 Misspecification of U

For our investigation of the impact of misspecifications of U, the baseline parameters are set to be $d = 20, n_S = 1000, n_T = 100, \sigma_S^2 = \sigma_T^2 = 1$. We randomly generate $\mathbf{X}$ where each row independently follows the standard multivariate Gaussian, and also independently generate $\mathbf{W}$ in a similar fashion. We consider β_T, the target parameter after the coordinate transformation (8), as the vector of all ones. Note that specifying θ_T and β_T is equivalent given the design matrices $\mathbf{X}$ and $\mathbf{W}$. Experiments are then performed with ground-truth values of $U \in \{0.5, 1.5, 2.5\}$ where, for each value of U, we generate β_S to be the Nash equilibrium in problem (13). We then repeat 1000 independent simulation runs where, in each run, the design matrices $\mathbf{X}, \mathbf{W}$ are kept unchanged and fresh copies of the response vectors $\mathbf{Y}, \mathbf{V}$ are resampled following the LRM (9) and (10).

In calculating the optimal interpolation scheme, we assume σ_S^2, σ_T^2 to be known and experiment with U_{guess} values that are equispaced within the interval $[0, 3]$ as (mis)specifications of U. Then, we compute the average of the estimation error for θ_T in (5) over the 1000 runs and plot the corresponding trends with respect to U_{guess}. The results are summarized in the left plot of Figure 1, where the horizontal lines are the averaged error of the basic pooling method (18) included for comparison. We observe that the performance gap between the proposed method (in solid lines) and the basic pooling method (in dashed lines) increases with

[1] All problems are modeled in Python and run on an Intel i5 CPU (1.4GHz) computer.

Table 1: Simulation Results Comparing the Proposed Method to Other Competing Methods in Moderate-Dimensions (Left-Half) and in High-Dimensions (Right-Half). "Basic" Represents the Lowest of the Errors Attained by the Three Basic Methods in Section 2.2. Numbers in Parentheses Are Standard Deviations.

U	Basic	Proposed	Two-Step	Trans-Lasso	Basic	Proposed	Two-Step	Trans-Lasso
0.5	1.9(0.6)	2.3(1.3)	3.0(2.5)	**0.9**(1.8)	**9.1**(1.7)	9.3(2.1)	11.8(3.9)	9.6(3.0)
1.5	3.9(1.1)	4.3(1.5)	5.7(3.2)	**2.8**(2.9)	11.4(2.0)	**11.4**(2.3)	14.5(4.6)	12.2(2.9)
2.5	6.9(1.5)	6.5(1.7)	7.9(2.6)	**5.3**(2.3)	15.2(1.9)	**14.6**(2.5)	18.1(3.8)	16.2(2.3)
3.5	11.7(1.9)	**9.2**(2.5)	11.9(3.5)	9.5(2.9)	20.5(2.5)	**18.6**(2.7)	23.8(3.4)	21.7(2.8)
4.5	18.6(2.7)	**12.4**(4.1)	15.4(4.9)	13.9(4.0)	27.6(2.3)	**23.6**(3.9)	31.2(3.1)	29.2(2.9)
5.5	20.3(6.1)	**14.8**(5.3)	17.9(5.9)	18.0(5.3)	36.9(3.1)	**29.4**(4.6)	41.3(3.9)	38.3(3.4)
6.5	20.9(6.6)	**15.8**(5.5)	18.4(6.6)	19.2(7.3)	46.1(3.2)	**33.7**(5.0)	42.1(4.8)	47.9(3.7)
7.5	19.3(5.8)	**16.2**(4.8)	18.8(7.8)	19.6(8.2)	54.6(12.5)	**40.7**(6.6)	65.6(5.9)	60.8(5.2)
8.5	20.8(6.6)	**18.8**(6.6)	20.8(8.5)	21.4(8.8)	53.8(10.5)	**43.6**(7.2)	80.9(6.3)	73.5(4.8)
9.5	20.4(6.6)	**19.1**(6.5)	19.6(7.5)	20.1(8.0)	51.7(11.3)	**48.3**(9.6)	97.7(9.1)	88.6(6.2)

the ground-truth value of U. While this gap is highest when the proposed method uses the correct U value, it is robust to misspecification in U.

We now vary the parameter values and observe that similar phenomena exist across the different settings:

The ground-truth value of U is changed to $\mathsf{f}5, 10, 15\mathsf{g}$, and we experiment with U_{guess} values equispaced within $[0, 20]$. The results are summarized in the right plot of Figure 1. Our proposed method works much better than the basic pooling method across the wider range of U values around the ground-truth.

The rows of $\mathbf{X}$ are generated independently by a zero-mean Gaussian with a Toeplitz covariance matrix (Li et al., 2020, Section 5.2), or both $\mathbf{X}$ and $\mathbf{W}$ are generated in this way; refer to the supplement and Figure 2. The introduction of correlation does not impact the performance of either method when compared to the uncorrelated case in the left plot of Figure 1.

The noise variances are changed to $\sigma_{\mathsf{S}}^2 = 1, \sigma_{\mathsf{T}}^2 = 5$, or $\sigma_{\mathsf{S}}^2 = 5, \sigma_{\mathsf{T}}^2 = 1$; refer to the supplement and Figure 3. The proposed method handles large variances in the source data much better than the basic pooling method, while high variance in (smaller sized) target data has no significant impact on either method.

The dimension d is changed to 5, or $d = 100$ but with β_{T} three-sparse (specifically, the first three elements of β_{T} are one and the rest are zero); refer to the supplement and Figure 4. Lower dimensionality seems to improve the performance gap between the two methods as compared to the left plot of Figure 1, while this gap shortens in high-dimensions with extreme sparsity.

4.2 Comparisons with Competing Methods

We have seen from Figures 1 – 4 that our method admits a broad tolerance to misspecifications of the value of U relative to its true value, especially if the true value is moderately large. We next develop heuristic procedures for estimating U, alongside with σ_{S}^2 and σ_{T}^2, from the datasets as follows.

We use the usual least squares MLE estimate

$$\widehat{\sigma_{\mathsf{S}}^2} = \frac{1}{n_{\mathsf{S}}} \sum_{i=1}^{n_{\mathsf{S}}} \left(y_i \quad x_i^{\mathsf{T}} \widehat{\theta}_{\mathsf{S}} \right)^2,$$

with $\widehat{\theta}_{\mathsf{S}}$ the ordinary least squares estimate of θ_{S} in (3).

For moderate dimension and θ_{T} not sparse, we use a similar least squares estimate $\widehat{\sigma_{\mathsf{T}}^2}$. However, in high-dimensional settings, it has been observed that a more accurate estimator is given by (Reid et al., 2016)

$$\widehat{\sigma_{\mathsf{T}}^2} = \frac{1}{n_{\mathsf{T}} \quad \hat{s}_{\wedge}} \sum_{i=1}^{n_{\mathsf{T}}} \left(v_i \quad w_i^{\mathsf{T}} \widehat{\theta}_{\mathsf{T};\wedge} \right)^2,$$

where $\widehat{\theta}_{\mathsf{T};\wedge}$ is the Lasso estimator with cross-validated penalization parameter $\hat{\gamma}$, and $\hat{s}_{\wedge}$ is the number of non-zero elements in $\widehat{\theta}_{\mathsf{T};\wedge}$.

We use a 5-fold cross-validation (CV) procedure to determine an estimate $\widehat{U}$, where the CV objective is the mean-squared test error on the hold-out set.

The experimental results in Section 4.1 demonstrate forms of robustness with respect to the misspecification of U in our approach. Beyond the above 5-fold CV approach to estimate U from the datasets, which delivers promising results below in comparison with state-of-the-art methods, we can use subsampling methods as an alternative to estimate U whenever the source and target samples are not too scarce. We plan to address this issue in more detail as part of future work.

Now we compare the results from our full-fledged method with those from the basic methods discussed

Table 2: Results Comparing the Proposed Method to Other Competing Methods on Uber&Lyft Data.

n_S	n_T	$\widehat{U}$	Basic	Proposed	Two-Step	Trans-Lasso
1000	100	19.79	34.54(15.10) (pooling)	**28.50**(16.53)	1.16(2.03) 10^5	93.80(57.22)
10000	1000	19.20	38.52(14.72) (target)	**36.30**(14.32)	1.62(2.35) 10^5	778.47(538.76)
1000	1000	15.05	37.17(10.66) (target)	**36.04**(11.08)	8.84(13.23) 10^5	353.49(347.70)
100	10000	47.97	30.72(9.35) (target)	**30.50**(9.35)	1.08(15.94) 10^6	47.59(44.61)
10000	100	9.29	29.86(6.37) (pooling)	**26.47**(8.10)	1.35(2.46) 10^4	101.31(67.10)

in Section 2.2 and two recent state-of-the-art transfer learning methods in the literature, namely the two-step joint estimator proposed by Bastani (2021) and its extension to Trans-Lasso by Li et al. (2020). For the latter case, since the setting is a single source domain from which learning is transferred to a target domain, we only include for comparison the Oracle Trans-Lasso algorithm in Li et al. (2020) (i.e., their Algorithm 1).

4.2.1 Comparisons in Moderate-Dimensions

The parameters for the case of moderate dimensions are set to be $d = 20, n_S = 1000, n_T = 100, \sigma_S^2 = \sigma_T^2 = 1$. We randomly generate $\mathbf{X}$ where each row independently follows the standard multivariate Gaussian, and independently generate $\mathbf{W}$ in a similar fashion. We consider θ_T to be the vector of all ones. Experiments are performed with ground-truth values of U in $\{0.5, 1.5, \ldots, 9.5\}$ where, for each value of U, we generate θ_S to be the Nash equilibrium in problem (13). We then repeat 1000 independent simulation runs where, in each run, the design matrices $\mathbf{X}, \mathbf{W}$ are kept unchanged and fresh copies of the response vectors $\mathbf{Y}, \mathbf{V}$ are resampled following the LRM (3) and (4). For the methods under consideration, we report the average estimation error of θ_T in (5), and its standard deviation, over the 1000 runs. The results are summarized in the left-half of Table 1. For small U, the Trans-Lasso method produces the best results. We note that these minimax optimality results of Li et al. (2020) are established under different assumptions, and that their results do not contradict our minimax optimality results. For moderate to larger U, our proposed method attains better performance on average. We also provide additional experiments with $n_S = 200, n_T = 100$ in the supplement, which exhibit consistent behaviors.

4.2.2 Comparisons in High-Dimensions

With all other parameters remaining the same, we now consider a more challenging high-dimensional setting where $d = 100$. Moreover, we set θ_T to be a sparse vector where the first 20 elements are one, and the remaining 80 elements are zero. To deal with this high-dimensional setup, we make a simple heuristic modification to the proposed interpolation scheme (12) by replacing the least squares estimator $\widehat{\beta}_T$ with $\widehat{\beta}_{T;\wedge}$, i.e., the Lasso estimator $\widehat{\theta}_{T;\wedge}$ after the coordinate transformation (8). We repeat 1000 independent simulation runs, and report the average estimation error of θ_T in (5) and its standard deviation for the methods under consideration. The results are summarized in the right-half of Table 1. Our proposed method outperforms the other two methods for all U values considered, somewhat surprisingly even for small U since the competing methods were designed to exploit sparsity. The supplement provides additional experiments with $n_S = 200, n_T = 100$ that exhibit consistent behaviors.

4.2.3 Comparisons on Real-World Dataset

Lastly, we compare the results from the different methods using the Uber&Lyft dataset[2] of Uber and Lyft cab rides collected in Boston, MA. The learning problem comprises prediction of the price using $d = 32$ numerical features, including hour-of-the-day, distance, weather, and demand factors. We consider UberX as the source model and standard Lyft service as the target. The entire dataset consists of 55094 observations for the source and 51235 observations for the target, from which we compute the ground truth regression parameters; see the supplement. Since we wish to study the benefit of transfer learning, we restrict ourselves to small random subsamples. We repeat 100 independent experiments and summarize the results in Table 2. Our proposed method attains a better performance on average, by a small margin relative to the basic methods and by a large margin relative to the two-step estimator and Trans-Lasso. Notice that for this problem l_q sparsity, $q \in [0, 1]$, (required by the last two methods) does not reasonably capture the contrast between the source and target models, due to the moderate dimensions and the existence of one dominating feature; see Table 4 and Figure 9 in the supplement.

