Robust Covariate Shift Regression

Xiangli Chen **Mathew Monfort** **Anqi Liu** **Brian D. Ziebart**
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607
{xchen40,mmonfo2,aliu33,bziebart}@uic.edu

Abstract

In many learning settings, the source data available to train a regression model differs from the target data it encounters when making predictions due to input distribution shift. Appropriately dealing with this situation remains an important challenge. Existing methods attempt to "reweight" the source data samples to better represent the target domain, but this introduces strong inductive biases that are highly extrapolative and can often err greatly in practice. We propose a robust approach for regression under covariate shift that embraces the uncertainty resulting from sample selection bias by producing regression models that are explicitly robust to it. We demonstrate the benefits of our approach on a number of regression tasks.

1 Introduction

Linear regression is prevalently employed across the data sciences to understand relationships between variables and to make predictions [18, 7, 27]. It is commonly assumed, in various ways, that the labeled source data available to train the linear regression model (pairs of input vectors and output scalars) is representative of the target data that it will use for making predictions (given only input vectors). However, in many datasets, this assumption is not valid; source data can often come from biased portions of the input space and measured relationships often violate the linearity assumption for the output variable made by the linear regression model. One form of this problem, where source data is biased and potentially non-representative, is known as **covariate shift**. It occurs when the (stochastic) mapping from inputs to output is shared by the source and target data (i.e., source/target data is distributed according to $f_{\text{src}}(\mathbf{x})f(y|\mathbf{x})$ and $f_{\text{trg}}(\mathbf{x})f(y|\mathbf{x})$), but the distribution of inputs can vary ($f_{\text{src}}(x) \neq f_{\text{trg}}(x)$).

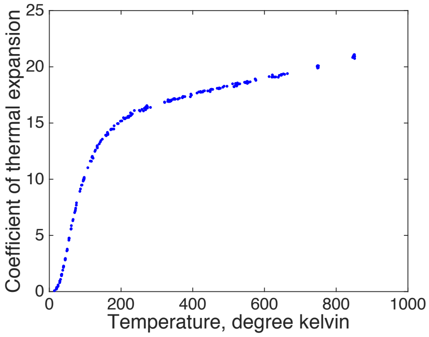

Figure 1: Hahn1 dataset [21] representing the result of a National Institute of Standards and Technology (NIST) study of the thermal expansion of copper.

Existing methods for addressing covariate shift employ importance weighting in an attempt to reweight available source data so that it may serve as an unbiased estimate for functions of the target data [30]. Unfortunately, when the amount of covariate shift is substantial and the number of source samples is small, importance weighting often leads to very high variance estimates [13] and inaccurate regression models. We illustrate the fundamental problems faced in regression under covariate shift using the Hahn1 dataset shown in Figure 1 as a running example in this paper. Locally, the datapoints appear linear in many portions of this dataset. A biased sample of datapoints from strictly within those regions will often mislead existing importance weighting methods into incorrectly linearly extrapolating beyond the data.

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51. Copyright 2016 by the authors.

Drawing on a recently developed method for robustly learning classifiers from data that has sample selection bias [26], we propose a novel robust approach for regression in the covariate shift setting. Our approach assumes that the stochastic mapping from inputs to the output variable is as similar as possible to a "zero-knowledge" reference distribution, except where statistics measured from source data indicate otherwise. We develop our **robust bias-aware regression** approach and illustrate its behavior on this dataset before evaluating it using additional higher-dimensional datasets.

2 Related Work

Domain adaptation tasks of learning from a source domain and predicting in a target domain have been an area of significant investigation [15, 4, 8, 19, 39, 13], but it is an inherently difficult task [5]. We focus on the special case of covariate shift [30, 38, 16, 17, 22, 32, 9, 37, 31, 36, 29], in which the conditional distribution of the predicted variable is shared in both source and target domains. In this section, we review the least squares linear regression formulation, existing covariate shift methods for regression, and robust minimax methods leveraged by our approach.

2.1 Least squares linear regression

Ordinary least squares (OLS) regression assumes a linear relationship between input variables $\mathbf{x}$ and output variable y, $\hat{y}_{a,\mathbf{b}}(\mathbf{x}) = \mathbf{b}^T\mathbf{x}+a$, that is parameterized by weights a (scalar) and $\mathbf{b}$ (vector). A standard method for estimating these parameters is to minimize the sum of squared residuals between estimator $\hat{y}_{a,\mathbf{b}}(\mathbf{x}_i)$ and the actual output variable y_i for each example i. This is equivalent to an expected squared loss over the empirical distribution $\tilde{f}(\mathbf{x})\tilde{f}(y|\mathbf{x})$:

$$\operatorname*{argmin}_{a,\mathbf{b}} \mathbb{E}_{\tilde{f}(\mathbf{x})\tilde{f}(y|\mathbf{x})} \left[(Y - \hat{y}_{a,\mathbf{b}}(\mathbf{X}))^2 \right]. \tag{1}$$