[2] https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma

Acknowledgements

J. Blanchet gratefully acknowledges support from the NSF via grant DMS-EPSRC 2118199 and AFOSR, as well as NSF-DMS 1915967 and AFOSR MURI FA9550-20-1-0397.

Part of this work was done while X. Zhang was at the IBM Thomas J. Watson Research Center.

References

Kahkashan Afrin, Bimal Nepal, and Leslie Monplaisir. A data-driven framework to new product demand prediction: Integrating product differentiation and transfer learning approach. *Expert Systems with Applications*, 108:246–257, 2018.

Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

T. Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100 – 128, 2021.

L. Lorne Campbell. An extended Cencov characterization of the information metric. *Proceedings of the American Mathematical Society*, 98(1):135–141, 1986.

Hal Daumé, III. Frustratingly easy domain adaptation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, 2007.

Wolfgang Förstner and Boudewijn Moonen. *A Metric for Covariance Matrices*, pages 299–309. Springer Berlin Heidelberg, 2003.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4th edition, 2013.

J. L. Hodges and E. L. Lehmann. Some problems in minimax point estimation. *The Annals of Mathematical Statistics*, 21(2):182–197, 1950.

Shunsuke Ihara. On the capacity of channels with additive non-Gaussian noise. *Information and Control*, 37(1):34–39, 1978.

W. James and Charles Stein. *Estimation with Quadratic Loss*, pages 443–460. Springer New York, 1992.

Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, and Mahdi Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

Samory Kpotufe and Guillaume Martinet. Marginal singularity and the benefits of labels in covariate-shift. *Annals of Statistics*, 49(6):3299–3323, 2021.

Kuan-Yun Lee and Thomas Courtade. Minimax bounds for generalized linear models. In *Advances in Neural Information Processing Systems*, volume 33, pages 9372–9382. Curran Associates, Inc., 2020.

Sai Li, T. Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *arXiv e-prints*, art. arXiv:2006.10593, 2020.

Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10), 2020.

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

Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in Lasso regression. *Statistica Sinica*, 26(1):35–67, 2016.

Maurice Sion. On general minimax theorems. *Paci c Journal of Mathematics*, 8(1):171 – 176, 1958.

Bahar Taskesen, Man-Chung Yue, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. Sequential domain adaptation by synthesizing distributionally robust experts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10162–10172. PMLR, 2021.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *arXiv e-prints*, art. arXiv:2105.14328, 2021.

L. Torrey and J. Shavlik. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, page 242–264. IGI Global, 2010.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008.

K. Weiss, T.M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Supplementary Material: A Class of Geometric Structures in Transfer Learning: Minimax Bounds and Optimality

In support of the main body of the paper, this supplement contains additional results, technical details, and complete proofs of our theoretical results. We start by presenting proofs and related results for Theorem 2.1 and Theorem 2.2 in Sections A.1 and A.2, respectively. We then present the proofs of Remark 2.1 and Proposition 2.1 in Sections A.3 and A.4, respectively. Next, we present the proofs of Theorem 3.1, Remark 3.1 and Corollary 3.1 in Sections A.5, A.6 and A.7, respectively. Each of these sections includes statements of the theoretical results from the main body of the paper in an effort to make the supplement self-contained. We also provide auxiliary results on the GLM upper bound in Section A.8 and on the comparison with Kalan et al. (2020) in Section A.9. Finally, in Section B, we present an additional set of simulation results together with additional details and results for the application of a real-world dataset that complement those in the main body of the paper.

A Proofs and Related Results

A.1 Proof of Theorem 2.1

Theorem 2.1. *Under Assumption 2.1, an upper bound B is given by*

$$\inf_{t_1,\dots,t_d} \sup_{D(\ _S;\ _T)\ \ U^2} \mathsf{E}_{P_S;P_T}[\ell(\widehat{\theta}_{t_1,\dots,t_d}, \theta_T)] \tag{13}$$

$$= \sum_{i=1}^{d} \frac{1}{\frac{1}{\frac{2}{T}} + \frac{1}{\frac{?}{i}U^2 + \frac{2}{\frac{S}{i}}}}, \tag{14}$$

where

$$\alpha_i^? = \begin{cases} \sum_{j=i}^{K^?} \kappa_j + \frac{1}{K^?+1}(1 \quad \sum_{j=1}^{K^?} j\kappa_j) & \text{if } i \quad K^?+1, \\ 0 & \text{if } i > K^?+1, \end{cases}$$

and

$$K^? = \ _{\mathsf{P}\ \frac{K}{i=1}\ i\ \ i} \max_{1;0\ \ K\ \ d\ \ 1} K,$$

with

$$\kappa_i = \frac{\sigma_S^2}{U^2}\left(\frac{1}{\lambda_{i+1}} \quad \frac{1}{\lambda_i}\right), \qquad i = 1,\dots,d \quad 1.$$

Moreover, the optimal estimator $\widehat{\theta}_{t_1^?,\dots,t_d^?}$ satis es

$$t_i^? = \frac{\sigma_T^2}{\sigma_T^2 + \alpha_i^? U^2 + \frac{2}{\frac{S}{i}}}. \tag{15}$$

Proof. It is more convenient to work with the reparametrization (11). Note that

$$\begin{aligned} \widehat{\beta}_S &= (E^\top \mathbf{X}^\top \mathbf{X} E)\ \ ^1 E^\top \mathbf{X}^\top \mathbf{Y} \qquad \mathsf{N}(\beta_S, \sigma_S^2 \mathrm{diag}(\lambda_1\ ^1, \dots, \lambda_d\ ^1)), \\ \widehat{\beta}_T &= (E^\top \mathbf{W}^\top \mathbf{W} E)\ \ ^1 E^\top \mathbf{W}^\top \mathbf{V} \qquad \mathsf{N}(\beta_T, \sigma_T^2 I). \end{aligned}$$

We therefore obtain

$$
\begin{aligned}
&\inf_{t_1,\dots,t_d} \sup_{D(\ S;\ T)\ \ U^2} \mathsf{E}_{P_S;P_T}[k\widehat{\beta} \quad \beta_T k_2^2] \\
&\qquad = \inf_{t_1,\dots,t_d} \sup_{k\ S\ \ T k_2^2\ \ U^2} \mathrm{Tr}\left(\mathrm{diag}(t_1^2,\dots,t_d^2)\sigma_S^2 \mathrm{diag}(\lambda_1{}^1,\dots,\lambda_d{}^1)\right) + \sum_{i=1}^d t_i^2((\beta_S)_i \quad (\beta_T)_i)^2 \\
&\qquad\qquad + \mathrm{Tr}\left(\mathrm{diag}((1 \quad t_1)^2,\dots,(1 \quad t_d)^2)\sigma_T^2\right) \\
&\qquad = \inf_{t_1,\dots,t_d} \sup_{{}_i\ 0;\ \sum_{i=1}^d {}_i = 1} \sum_{i=1}^d t_i^2\left(\frac{\sigma_S^2}{\lambda_i} + \alpha_i U^2\right) + (1 \quad t_i)^2\sigma_T^2 \\
&\qquad = \sup_{{}_i\ 0;\ \sum_{i=1}^d {}_i = 1} \inf_{t_1,\dots,t_d} \sum_{i=1}^d t_i^2\left(\frac{\sigma_S^2}{\lambda_i} + \alpha_i U^2\right) + (1 \quad t_i)^2\sigma_T^2 \\
&\qquad = \sup_{{}_i\ 0;\ \sum_{i=1}^d {}_i = 1} \sum_{i=1}^d \frac{1}{\frac{1}{\sigma_T^2} + \frac{1}{{}_i U^2 + \frac{\sigma_S^2}{\lambda_i}}} \\
&\qquad = d\sigma_T^2 \quad \sigma_T^4 \left(\inf_{{}_i\ 0;\ \sum_{i=1}^d {}_i = 1} \sum_{i=1}^d \frac{1}{\alpha_i U^2 + \sigma_T^2 + \frac{\sigma_S^2}{\lambda_i}} \right),
\end{aligned}
$$

where Sion's minimax theorem (Sion, 1958) is employed to swap the supremum and infinum. Further note that

$$
\sigma_T^2 + \frac{\sigma_S^2}{\lambda_1} \qquad \sigma_T^2 + \frac{\sigma_S^2}{\lambda_d}.
$$

Let

$$
\kappa_i = \frac{\sigma_S^2}{U^2}\left(\frac{1}{\lambda_{i+1}} \quad \frac{1}{\lambda_i}\right), \qquad i = 1,\dots,d \quad 1,
$$

and

$$
K^? = \sum_{i=1}^{K} {}_i \max_{1;0 \ \ K \ \ d \ \ 1} K.
$$

It is then easy to see that the solution of

$$
\inf_{{}_i\ 0;\ \sum_{i=1}^d {}_i = 1} \sum_{i=1}^d \frac{1}{\alpha_i U^2 + \sigma_T^2 + \frac{\sigma_S^2}{\lambda_i}}
$$

is given by

$$
\alpha_i^? = \begin{cases} \sum_{j=i}^{K^?} \kappa_j + \frac{1}{K^?+1}\left(1 \quad \sum_{j=1}^{K^?} j\kappa_j\right) & \text{if } i \quad K^?+1, \\ 0 & \text{if } i > K^?+1. \end{cases}
$$

Hence, we have

$$
\inf_{t_1,\dots,t_d} \sup_{D(\ S;\ T)\ \ U^2} \mathsf{E}_{P_S;P_T}[k\widehat{\beta} \quad \beta_T k_2^2] = \sum_{i=1}^d \frac{1}{\frac{1}{\sigma_T^2} + \frac{1}{{}_i^? U^2 + \frac{\sigma_S^2}{\lambda_i}}},
$$

and

$$
t_i^? = \frac{\frac{1}{{}_i^? U^2 + \frac{\sigma_S^2}{\lambda_i}}}{\frac{1}{\sigma_T^2} + \frac{1}{{}_i^? U^2 + \frac{\sigma_S^2}{\lambda_i}}} = \frac{\sigma_T^2}{\sigma_T^2 + \alpha_i^? U^2 + \frac{\sigma_S^2}{\lambda_i}}, \qquad i = 1,\dots,d,
$$

thus completing the proof. □

A.2 Proof of Theorem 2.2

In order to obtain our lower bound results, we shall make use of the following lemma which concerns the admissibility of the hypotheses in applying Le Cam's or Fano's method (Tsybakov, 2008, Chapter 2).

Lemma A.1. *For any $\gamma_i > 0$, $U_i > 0, i = 0, \ldots, M$, and $K \quad \frac{1}{4}$, there exists $\beta_i^{(j)}$ 2 R, $i = 0, \ldots, M, j = 0, 1$, such that the following conditions hold simultaneously:*

1. $\left(\beta_0^{(0)} \quad \beta_0^{(1)}\right)^2 = \frac{1}{K} \frac{1}{0 + \sum_{k=1}^{M} \frac{1}{U_k^2 + \frac{1}{k}}}$;
2. $\left|\beta_0^{(j)} \quad \beta_i^{(j)}\right| \quad U_i, \qquad i = 1, \ldots, M, j = 0, 1$;
3. $\left(\beta_i^{(0)} \quad \beta_i^{(1)}\right)^2 \quad \frac{1}{K} \frac{\frac{1}{U_i^2 + \frac{1}{i}} \frac{1}{i}}{0 + \sum_{k=1}^{M} \frac{1}{U_k^2 + \frac{1}{k}}}, \qquad i = 1, \ldots, M.$

Under these conditions, we have

$$\sum_{i=0}^{M} \gamma_i \left(\beta_i^{(0)} \quad \beta_i^{(1)}\right)^2 \quad \frac{1}{K}.$$

Proof. For the three conditions to hold simultaneously, it suffices to show (using the triangle inequality twice to bound condition 2 in terms of conditions 1 and 3) that, for $i = 1, \ldots, M$,

$$\sqrt{\frac{1}{\gamma_0 + \sum_{k=1}^{M} \frac{1}{U_k^2 + \frac{1}{k}}}} \quad \sqrt{\frac{\frac{1}{U_i^2 + \frac{1}{i}} \frac{1}{i}}{\gamma_0 + \sum_{k=1}^{M} \frac{1}{U_k^2 + \frac{1}{k}}}} \quad 2^{p} \overline{K} U_i.$$

This problem is equivalent to showing that

$$1 \quad \sqrt{\frac{1}{\gamma_i} \frac{1}{U_i^2 + \frac{1}{i}}} \quad 2^{p} \overline{K} U_i \sqrt{\gamma_0 + \sum_{k=1}^{M} \frac{1}{U_k^2 + \frac{1}{k}}}.$$

In fact, we can show a tighter bound

$$1 \quad \sqrt{\frac{1}{\gamma_i} \frac{1}{U_i^2 + \frac{1}{i}}} \quad 2^{p} \overline{K} U_i \sqrt{\gamma_0 + \frac{1}{U_i^2 + \frac{1}{i}}},$$

which can be rewritten as

$$\sqrt{U_i^2 + \frac{1}{\gamma_i}} \quad \sqrt{\frac{1}{\gamma_i}} \quad 2^{p} \overline{K} U_i \sqrt{1 + \gamma_0 (U_i^2 + \frac{1}{\gamma_i})},$$

or further rewritten as

$$\frac{U_i^2}{\sqrt{U_i^2 + \frac{1}{i}} + \sqrt{\frac{1}{i}}} \quad 2^{p} \overline{K} U_i \sqrt{1 + \gamma_0 (U_i^2 + \frac{1}{\gamma_i})}. \tag{21}$$

Since $K \quad \frac{1}{4}$ and

$$\frac{U_i}{\sqrt{U_i^2 + \frac{1}{i}} + \sqrt{\frac{1}{i}}} \quad 1,$$

we conclude that the desired result (21) holds. The last part of the claim in Lemma A.1 follows easily. □

We now are in a position to prove Theorem 2.2 via Le Cam's method (Tsybakov, 2008, Chapter 2), which we restate as follows.

Theorem 2.2. *Under Assumption 2.1, a lower bound L is given by*

$$\inf_{\hat{\theta}} \sup_{D(P_S, P_T) \in U^2} \mathbb{E}_{P_S, P_T}[\ell(\hat{\theta}, \theta_T)] \geq \frac{\exp(-\frac{1}{2})}{16} \sum_{i=1}^{d} \frac{1}{\frac{1}{\sigma_T^2} + \frac{1}{\lambda_i U^2 + \frac{\sigma_S^2}{\lambda_i}}}. \tag{16}$$

Proof. It is more convenient to work with the reparametrization (11). Note that

$$\inf_{\hat{\theta}} \sup_{D(P_S, P_T) \in U^2} \mathbb{E}_{P_S, P_T}[\tilde{\ell}(\hat{\beta}, \beta_T)] \geq \inf_{\hat{\theta}} \sup_{((P_S)_i, (P_T)_i) \in \Lambda U^2 \otimes I} \mathbb{E}_{P_S, P_T}[\tilde{\ell}(\hat{\beta}, \beta_T)]$$
$$\geq \sum_{i=1}^{d} \inf_{\hat{\theta}_i} \sup_{((P_S)_i, (P_T)_i) \in \Lambda U^2} \mathbb{E}_{P_S, P_T}[(\hat{\beta}_i - (\beta_T)_i)^2].$$

Consider the singular value decomposition of $\mathbf{W}E$, for which we obtain

$$\mathbf{W}E = PDQ = P \begin{pmatrix} Q \\ 0 \end{pmatrix},$$

since $E^\top \mathbf{W}^\top \mathbf{W} E = I$ and where P, Q are orthogonal matrices of appropriate dimensions. We therefore have

$$\begin{pmatrix} E^\top \mathbf{W}^\top \\ (0\ I)P^\top \end{pmatrix} \mathbf{W}E = \begin{pmatrix} Q^\top & 0 \\ 0 & I \end{pmatrix} P^\top \mathbf{W}E = \begin{pmatrix} I \\ 0 \end{pmatrix},$$

and thus

$$\tilde{\mathbf{V}} = \begin{pmatrix} E^\top \mathbf{W}^\top \\ (0\ I)P^\top \end{pmatrix} \mathbf{V} = \begin{pmatrix} I \\ 0 \end{pmatrix} \beta_T + \tilde{\eta}, \quad \tilde{\eta} \sim N\left(0, \sigma_T^2 \begin{pmatrix} I & 0 \\ 0 & \Gamma \end{pmatrix}\right), \tag{22}$$

where Γ is some positive-definite matrix that does not concern us in the following calculations. Similarly, considering the singular value decomposition of $\mathbf{X}E$, we obtain

$$\mathbf{X}E = \tilde{P}\tilde{D}\tilde{Q} = \tilde{P} \begin{pmatrix} \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_d^{1/2})\tilde{Q} \\ 0 \end{pmatrix},$$

since $E^\top \mathbf{X}^\top \mathbf{X} E = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ and where $\tilde{P}, \tilde{Q}$ are orthogonal matrices of appropriate dimensions. We therefore have

$$\tilde{\mathbf{Y}} = \begin{pmatrix} E^\top \mathbf{X}^\top \\ (0\ I)\tilde{P}^\top \end{pmatrix} \mathbf{Y} = \begin{pmatrix} \mathrm{diag}(\lambda_1, \ldots, \lambda_d) \\ 0 \end{pmatrix} \beta_S + \tilde{\epsilon}, \quad \tilde{\epsilon} \sim N\left(0, \sigma_S^2 \begin{pmatrix} \mathrm{diag}(\lambda_1, \ldots, \lambda_d) & 0 \\ 0 & \tilde{\Gamma} \end{pmatrix}\right), \tag{23}$$

where $\tilde{\Gamma}$ is some positive-definite matrix that does not concern us in the following calculations. By Le Cam's method (Tsybakov, 2008, Chapter 2), we then obtain

$$\inf_{\hat{\beta}_i} \sup_{((P_S)_i, (P_T)_i) \in \Lambda U^2} \mathbb{E}_{P_S, P_T}[(\hat{\beta}_i - (\beta_T)_i)^2] \geq \frac{((\beta_T)_i^0 - (\beta_T)_i^1)^2}{16} \exp\left(-\frac{((\beta_T)_i^0 - (\beta_T)_i^1)^2}{2\sigma_T^2} - \frac{\lambda_i((\beta_S)_i^0 - (\beta_S)_i^1)^2}{2\sigma_S^2}\right),$$

for any

$$|(\beta_T)_i^0 - (\beta_S)_i^0| = |(\beta_T^0 - \beta_S^0)_i| \leq \sqrt{\alpha_i \wedge \lambda_i} U$$

and

$$|(\beta_T)_i^1 - (\beta_S)_i^1| = |(\beta_T^1 - \beta_S^1)_i| \leq \sqrt{\alpha_i \wedge \lambda_i} U.$$

From Lemma A.1, upon choosing $K = 1$, we know there exists $(\beta_T)_i^0, (\beta_T)_i^1, (\beta_S)_i^0, (\beta_S)_i^1$ such that

$$(\beta_T^0 - \beta_S^0)_i^2 \leq \alpha_i \wedge \lambda_i U^2, (\beta_T^1 - \beta_S^1)_i^2 \leq \alpha_i \wedge \lambda_i U^2, (\beta_T^0 - \beta_T^1)_i^2 = \frac{1}{\frac{1}{\lambda_i U^2 + \frac{\sigma_S^2}{\lambda_i}} + \frac{1}{\sigma_T^2}}$$

and

$$(\beta_S^0 \quad \beta_S^1)_i^2 \quad \frac{\frac{1}{\frac{?}{i}U^2+\frac{2}{S}_i}\frac{2}{S}_i}{\frac{1}{\frac{?}{i}U^2+\frac{2}{S}_i}+\frac{1}{2}_T}.$$

We therefore have

$$\frac{((\beta_T)_i^0 \quad (\beta_T)_i^1)^2}{16}\exp\left(\quad \frac{((\beta_T)_i^0 \quad (\beta_T)_i^1)^2}{2\sigma_T^2} \quad \frac{\lambda_i((\beta_S)_i^0 \quad (\beta_S)_i^1)^2}{2\sigma_S^2}\right) \quad \frac{\exp\left(\quad \frac{1}{2}\right)}{16}\frac{1}{\frac{1}{\frac{?}{i}U^2+\frac{2}{S}_i}+\frac{1}{2}_T},$$

and thus conclude the lower bound

$$\inf_b \sup_{D(\ S;\ T)\ \ U^2} \mathsf{E}_{P_S;P_T}[\tilde{\ell}(\widehat{\beta},\beta_T)] \quad \frac{\exp\left(\quad \frac{1}{2}\right)}{16}\sum_{i=1}^{d}\frac{1}{\frac{1}{\frac{?}{i}U^2+\frac{2}{S}_i}+\frac{1}{2}_T}.$$

□

A.3 Proof of Remark 2.1

Remark 2.1. *Using the channel capacity of a non-Gaussian additive noise channel (Ihara, 1978), we can improve the uniform constant* $\frac{\exp\left(\quad \frac{1}{2}\right)}{16}$ *to*

$$\max\left\{\frac{\exp\left(\quad \frac{1}{2}\right)}{16}, \left(\frac{\frac{2}{S}_i}{\alpha_i^? U^2+\frac{2}{S}_i}\right)^2\right\}$$

for the i*-th summand in* (16)*. Note that the second term is* 1 *if* $\alpha_i^? = 0$*, and it is arbitrarily close to* 1 *if* U *is su ciently small.*

Proof. It is well known that the minimax risk is lower bounded by the Bayesian risk

$$\inf_b \sup_{D(\ S;\ T)\ \ U^2} \mathsf{E}_{P_S;P_T}[\tilde{\ell}(\widehat{\beta},\beta_T)] \quad \inf_b \mathsf{E}[k\widehat{\beta} \quad \beta_T k_2^2],$$

where the expectation on the right-hand side refers to a fixed design model (i.e., the predictors in the source and target are given), there is a prior on both β_S and β_T, and the responses, conditional on the prior, follow the P_S and P_T models for source and target environments, respectively. We assume independent priors $(\beta_T)_i \quad N(0,\sigma^2)$, and further assume $(\beta_S)_i = (\beta_T)_i + \sqrt{\alpha_i^?}\Delta_i$ where Δ_i assigns a probability of 0.5 to $\quad U$ and a probability of 0.5 to U. By the maximum entropy of the Gaussian distribution and the data processing inequality, we have

$$\inf_b \mathsf{E}[k\widehat{\beta} \quad \beta_T k_2^2] \quad \frac{1}{2\pi e}\sum_{i=1}^{d} e^{2h((\ T)_i)\ \ 2I(\mathbf{Y};\mathbf{V};(\ T)_i)}.$$

Since mutual information is invariant under invertible transformations, we obtain

$$I(\mathbf{Y},\mathbf{V};(\beta_T)_i) = I(\tilde{\mathbf{V}},\tilde{\mathbf{Y}};(\beta_T)_i) = I(\tilde{\mathbf{V}}_i,\tilde{\mathbf{Y}}_i;(\beta_T)_i),$$

where $\tilde{V}$ and $\tilde{Y}$ are invertible transformations of V and Y; refer to (22) and (23). We also know that

$$\tilde{\mathbf{V}}_i = (\beta_T)_i + \tilde{\eta}_i$$

and

$$\frac{\tilde{\mathbf{Y}}_i}{\lambda_i} = (\beta_T)_i + \sqrt{\alpha_i^?}\Delta_i + \frac{1}{\lambda_i}\tilde{\epsilon}_i.$$

Noting the decomposition

$$I(\tilde{\mathbf{V}}_i,\tilde{\mathbf{Y}}_i;(\beta_T)_i) = I(\tilde{\mathbf{V}}_i;(\beta_T)_i) + I(\tilde{\mathbf{Y}}_i;(\beta_T)_i j\tilde{\mathbf{V}}_i),$$

then, as $\sigma^2 \to 1$, we have

$$I(\tilde{\mathbf{V}}_i; (\beta_T)_i) \to \frac{1}{2}(\log(\sigma^2) - \log(\sigma_T^2)).$$

For the second term of this decomposition, we obtain

$$I(\tilde{\mathbf{Y}}_i; (\beta_T)_i | \tilde{\mathbf{V}}_i) = \mathsf{E}_{\tilde{\mathbf{V}}_i}[I(\tilde{\mathbf{Y}}_i; (\beta_T)_i | \tilde{\mathbf{V}}_i = \tilde{v}_i)].$$

Due to conditional independence, we know that $\tilde{\mathbf{Y}}_i | ((\beta_T)_i, \tilde{\mathbf{V}}_i = \tilde{v}_i)$ has the same distribution as $\tilde{\mathbf{Y}}_i | (\beta_T)_i$. Hence, we see that

$$I(\tilde{\mathbf{Y}}_i; (\beta_T)_i | \tilde{\mathbf{V}}_i = \tilde{v}_i) = I(\tilde{\mathbf{Y}}_i; (\tilde{\beta}_T)_i),$$

where

$$(\tilde{\beta}_T)_i \sim \mathcal{N}((1 + \frac{1}{\sigma^2})^{-1} v_i, \sigma_T^2 (1 + \frac{1}{\sigma^2})^{-1}).$$

From the non-Gaussian additive noise channel capacity (Ihara, 1978), we have

$$I(\tilde{\mathbf{Y}}_i; (\tilde{\beta}_T)_i) \leq \frac{1}{2} \log(1 + \frac{\sigma_T^2 (1 + \sigma^2)^{-1}}{\alpha_i^2 U^2 + \frac{\sigma_S^2}{\lambda_i}}) + \mathrm{KL}(P_{\alpha_i \beta_i + \frac{1}{\lambda_i} \epsilon_i} \| \mathcal{N}(0, \alpha_i^2 U^2 + \frac{\sigma_S^2}{\lambda_i})),$$

and by the convexity of KL divergence, we obtain

$$\begin{aligned}
&\mathrm{KL}(P_{\alpha_i \beta_i + \frac{1}{\lambda_i} \epsilon_i} \| \mathcal{N}(0, \alpha_i^2 U^2 + \frac{\sigma_S^2}{\lambda_i})) \\
&\quad \leq \frac{1}{2} \mathrm{KL}(\mathcal{N}(\sqrt{\alpha_i^2} U, \frac{\sigma_S^2}{\lambda_i}) \| \mathcal{N}(0, \alpha_i^2 U^2 + \frac{\sigma_S^2}{\lambda_i})) + \frac{1}{2} \mathrm{KL}(\mathcal{N}(-\sqrt{\alpha_i^2} U, \frac{\sigma_S^2}{\lambda_i}) \| \mathcal{N}(0, \alpha_i^2 U^2 + \frac{\sigma_S^2}{\lambda_i})) \\
&\quad \to \log(\alpha_i^2 U^2 + \frac{\sigma_S^2}{\lambda_i}) - \log(\frac{\sigma_S^2}{\lambda_i}).
\end{aligned}$$

We therefore have

$$\lim_{\sigma^2 \to 1} h((\beta_T)_i) - I(\tilde{\mathbf{V}}_i, \tilde{\mathbf{Y}}_i; (\beta_T)_i) \geq \frac{1}{2} \left(\log(2\pi e) - \log(\frac{1}{\sigma_T^2} + \frac{1}{\alpha_i^2 U^2 + \frac{\sigma_S^2}{\lambda_i}}) \right) + \log(\frac{\sigma_S^2}{\lambda_i}) - \log(\alpha_i^2 U^2 + \frac{\sigma_S^2}{\lambda_i}),$$

and thus conclude the lower bound

$$\inf_{\hat{\beta}} \mathsf{E}[\|\hat{\beta} - \beta_T\|_2^2] \geq \sum_{i=1}^{d} \left(\frac{\frac{\sigma_S^2}{\lambda_i}}{\alpha_i^2 U^2 + \frac{\sigma_S^2}{\lambda_i}} \right)^2 \frac{1}{\frac{1}{\sigma_T^2} + \frac{1}{\alpha_i^2 U^2 + \frac{\sigma_S^2}{\lambda_i}}}.$$

□

A.4 Proof of Proposition 2.1

Proposition 2.1.

1. *For the LRM based solely on the source dataset, the estimator $\widehat{\theta}_S = E\widehat{\beta}_S$ satisfies*

$$\sup_{D(P_S, P_T) \leq U^2} \mathsf{E}_{P_S, P_T}[\ell(\widehat{\theta}_S, \theta_T)] = U^2 + \sigma_S^2 \sum_{i=1}^{d} \lambda_i^{-1}.$$

2. *For the LRM based solely on the target dataset, the estimator $\widehat{\theta}_T = E\widehat{\beta}_T$ satisfies*

$$\sup_{D(P_S, P_T) \leq U^2} \mathsf{E}_{P_S, P_T}[\ell(\widehat{\theta}_T, \theta_T)] = d\sigma_T^2.$$

3. *Finally, for the LRM based on pooling the source and target datasets, the estimator*

$$\widehat{\theta}_P = \left(\begin{pmatrix} \mathbf{X}^\top & \mathbf{W}^\top \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{W} \end{pmatrix} \right)^{-1} \begin{pmatrix} \mathbf{X}^\top & \mathbf{W}^\top \end{pmatrix} \begin{pmatrix} \mathbf{Y} \\ \mathbf{V} \end{pmatrix} \tag{18}$$

satisfies

$$\sup_{D(P_S;P_T) \le U^2} \mathbb{E}_{P_S;P_T}[\ell(\widehat{\theta}_P, \theta_T)] = U^2 \max_{1 \le i \le d} \left\{ \left(\frac{\lambda_i}{1+\lambda_i} \right)^2 \right\} + \sigma_T^2 \sum_{i=1}^{d} \left(\frac{1}{1+\lambda_i} \right)^2 + \sigma_S^2 \sum_{i=1}^{d} \frac{\lambda_i}{(1+\lambda_i)^2}.$$

Proof. It is more convenient to work with the reparametrization (11). Note that

$$\begin{aligned} \widehat{\beta}_S &= (E^\top \mathbf{X}^\top \mathbf{X} E)^{-1} E^\top \mathbf{X}^\top \mathbf{Y} \qquad \mathcal{N}(\beta_S, \sigma_S^2 \operatorname{diag}(\lambda_1^{-1}, \ldots, \lambda_d^{-1})), \\ \widehat{\beta}_T &= (E^\top \mathbf{W}^\top \mathbf{W} E)^{-1} E^\top \mathbf{W}^\top \mathbf{V} \qquad \mathcal{N}(\beta_T, \sigma_T^2 I). \end{aligned}$$

We then have for the estimator $\widehat{\theta}_S$

$$\begin{aligned} \sup_{D(P_S;P_T) \le U^2} \mathbb{E}_{P_S;P_T}[\|\widehat{\beta}_S - \beta_T\|_2^2] &= \sup_{\|\beta_S - \beta_T\|_2^2 \le U^2} \operatorname{Tr}\left(\sigma_S^2 \operatorname{diag}(\lambda_1^{-1}, \ldots, \lambda_d^{-1})\right) + \sum_{i=1}^{d} ((\beta_S)_i - (\beta_T)_i)^2 \\ &= \sigma_S^2 \sum_{i=1}^{d} \lambda_i^{-1} + U^2, \end{aligned}$$

and similarly for the estimator $\widehat{\theta}_T$

$$\sup_{D(P_S;P_T) \le U^2} \mathbb{E}_{P_S;P_T}[\|\widehat{\beta}_T - \beta_T\|_2^2] = \sup_{\|\beta_S - \beta_T\|_2^2 \le U^2} \operatorname{Tr}\left(\sigma_T^2 I\right) = \sigma_T^2 d.$$

For the pooling estimator (18), consider its reparametrization

$$\widehat{\beta}_P = \left(\begin{pmatrix} E^\top \mathbf{X}^\top & E^\top \mathbf{W}^\top \end{pmatrix} \begin{pmatrix} \mathbf{X} E \\ \mathbf{W} E \end{pmatrix} \right)^{-1} \begin{pmatrix} E^\top \mathbf{X}^\top & E^\top \mathbf{W}^\top \end{pmatrix} \begin{pmatrix} \mathbf{Y} \\ \mathbf{V} \end{pmatrix},$$

whose bias we can compute as

$$\begin{aligned} \mathbb{E}_{P_S;P_T}[\widehat{\beta}_P] - \beta_T &= \left(E^\top \mathbf{X}^\top \mathbf{X} E + E^\top \mathbf{W}^\top \mathbf{W} E\right)^{-1} \left(E^\top \mathbf{X}^\top \mathbf{X} E \beta_S + E^\top \mathbf{W}^\top \mathbf{W} E \beta_T\right) - \beta_T \\ &= \operatorname{diag}(1+\lambda_1, \ldots, 1+\lambda_d)^{-1} \left(\operatorname{diag}(\lambda_1, \ldots, \lambda_d) \beta_S + \beta_T\right) - \beta_T \\ &= \operatorname{diag}\left(\lambda_1/(1+\lambda_1), \ldots, \lambda_d/(1+\lambda_d)\right) (\beta_S - \beta_T), \end{aligned}$$

and whose variance we can compute as

$$\begin{aligned} \mathbb{E}_{P_S;P_T}[\|\widehat{\beta}_P - \beta_T\|_2^2] &= \operatorname{Tr}\left(\operatorname{diag}(1+\lambda_1, \ldots, 1+\lambda_d)^{-1} \left(\operatorname{diag}(\lambda_1, \ldots, \lambda_d) \sigma_S^2 + \sigma_T^2 I \right) \operatorname{diag}(1+\lambda_1, \ldots, 1+\lambda_d)^{-1} \right) \\ &= \sigma_T^2 \sum_{i=1}^{d} \left(\frac{1}{1+\lambda_i} \right)^2 + \sigma_S^2 \sum_{i=1}^{d} \frac{\lambda_i}{(1+\lambda_i)^2}. \end{aligned}$$

We therefore obtain

$$\begin{aligned} \sup_{D(P_S;P_T) \le U^2} \mathbb{E}_{P_S;P_T}[\|\widehat{\beta}_P - \beta_T\|_2^2] &= \sup_{\|\beta_S - \beta_T\|_2^2 \le U^2} \sigma_T^2 \sum_{i=1}^{d} \left(\frac{1}{1+\lambda_i} \right)^2 + \sigma_S^2 \sum_{i=1}^{d} \frac{\lambda_i}{(1+\lambda_i)^2} + \sum_{i=1}^{d} \left(\frac{\lambda_i}{1+\lambda_i} \right)^2 (\beta_S - \beta_T)_i^2 \\ &= U^2 \max_{1 \le i \le d} \left\{ \left(\frac{\lambda_i}{1+\lambda_i} \right)^2 \right\} + \sigma_T^2 \sum_{i=1}^{d} \left(\frac{1}{1+\lambda_i} \right)^2 + \sigma_S^2 \sum_{i=1}^{d} \frac{\lambda_i}{(1+\lambda_i)^2}, \end{aligned}$$

thus completing the proof. □

To see that the worst-case risk (14) is also smaller than that of the pooling method, we can compute

$$\frac{1}{\frac{1}{{}_T^2}+\frac{1}{{}_i^? U^2+\frac{{}_S^2}{{}_i}}} \quad \sigma_T^2\left(\frac{1}{1+\lambda_i}\right)^2 \quad \sigma_S^2\frac{\lambda_i}{(1+\lambda_i)^2}$$

$$= \left(\frac{\lambda_i}{1+\lambda_i}\right)^2\frac{\lambda_i\sigma_T^2+(2\sigma_T^2 \quad \sigma_S^2)}{\lambda_i\sigma_T^2+\lambda_i\alpha_i^? U^2+\sigma_S^2}\alpha_i^? U^2+\frac{\lambda_i}{(1+\lambda_i)^2}\frac{2\sigma_S^2\sigma_T^2 \quad \sigma_S^4 \quad \sigma_T^4}{\lambda_i\sigma_T^2+\lambda_i\alpha_i^? U^2+\sigma_S^2}$$

$$= \left(\frac{\lambda_i}{1+\lambda_i}\right)^2\frac{\lambda_i\sigma_T^2+(2\sigma_T^2 \quad \sigma_S^2)+\frac{2\ {}_S^2\ {}_T^2 \quad {}_S^4 \quad {}_T^4}{{}_i\ {}_i^? U^2}}{\lambda_i\sigma_T^2+\lambda_i\alpha_i^? U^2+\sigma_S^2}\alpha_i^? U^2.$$

It is then readily verified that

$$(2\sigma_T^2 \quad \sigma_S^2)+\frac{2\sigma_S^2\sigma_T^2 \quad \sigma_S^4 \quad \sigma_T^4}{\lambda_i\alpha_i^? U^2} \quad \lambda_i\alpha_i^? U^2+\sigma_S^2,$$

as this is equivalent to

$$2(\sigma_T^2 \quad \sigma_S^2)\lambda_i\alpha_i^? U^2 \quad \left(\lambda_i\alpha_i^? U^2\right)^2+\left(\sigma_S^2 \quad \sigma_T^2\right)^2.$$

Hence, we have the desired inequality

$$\sum_{i=1}^d \frac{1}{\frac{1}{{}_T^2}+\frac{1}{{}_i^? U^2+\frac{{}_S^2}{{}_i}}} \quad U^2 \max_{1 \ i \ d}\left\{\left(\frac{\lambda_i}{1+\lambda_i}\right)^2\right\}+\sigma_T^2\sum_{i=1}^d\left(\frac{1}{1+\lambda_i}\right)^2+\sigma_S^2\sum_{i=1}^d\frac{\lambda_i}{(1+\lambda_i)^2}.$$

A.5 Proof of Theorem 3.1

Theorem 3.1. *A lower bound of the minimax risk corresponding to the GLMs is given by*

$$\inf_{b} \sup_{D(\ {}_{S_m};\ {}_T)\ U_m^2;8m} \mathsf{E}[\ell(\widehat{\theta},\theta_T)] \quad \frac{e^{\ 1}}{800}\frac{d}{\sum_{m=1}^M \frac{1}{\frac{U_m^2}{d}+\frac{a^{(m)}(\ {}_{S_m})}{C_{S_m}\ {}_1^{(m)}}}+\frac{C_T}{b(\ {}_T)}}.$$

Proof. First consider the case where $d \quad 100$, for which we use Le Cam's method (Tsybakov, 2008, Chapter 2). For two pairs of parameters $(\theta^0_{S_1},\ldots,\theta^0_{S_M},\theta^0_T)$ and $(\theta^1_{S_1},\ldots,\theta^1_{S_M},\theta^1_T)$, we have

$$\inf_{b} \sup_{D(\ {}_{S_m};\ {}_T)\ U_m^2 8m} \mathsf{E}[\ell(\widehat{\theta},\theta_T)] \quad \frac{\ell(\theta_T^0,\theta_T^1)}{16}\exp\left\{\quad \mathrm{KL}((\theta^0_{S_1},\ldots,\theta^0_{S_M},\theta^0_T);(\theta^1_{S_1},\ldots,\theta^1_{S_M},\theta^1_T))\right\}.$$

By independence, we note that

$$\mathrm{KL}((\theta^0_{S_1},\ldots,\theta^0_{S_M},\theta^0_T);(\theta^1_{S_1},\ldots,\theta^1_{S_M},\theta^1_T)) = \sum_{m=1}^M \mathrm{KL}(\theta^0_{S_m};\theta^1_{S_m})+\mathrm{KL}(\theta^0_T;\theta^1_T)$$

and

$$\mathrm{KL}(\theta^0_{S_m};\theta^1_{S_m}) = \frac{1}{a^{(m)}(\sigma_{S_m})}\sum_{i=1}^{n_{S_m}}\Big(\Psi^{(m)}(hx_i^{(m)},\theta^1_{S_m}i) \quad \Psi^{(m)}(hx_i^{(m)},\theta^0_{S_m}i)$$
$$h(\Psi^{(m)})^{\theta}(hx_i^{(m)},\theta^0_{S_m}i)x_i^{(m)},\theta^1_{S_m} \quad \theta^0_{S_m}i\Big)$$
$$\frac{1}{a^{(m)}(\sigma_{S_m})}\sum_{i=1}^{n_{S_m}}\frac{1}{2}C_{S_m}\sum_{j;k}x_{ij}^{(m)}x_{ik}^{(m)}(\theta^1_{S_m} \quad \theta^0_{S_m})_j(\theta^1_{S_m} \quad \theta^0_{S_m})_k$$
$$\frac{C_{S_m}}{2a^{(m)}(\sigma_{S_m})}(\theta^1_{S_m} \quad \theta^0_{S_m})^{>}(\mathbf{X}^{(m)})^{>}\mathbf{X}^{(m)}(\theta^1_{S_m} \quad \theta^0_{S_m})$$
$$\frac{\lambda_1^{(m)}C_{S_m}}{2a^{(m)}(\sigma_{S_m})}(\theta^1_{S_m} \quad \theta^0_{S_m})^{>}\mathbf{W}^{>}\mathbf{W}(\theta^1_{S_m} \quad \theta^0_{S_m}).$$

Similarly, we obtain

$$\mathrm{KL}(\theta_T^0;\theta_T^1) \quad \frac{C_T}{2b(\sigma_T)}(\theta_T^1 \quad \theta_T^0)^\top \mathbf{W}^\top \mathbf{W}(\theta_T^1 \quad \theta_T^0).$$

Then, by Le Cam's bound (Tsybakov, 2008, Chapter 2), we have

$$\begin{aligned}\inf_{\mathsf{b}} \sup_{D(S_m;T)\ U_m^2 8m} \mathsf{E}[\ell(\widehat{\theta},\theta_T)] \quad & \frac{\ell(\theta_T^0,\theta_T^1)}{16}\exp\Big\{ \quad \mathrm{KL}((\theta_{S_1}^0,\ldots,\theta_{S_M}^0,\theta_T^0);(\theta_{S_1}^1,\ldots,\theta_{S_M}^1,\theta_T^1))\Big\} \\ & \frac{(\theta_T^1 \quad \theta_T^0)^\top \mathbf{W}^\top \mathbf{W}(\theta_T^1 \quad \theta_T^0)}{16}\exp\left\{ \quad \frac{C_T}{2b(\sigma_T)}(\theta_T^1 \quad \theta_T^0)^\top \mathbf{W}^\top \mathbf{W}(\theta_T^1 \quad \theta_T^0)\right\} \\ & \exp\left\{ \quad \sum_{m=1}^M \frac{\lambda_1^{(m)} C_{S_m}}{2a^{(m)}(\sigma_{S_m})}(\theta_{S_m}^1 \quad \theta_{S_m}^0)^\top \mathbf{W}^\top \mathbf{W}(\theta_{S_m}^1 \quad \theta_{S_m}^0)\right\} \\ = & \frac{k\beta_T^1 \quad \beta_T^0 k_2^2}{16}\exp\left\{ \quad \sum_{i=1}^d \left(\frac{C_T}{2b(\sigma_T)}(\beta_T^1 \quad \beta_T^0)_i^2 + \sum_{m=1}^M \frac{\lambda_1^{(m)} C_{S_m}}{2a^{(m)}(\sigma_{S_m})}(\beta_{S_m}^1 \quad \beta_{S_m}^0)_i^2\right)\right\},\end{aligned}$$

where

$$\theta_T^j = E\beta_T^j, \theta_{S_m}^j = E\beta_{S_m}^j$$

and $E \ 2 \ \mathsf{R}^{d \ \ d}$ is any matrix that satisfies

$$E^\top \mathbf{W}^\top \mathbf{W} E = I.$$

By Lemma A.1, for any $K \quad \frac{1}{4}$, we can choose $\beta_T^0, \beta_T^1, \beta_{S_m}^0, \beta_{S_m}^1, m \ 2 \ [M]$, such that

$$(\beta_T^0 \quad \beta_{S_m}^0)_i^2 \quad \frac{U_m^2}{d}, (\beta_T^1 \quad \beta_{S_m}^1)_i^2 \quad \frac{U_m^2}{d}, (\beta_T^0 \quad \beta_T^1)_i^2 = \frac{1}{K}\frac{1}{\sum_{k=1}^M \frac{1}{\frac{U_k^2}{d}+\frac{a^{(k)}(S_k)}{c_{S_k\ 1}^{(k)}}} + \frac{C_T}{b(T)}}$$

and

$$(\beta_{S_m}^0 \quad \beta_{S_m}^1)_i^2 \quad \frac{1}{K}\frac{\frac{1}{\frac{U_m^2}{d}+\frac{a^{(m)}(S_m)}{c_{S_m\ 1}^{(m)}}}\frac{a^{(m)}(S_m)}{c_{S_m\ 1}^{(m)}}}{\sum_{k=1}^M \frac{1}{\frac{U_k^2}{d}+\frac{a^{(k)}(S_k)}{c_{S_k\ 1}^{(k)}}} + \frac{C_T}{b(T)}}.$$

We therefore conclude

$$\begin{aligned}& \frac{k\beta_T^1 \quad \beta_T^0 k_2^2}{16}\exp\left\{ \quad \sum_{i=1}^d \left(\frac{C_T}{2b(\sigma_T)}(\beta_T^1 \quad \beta_T^0)_i^2 + \sum_{m=1}^M \frac{\lambda_1^{(m)} C_{S_m}}{2a^{(m)}(\sigma_{S_m})}(\beta_{S_m}^1 \quad \beta_{S_m}^0)_i^2\right)\right\} \\ & \qquad \frac{\exp\left(\ \frac{d}{2K}\right)}{16K}\frac{d}{\sum_{k=1}^M \frac{1}{\frac{U_k^2}{d}+\frac{a^{(k)}(S_k)}{c_{S_k\ 1}^{(k)}}} + \frac{C_T}{b(T)}} \\ & \qquad \frac{\exp\left(\ \frac{50}{K}\right)}{16K}\frac{d}{\sum_{k=1}^M \frac{1}{\frac{U_k^2}{d}+\frac{a^{(k)}(S_k)}{c_{S_k\ 1}^{(k)}}} + \frac{C_T}{b(T)}}, \qquad \text{for } d \quad 100, \\ & \qquad \frac{e^{\ 1}}{800}\frac{d}{\sum_{k=1}^M \frac{1}{\frac{U_k^2}{d}+\frac{a^{(k)}(S_k)}{c_{S_k\ 1}^{(k)}}} + \frac{C_T}{b(T)}},\end{aligned}$$

where $K = 50$ is chosen.

Now, for $d \quad 100$, we use Fano's method (Tsybakov, 2008, Chapter 2). Let

$$h = \sqrt{\frac{1}{4K}\frac{1}{\sum_{k=1}^M \frac{1}{\frac{U_k^2}{d}+\frac{a^{(k)}(S_k)}{c_{S_k\ 1}^{(k)}}} + \frac{C_T}{b(T)}}},$$

and consider the hypercube

$$C = \{\beta \in \mathbb{R}^d : \beta_i \in \{h, h\}, i = 1, \ldots, d\}.$$

Then, by the Varshamov-Gilbert Lemma, since $d \geq 8$, there exists a pruned hypercube $\beta_T^0, \ldots, \beta_T^J \in C$ such that $J \geq 2^{d/8}$ and $H(\beta_T^j, \beta_T^k) \geq \frac{d}{8}$ for $0 \leq j < k \leq J$, where H denotes the Hamming distance, namely

$$H(\beta_T^j, \beta_T^k) = \sum_{i=1}^{d} Id_{\{(\beta_T^j)_i \neq (\beta_T^k)_i\}}.$$

We therefore have

$$\min_{j \neq k} \|\beta_T^j - \beta_T^k\|_2^2 \geq \frac{d}{8K} \frac{1}{\sum_{k=1}^{M} \frac{1}{\frac{U_k^2}{d} + \frac{a^{(k)}(S_k)}{c_{S_k 1}^{(k)}}} + \frac{C_T}{b(\sigma_T)}}.$$

By Lemma A.1, for any $m \in [M]$, there exists k_m such that $0 \leq k_m \leq h$ and

$$(h - k_m)^2 \frac{U_m^2}{d}, (2k_m)^2 \frac{1}{K} \frac{\frac{1}{\frac{U_m^2}{d} + \frac{a^{(m)}(S_m)}{c_{S_m 1}^{(m)}}} \frac{a^{(m)}(S_m)}{c_{S_m 1}^{(m)}}}{\sum_{k=1}^{M} \frac{1}{\frac{U_k^2}{d} + \frac{a^{(k)}(S_k)}{c_{S_k 1}^{(k)}}} + \frac{C_T}{b(\sigma_T)}}.$$

Hence, choosing $\beta_{S_m}^j, j = 0, \ldots, J$, such that

$$\forall 1 \leq i \leq d, (\beta_{S_m}^j)_i = \begin{cases} k_m & \text{if } (\beta_T^j)_i = h \\ k_m & \text{if } (\beta_T^j)_i = -h \end{cases},$$

we obtain

$$\|\beta_T^j - \beta_{S_m}^j\|_2^2 \leq U_m^2, \quad j = 0, \ldots, J, \quad m \in [M],$$

and

$$\mathrm{KL}((\theta_{S_1}^j, \ldots, \theta_{S_M}^j, \theta_T^j); (\theta_{S_1}^k, \ldots, \theta_{S_M}^k, \theta_T^k)) \leq \frac{d}{2K}, \quad \forall 0 \leq j < k \leq M.$$

We therefore have, by Fano's bound (Tsybakov, 2008, Chapter 2),

$$\begin{aligned} \inf_{\widehat{\theta}} \sup_{D(S_m; \sigma_T) \leq U_m^2 \forall m} \mathbb{E}[\ell(\widehat{\theta}, \theta_T)] &\geq \frac{\min_{j \neq k} \|\beta_T^j - \beta_T^k\|_2^2}{4} \left(1 - \frac{\frac{d}{2K} + \log 2}{\log J}\right) \\ &\geq \frac{1}{32K} \left(1 - \frac{4}{K \log 2} - \frac{8}{d}\right) \frac{d}{\sum_{k=1}^{M} \frac{1}{\frac{U_k^2}{d} + \frac{a^{(k)}(S_k)}{c_{S_k 1}^{(k)}}} + \frac{C_T}{b(\sigma_T)}} \\ &\geq \frac{3}{3200} \frac{d}{\sum_{k=1}^{M} \frac{1}{\frac{U_k^2}{d} + \frac{a^{(k)}(S_k)}{c_{S_k 1}^{(k)}}} + \frac{C_T}{b(\sigma_T)}}, \end{aligned}$$

where $K = \frac{50}{3}$ is chosen and recalling that $d \geq 100$. □

A.6 Proof of Remark 3.1

Remark 3.1. *For the non-transfer learning setting considered in Lee and Courtade (2020), our proof method gives rise to a lower bound of*

$$d \geq b(\sigma_T)/C_T$$

which is sharper than their lower bound of

$$\max\left\{\frac{\|\Lambda_{\mathbf{W}}\|_1^2}{\|\Lambda_{\mathbf{W}}\|_2^2}, \lambda_{min}(\mathbf{W}^\top \mathbf{W}) \|\Lambda_{\mathbf{W}}^{-1}\|_1\right\} \geq b(\sigma_T)/C_T,$$

where $\Lambda_{\mathbf{W}}$ is the vector of eigenvalues of the positive-definite matrix $\mathbf{W}^\top \mathbf{W}$, and $\Lambda_{\mathbf{W}}^{-1}$ denotes its coordinate-wise inverse.

Proof. By Holder's inequality, we have

$$k\Lambda_{\mathbf{W}}k_1^2 \quad k\Lambda_{\mathbf{W}}k_2^2 d.$$

It is also readily verified that

$$\lambda_{\min}(\mathbf{W}^{>}\mathbf{W})k\Lambda_{\mathbf{W}}^{1}k_1 \quad d,$$

and the desired result follows. □

A.7 Proof of Corollary 3.1

Corollary 3.1. *When the GLMs considered are Gaussian LRMs, then a lower bound of the minimax risk is*

$$\inf_{\mathsf{b}} \sup_{D(\ _{S_m};\ _T)\ \ U_m^2;\delta m} \mathsf{E}[\ell(\widehat{\theta},\theta_T)] \quad \frac{e^{\ 1}}{800}\frac{d}{\sum_{m=1}^{M}\frac{1}{\frac{U_m^2}{d}+\frac{\ ^2_{S_m}}{\ ^{(m)}_1}}+\frac{1}{\ ^2_T}}.$$

Proof. Upon simply noting that, for Gaussian LRMs, we have $a^{(m)}(\sigma_{S_m}) = \sigma_{S_m}^2$, $b(\sigma_T) = \sigma_T^2$ and $C_{S_m} = C_T = 1$, the desired result then follows. □

A.8 Auxiliary Result on the GLM Upper Bound

We now provide details on the GLM upper bound in equation (20), where we additionally assume that

$$\inf_z(\Psi^{(m)})^{\theta\theta}(z) \quad L_{S_m}\ \delta m, \qquad \inf_z \Gamma^{\theta\theta}(z) \quad L_T.$$

Consider the usual MLE estimator $\widehat{\theta}_{S_m}$ and $\widehat{\theta}_T$ for the source and target domains, and further consider a simplified interpolator

$$\widehat{\theta}_t = \sum_{m=1}^{M} t_m\widehat{\theta}_{S_m} + t_{M+1}\widehat{\theta}_T, \qquad \sum_{m=1}^{M+1} t_m = 1, \quad t_m \quad 0.$$

For any fixed admissible parameters satisfying $D(\theta_{S_m},\theta_T) \quad U_m^2,\delta m$, we claim that with probability at least $1 \quad e^{\ c}$, it holds that

$$\ell(\widehat{\theta}_{t^?},\theta_T) = (\widehat{\theta}_{t^?} \quad \theta_T)\mathbf{W}^{>}\mathbf{W}(\widehat{\theta}_{t^?} \quad \theta_T) \quad \frac{d}{\sum_{m=1}^{M}\frac{1}{\frac{U_m^2}{d}+\frac{2C_{S_m}a^{(m)}(\ _{S_m})}{L_{S_m}^2\ ^{(m)}_d}(c+\log(2dm))}+\frac{1}{\frac{2C_T b(\ _T)}{L_T^2}(c+\log(2dm))}},$$

where $t^?$ solves

$$\inf_{t=(t_1;:::;t_{M+1})\ \ 0;\ ^{\mathsf{P}}_{m=1}{}^{M+1}t_m=1}\sum_{m=1}^{M} t_m^2\left(\frac{2dC_{S_m}a^{(m)}(\sigma_{S_m})}{L_{S_m}^2\lambda_d^{(m)}}(c+\log(2dm))+U_m^2\right)+t_{M+1}^2\frac{2dC_T b(\sigma_T)}{L_T^2}(c+\log(2dm)).$$

Given that the dimension and number of sources are fixed, we consider $c \quad \log(2dm)$ and compare against our lower bound in Theorem 3.1, namely

$$\frac{e^{\ 1}}{800}\frac{d}{\sum_{m=1}^{M}\frac{1}{\frac{U_m^2}{d}+\frac{a^{(m)}(\ _{S_m})}{C_{S_m}\ ^{(m)}_1}}+\frac{C_T}{b(\ _T)}},$$

from which we find that there are gaps due to the ratios $\frac{C_{S_m}}{L_{S_m}}$, $\frac{C_T}{L_T}$ and the eigen gap $\frac{\ ^{(m)}_1}{\ ^{(m)}_d}$.

Proof of Upper Bound. Using the sub-Gaussian concentration bound for GLM noise and the trick in Lemma 8 of Bastani (2021), we have the concentration bounds

$$P\left((\widehat{\theta}_{S_m} \quad \theta_{S_m})\mathbf{W}^{>}\mathbf{W}(\widehat{\theta}_{S_m} \quad \theta_{S_m}) \quad \frac{2dC_{S_m}a^{(m)}(\sigma_{S_m})}{L_{S_m}^2\lambda_d^{(m)}}(c_1+\log(2d))\right) > 1 \quad e^{\ c_1}, \qquad \delta m\ 2\ [M],$$

and

$$P\left((\widehat{\theta}_T \quad \theta_T)\mathbf{W}^{>}\mathbf{W}(\widehat{\theta}_T \quad \theta_T) \quad \frac{2dC_T b(\sigma_T)}{L_T^2}(c_1+\log(2d))\right) > 1 \quad e^{\ c_1}.$$

Due to independence, the probability of the intersection of the events happening is greater than $(1 \quad e^{\ c_1})^m$ $1 \quad 2me^{\ c_1}$ for large c_1. On the intersection of the events, we solve the problem

$$\inf_{t=(t_1;:::;t_{M+1}) \ 0;\ \sum_{m=1}^{M+1} t_m = 1} \sum_{m=1}^{M} t_m^2 \left(\frac{2dC_{S_m} a^{(m)}(\sigma_{S_m})}{L_{S_m}^2 \lambda_d^{(m)}}(c_1+\log(2d)) + U_m^2\right) + t_{M+1}^2 \frac{2dC_T b(\sigma_T)}{L_T^2}(c_1+\log(2d)),$$

with optimal solution $t^?$. We have that the optimal interpolator defined by

$$\widehat{\theta}_{t^?} = \sum_{m=1}^{M} t_m^? \widehat{\theta}_{S_m} + t_{M+1}^? \widehat{\theta}_T,$$

satisfies

$$(\widehat{\theta}_{t^?} \quad \theta_T)\mathbf{W}^{>}\mathbf{W}(\widehat{\theta}_{t^?} \quad \theta_T) \quad \frac{d}{\sum_{m=1}^{M} \frac{1}{\frac{U_m^2}{d} + \frac{2C_{S_m} a^{(m)}(\ S_m)}{L_{S_m}^2 \ \ (m)\ d}(c_1+\log(2d))} + \frac{1}{\frac{2C_T b(\ T)}{L_T^2}(c_1+\log(2d))}}.$$

We choose $c_1 = c_1^{\theta} + \log(2m)$ so that this event happens with probability at least $1 \quad e^{\ c_1^{\theta}}$. □

A.9 Auxiliary Result on the Comparison with Kalan et al. (2020)

Kalan et al. (2020) study a minimax lower bound for the linear regression setting (albeit under random design) and involve the spectral gap of the generalized eigenvalue problem we consider, with analogous definitions for the population distribution of their random design setting. The significance of our geometric perspective is best illustrated in comparison with their results where, in strong contrast, our analysis (in fixed design) takes care of the entire spectrum of the generalized eigenvalues. More precisely, our lower bound scaled by $1/n_T$ is lower bounded by (ignoring the constant $\exp(\ 1/2)/16$):

$$\frac{1}{n_T}\sum_{i=1}^{d} \frac{1}{\frac{1}{2} + \frac{1}{\frac{?}{i}U^2 + \frac{2}{i}}} \ \backslash\text{spectral gap"}\ \frac{1}{n_T}\frac{d\sigma^2}{1+\frac{1}{\frac{U^2}{d^{\ 2}} + \frac{1}{1}}} \quad \frac{1}{n_T}\begin{cases} c_1\sigma^2 d & \text{if } U^2 \quad \tilde{c}_1\sigma^2 d, \\ c_2 U^2 & \text{if } \tilde{c}_2\frac{^2d}{1+\ _1} \quad U^2 \quad \tilde{c}_1\sigma^2 d, \\ c_3\frac{^2d}{1+\ _1} & \text{if } U^2 \quad \tilde{c}_2\frac{^2d}{1+\ _1}, \end{cases}$$

where $c_i, \tilde{c}_i$ are universal constants. The last expression essentially amounts to the lower bound in equation (3.1) in Kalan et al. (2020), adjusting for the random designs and the scaling of U. We emphasize that the spectral gap may cause the last expression to be arbitrarily suboptimal, e.g., when $U = o(1)$, $\lambda_1 \ ! \ 1$ and $\lambda_d = O(1)$, while our analysis is sharp (i.e., the upper bound and lower bound match).

B Simulation Results

We first describe the choice of the algorithm from Li et al. (2020) used for comparison against our proposed method. Since the setting is a single source domain from which learning is transferred to a target domain, we only consider Algorithm 1 from Li et al. (2020), i.e., the (original) Oracle Trans-Lasso algorithm. Algorithm 4 from Li et al. (2020) uses the l_0-norm to quantify the difference between the source and target parameters. However, for this algorithm, they require that the l_0-difference (denoted by h_0) is much smaller than the l_0 sparsity of the target parameter (denoted by s) for the learning to be effective, which does not hold in our simulation settings and in the real-world dataset (where $h_0 = 8$ and $s \quad 8$ from Table 4 and Figure 8 in Section B.2). The Oracle Trans-Lasso algorithm from Appendix C.2 in Li et al. (2020) uses the l_q-norm, $q \ 2 \ (0,1)$, to quantify the difference between the source and target parameters. However, for this algorithm, they require that the l_q-difference (denoted by h_q) is much smaller than $\sqrt{s\log d/n_T}$ for the learning to be effective, which also does not hold in our simulation settings and in the real-world dataset (refer to Table 4 in Section B.2 and note that $\sqrt{\log d/n_T}$ is between 0.018 and 0.2). We therefore choose to compare our proposed method with Algorithm 1 from Li et al. (2020), which uses the l_1-norm to quantify the difference.

B.1 Additional Simulation Comparisons

As part of our additional simulation results related to the misspecification of U, we vary the parameter values from the baseline in Section 4.1 such that the rows of $\mathbf{X}$ are generated independently by a zero-mean Gaussian with a Toeplitz covariance matrix (Li et al., 2020, Section 5.2), or that both $\mathbf{X}$ and $\mathbf{W}$ are generated in this way. The corresponding results are provided in the left plot and right plot of Figure 2, respectively. We observe from these results that the introduction of correlation, either in only $\mathbf{X}$ or in both $\mathbf{X}$ and $\mathbf{W}$, does not impact the performance of either method when compared to the uncorrelated case in the left plot of Figure 1.

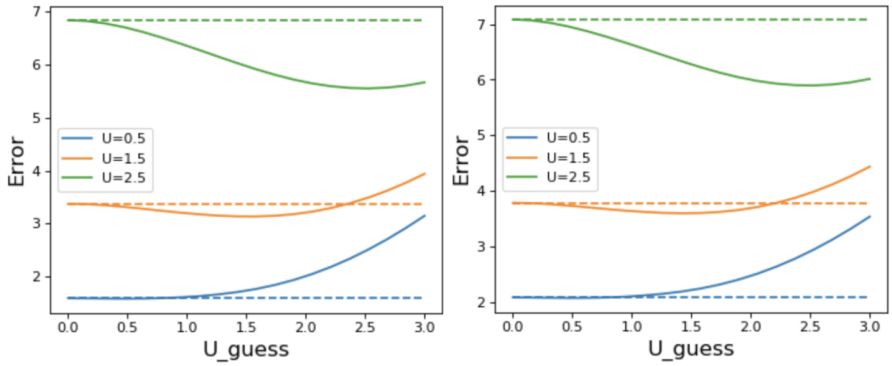

Figure 2: Simulation Results with Toeplitz Covariance Matrix for Designs. Solid Lines (Dashed Lines) Represent the Proposed Method (Basic Pooling Method).

We also consider varying the noise variances. In particular, the noise variances are changed to $\sigma_S^2 = 1, \sigma_T^2 = 5$ or $\sigma_S^2 = 5, \sigma_T^2 = 1$, the results of which are provided in the left plot and right plot of Figure 3, respectively. The proposed method handles large variances in the source data much better than the basic pooling method, while high variance in (smaller sized) target data has no significant impact on either method.

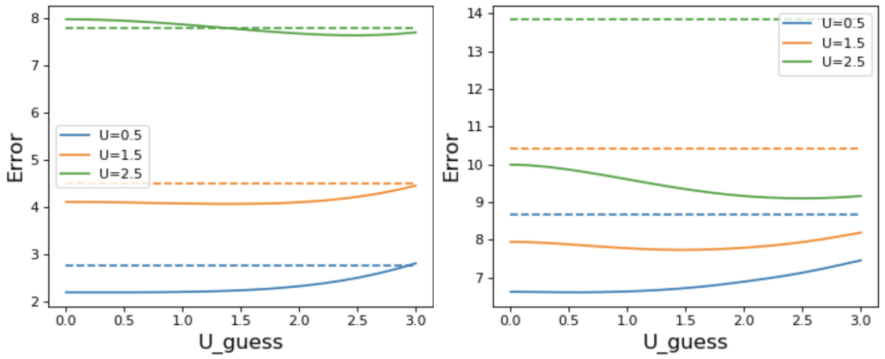

Figure 3: Simulation Results with Unequal Noise Variances. Solid Lines Represent the Proposed Method. Dashed Lines Represent the Basic Pooling Method.

We further consider varying the different dimensions. In particular, the dimension d is changed to 5 or $d = 100$ with β_T three-sparse; specifically, the first three elements of β_T are one and the rest are zero. The corresponding results are provided in the left plot and right plot of Figure 4, respectively. Lower dimensionality (left) seems to improve the performance gap between the two methods as compared to the left plot of Figure 1, while this gap shortens in high-dimensions with extreme sparsity (right).

We additionally consider varying the magnitude (sup-norm) of the ground truth value of the parameter β_T from the baseline in Section 4.1 such that β_T is the vector whose component values are all 0.1, or that β_T is the vector whose component values are all 10. The corresponding results are provided in the left plot and right plot of Figure 5, respectively. We observe that the plots remain exactly the same as the left plot of Figure 1, which is

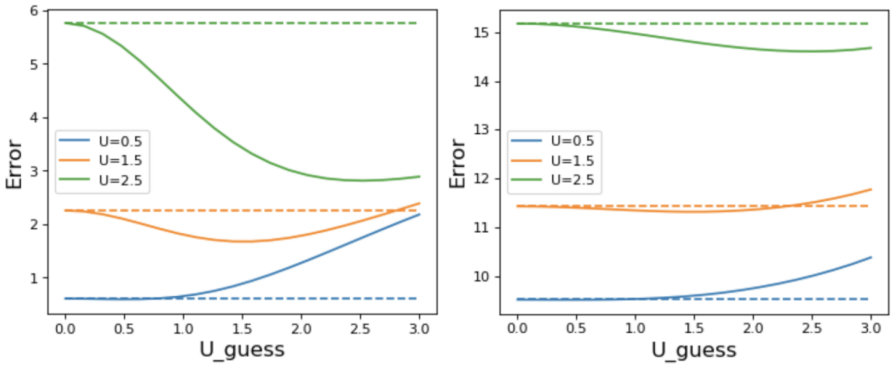

Figure 4: Simulation Results with Different Dimensions. Solid Lines Represent the Proposed Method. Dashed Lines Represent the Basic Pooling Method.

consistent with Theorem 2.1 that the worst-case performance of our method does not depend on the magnitude of $\beta_{\mathcal{T}}$, and similarly for the basic pooling method.

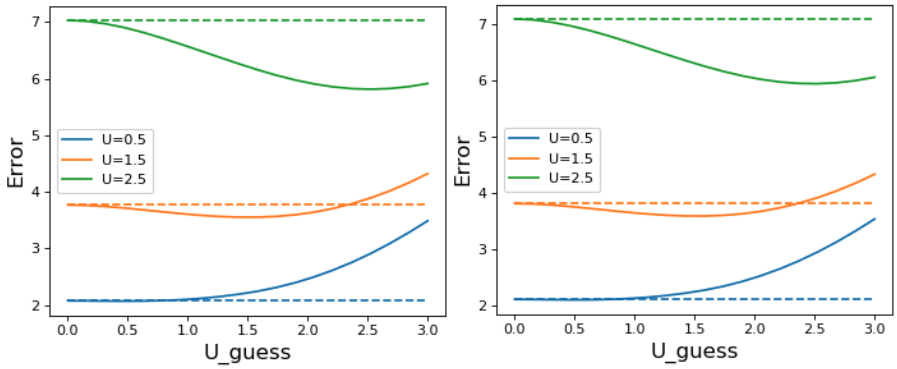

Figure 5: Simulation Results with Different Magnitudes of the Ground Truth $\beta_{\mathcal{T}}$. Solid Lines (Dashed Lines) Represent the Proposed Method (Basic Pooling Method).

We next investigate the behavior of the performance gap as the sample size increases. In particular, we vary the sample sizes to be $n_{\mathcal{S}} = 10000$ and $n_{\mathcal{T}} = 1000$, and present the corresponding results in Figure 6. We observe from these results that the plot remains the same as the left plot of Figure 1, demonstrating the robustness of the performance gap.

As part of our additional simulation results related to the comparisons against competing methods in the literature, we consider the same settings in Section 4.2.1 and Section 4.2.2 with a smaller $n_{\mathcal{S}}/n_{\mathcal{T}}$ ratio. In particular, we vary the sample sizes to be $n_{\mathcal{S}} = 200$ and $n_{\mathcal{T}} = 100$. For the methods under consideration, we report the average estimation error of $\theta_{\mathcal{T}}$ in (5), and its standard deviation, over 1000 simulation runs. The results are summarized in the left-half of Table 3 for moderate-dimensions and the right-half of Table 3 for high-dimensions. We observe behaviors of our proposed method relative to the competing methods to be consistent with those in Section 4.2.1 and Section 4.2.2.

B.2 Real-World Dataset Experiments

As part of our final set of empirical results considered in Section 4.2.3, we compare our proposed method against the other competing methods using the Uber&Lyft dataset (https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma) of Uber and Lyft cab rides collected in Boston, MA. Recall that we consider UberX to be the source model and standard Lyft service to be the target model where the learning problem comprises prediction of the price using $d = 32$ numerical features. Given that the focus of our study is on the benefit of transfer learning, we restrict our experiments to small random subsamples and we summarize in Table 2 the

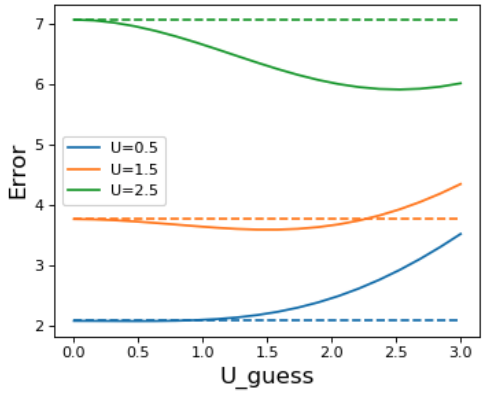

Figure 6: Simulation Results with A Larger Sample Size. Solid Lines (Dashed Lines) Represent the Proposed Method (Basic Pooling Method).

Table 3: Simulation Results Comparing the Proposed Method to Other Competing Methods in Moderate-Dimensions (Left-Half) and in High-Dimensions (Right-Half). "Basic" Represents the Lowest of the Errors Attained by the Three Basic Methods in Section 2.2. Numbers in Parentheses Are Standard Deviations.

U	Basic	Proposed	Two-Step	Trans-Lasso	Basic	Proposed	Two-Step	Trans-Lasso
0.5	6.8(2.1)	7.1(2.2)	10.6(3.6)	**1.1**(2.4)	33.9(7.2)	**30.9**(7.6)	61.1(23.7)	34.2(7.5)
1.5	8.4(2.5)	8.5(2.7)	11.4(3.5)	**1.7**(2.2)	36.3(6.2)	**32.6**(6.6)	62.9(22.2)	37.4(6.6)
2.5	10.5(2.7)	9.9(2.5)	14.1(4.5)	**2.5**(2.6)	38.6(5.9)	**34.3**(6.1)	69.8(22.1)	40.5(6.3)
3.5	13.1(3.2)	11.6(3.1)	15.5(4.6)	**4.8**(3.4)	44.6(5.6)	**38.3**(6.4)	77.1(23.7)	47.6(6.0)
4.5	17.5(3.7)	14.5(3.9)	17.6(6.2)	**8.7**(4.5)	50.8(6.1)	**41.5**(5.8)	84.8(23.5)	55.2(7.4)
5.5	19.3(5.8)	15.3(4.8)	18.7(7.2)	**11.7**(4.9)	58.3(7.2)	**45.4**(8.1)	98.8(24.9)	63.6(8.1)
6.5	19.3(6.4)	17.3(6.0)	19.7(7.7)	**16.6**(6.6)	55.9(12.5)	**49.3**(8.7)	115.9(29.4)	74.4(7.5)
7.5	19.4(6.8)	**18.3**(6.4)	19.3(7.5)	20.4(7.2)	57.1(11.6)	**52.9**(9.4)	130.3(30.1)	87.3(9.8)
8.5	20.1(5.5)	**18.5**(6.1)	19.0(6.4)	21.7(9.4)	56.3(12.8)	**53.4**(10.7)	151.8(28.3)	98.9(9.8)
9.5	21.4(12.4)	**18.8**(6.9)	20.5(7.6)	23.2(9.0)	57.7(13.9)	**56.9**(9.6)	167.6(34.5)	112.5(10.8)

results taken over 100 repeated independent experiments.

The Uber&Lyft dataset consists of 55094 observations for the source and 51235 observations for the target, from which we obtain and present in Figures 7 and 8 the corresponding ground-truth regression parameters as bar plots for the source and target models, respectively. We observe that the parameters are moderately sparse with the feature "surge_multiplier" having much larger magnitudes in comparison with the other features. Figure 9 plots the difference between these two parameters, from which we again observe that the difference is sparse. We also compute the l_q-distance, q *2* $f0, 0.5, 1g$, between the source and target parameters and summarize these results in Table 4.

The results in Table 2 show that our proposed method attains a better performance on average, by a small margin relative to the basic methods and by a large margin relative to the two-step estimator and Trans-Lasso. We note that the l_q sparsity, q *2* $[0, 1]$, required by the last two methods does not reasonably capture the contrast between the source and target models of the real-world dataset, due to the moderate dimensions and the existence of one dominating feature In particular, while Bastani (2021) shows that the two-step joint estimator performs well when the difference of regression parameters is l_0 sparse, their result applies to high-dimensions which is in contrast to the 32 dimensions of the dataset at hand. This helps to explain why their two-step joint estimator does not yield good performance in Table 2 where the moderate dimensions and a single feature of "surge_multiplier" significantly affects the model. Li et al. (2020) show that their Trans-Lasso method performs well when the l_1-difference of the regression parameters (denoted by h_1) satisfies h_1 $s\sqrt{\log d/n_T}$, where s is the l_0 sparsity of the target regression parameters. However, for the real-world dataset, we observe that s 8 (see Figure 8), that the factor $\sqrt{\log(d)/n_T}$ is smaller than 0.2, that h_q (i.e., l_q-difference of the regression parameters) increases

as q decreases from 1 toward 0, and that h_q presents a discontinuity at $q = 0$ with $h_0 = 8$, some of which is illustrated in Table 4. Hence, the l_1 (or l_q in general) relationship assumed by Li et al. (2020) does not appear to hold for the dataset at hand. This in turn helps to explain why Trans-Lasso does not yield good performance in Table 2.

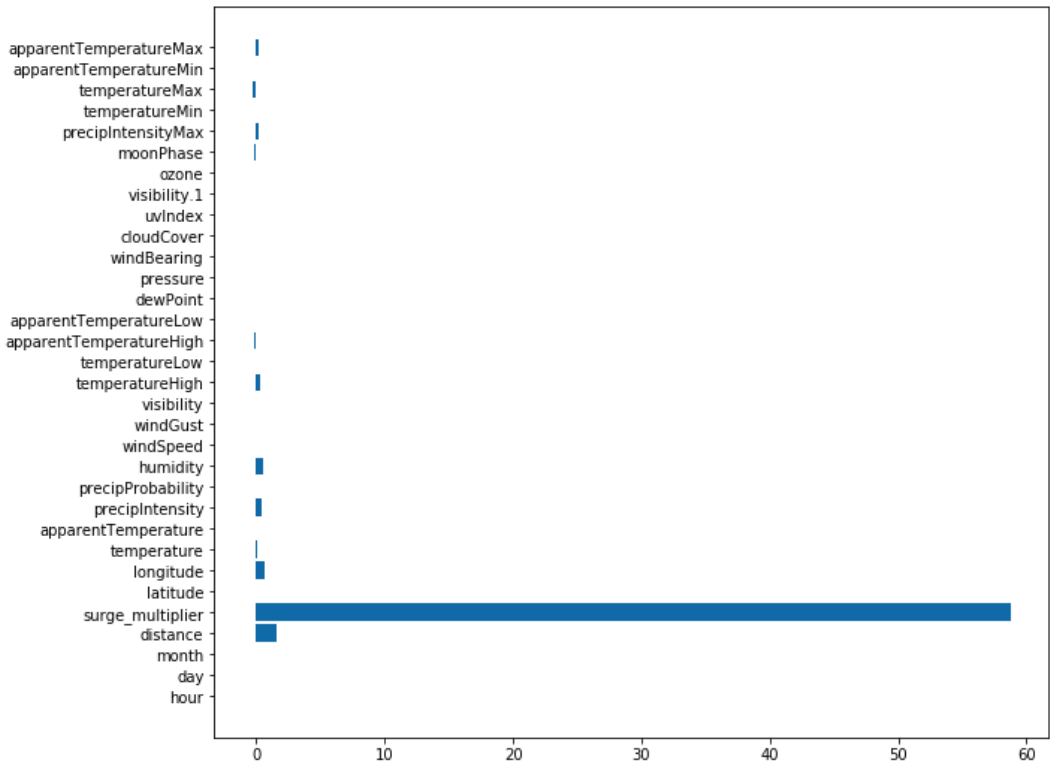

Figure 7: Estimated Regression Parameters of Source Model From the Entire Source Dataset.

Table 4: Results For Distance (under Different Sparsity Norms) between Source and Target Ground-truth Parameters on Uber&Lyft Data. The l_0-distance Has a Threshold of 0.1 to Determine Non-zero Entries.

l_0-distance (h_0)	$l_{0:5}$-distance ($h_{0:5}$)	l_1-distance (h_1)
8	189.67	53.65

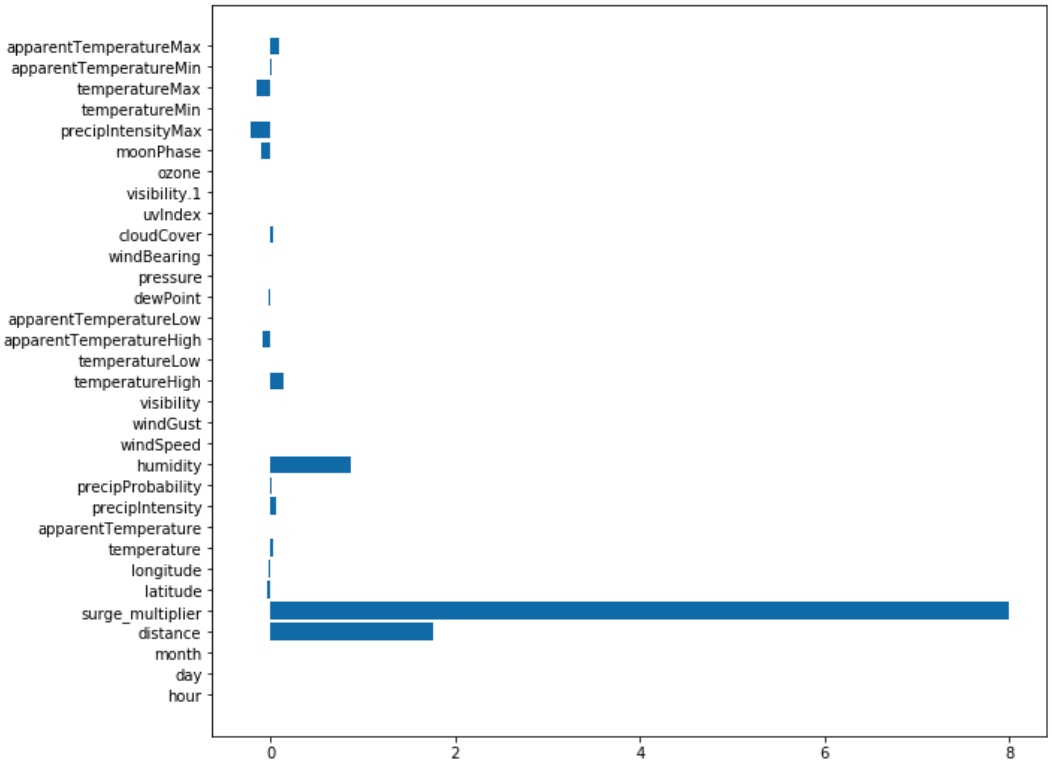

Figure 8: Estimated Regression Parameters of Target Model From the Entire Target Dataset.

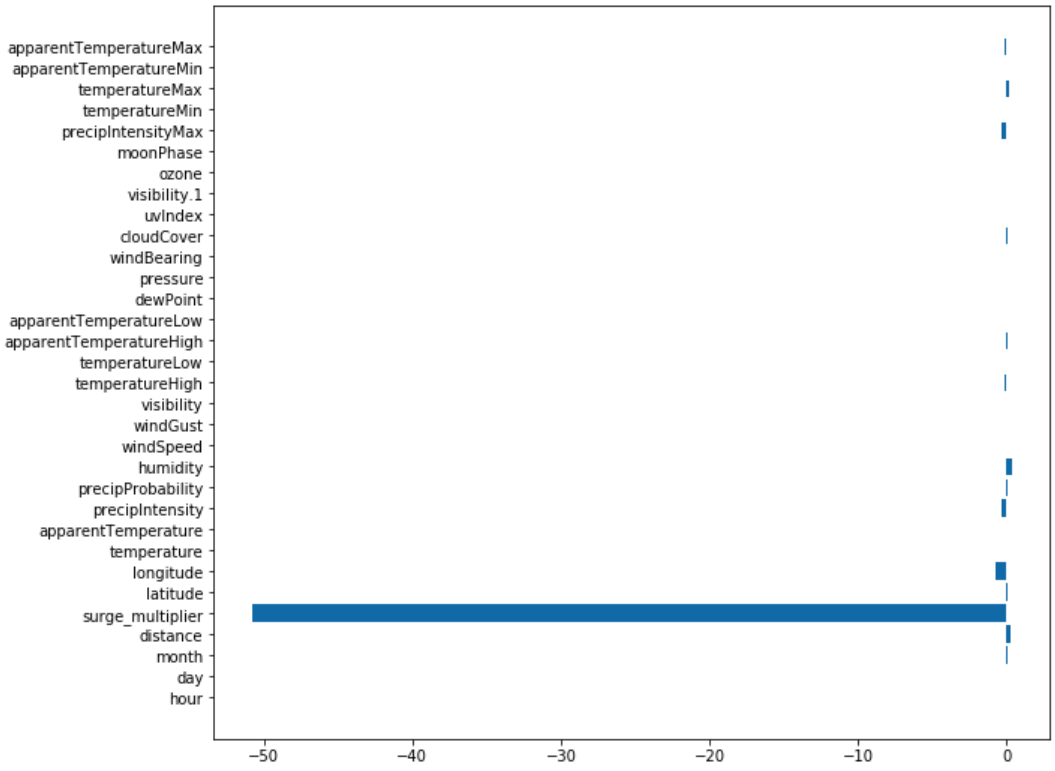

Figure 9: Difference in Regression Parameters of Target versus Source Model From the Entire Dataset.