This optimization (1), is equivalent to maximizing the conditional log likelihood of sample data,

$$\operatorname*{argmax}_{a,\mathbf{b},\sigma} \mathbb{E}_{\tilde{f}(\mathbf{x})\tilde{f}(y|\mathbf{x})} \left[\log \hat{f}_{a,\mathbf{b},\sigma}(Y|\mathbf{X}) \right], \tag{2}$$

with $\hat{f}(y|\mathbf{x})$ normally distributed with mean $\hat{y}_{a,\mathbf{b}}(\mathbf{x})$ and variance σ^2 [7]. This corresponds to a zero-mean Gaussian residual error model with variance σ^2.

2.2 Importance weighted linear regression

Under covariate shift, the input of source data for estimating the linear regression model comes from a distribution $f_{\text{src}}(\mathbf{x})$ that differs from the target data input distribution $f_{\text{trg}}(\mathbf{x})$ on which it will be employed to make predictions. OLS under covariate shift (1) does not minimize the residual error (equivalently maximize the likelihood) of data drawn from the target distribution, $f_{\text{trg}}(\mathbf{x})f(y|\mathbf{x})$, and it is not a consistent estimator. Importance weighted least squares (IWLS) [31] is often employed to reweight the source data by the importance ratio, $f_{\text{trg}}(\mathbf{x})/f_{\text{src}}(\mathbf{x})$, to estimate the target distribution's residual error, which is then minimized:

$$\operatorname*{argmin}_{a,\mathbf{b}} \mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\tilde{f}(y|\mathbf{x})} \left[\frac{f_{\text{trg}}(\mathbf{X})}{f_{\text{src}}(\mathbf{X})} (Y - \hat{y}_{a,\mathbf{b}}(\mathbf{X}))^2 \right]. \tag{3}$$

This provides consistent estimates that minimize the target residual error asymptotically:

$$\begin{aligned} &\lim_{n\to\infty} \min_{a,\mathbf{b}} \mathbb{E}_{f_{\text{src}}^{(n)}(x)\tilde{f}(y|x)} \left[\frac{f_{\text{trg}}(\mathbf{X})}{f_{\text{src}}(X)} (Y - \hat{y}_{a,\mathbf{b}}(\mathbf{X}))^2 \right] \\ &= \min_{a,\mathbf{b}} \mathbb{E}_{f_{\text{trg}}(x)f(y|x)} \left[(Y - \hat{y}_{a,\mathbf{b}}(\mathbf{X})) \right]. \end{aligned} \tag{4}$$

However, finite sample generalization bounds require the importance ratio's first moment to be finite [12], $\mathbb{E}_{f_{\text{trg}}(\mathbf{x})} \left[f_{\text{trg}}(\mathbf{X})/f_{\text{src}}(\mathbf{X}) \right] < \infty$. Unfortunately, many biased source distributions violate this requirement (e.g., when the target distribution has larger variance than the source distribution). Estimates that result from reweighting under those distributions typically have high variance, as a small number of source datapoints with large importance weights, $f_{\text{trg}}(\mathbf{x})/f_{\text{src}}(\mathbf{x})$, determine the regression model's parameters.

2.3 Adaptive importance weighted regression

In practice, the adaptive importance weighted least squares (AIWLS) method, which is a slightly stabilized variant of IWLS, is often preferable [31]. Following Eq. (2), the model maximizes the weighted likelihood:

$$\operatorname*{argmax}_{a,\mathbf{b},\sigma} \mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\tilde{f}(y|\mathbf{x})} \left[\left(\frac{f_{\text{trg}}(\mathbf{X})}{f_{\text{src}}(\mathbf{X})} \right)^{\gamma} \log \hat{f}_{a,\mathbf{b},\sigma}(Y|\mathbf{X}) \right], \tag{5}$$

where $\gamma \in (0, 1)$ is the flattening parameter. It balances the estimator's consistency and stability, with ordinary least squares ($\gamma = 0$), and importance weighted least squares ($\gamma = 1$) at its extremes.

2.4 Robust minimax learning

Minimax robust estimation [33, 20] prescribes the predictor that minimizes the worst-case prediction loss [2, 35]. When the logarithmic loss (log loss), $\mathbb{E}_{f((\mathbf{x})f(y|\mathbf{x})} \left[-\log \hat{f}(y|\mathbf{x}) \right]$, is employed as the loss function, the minimax robust estimation approach reduces to the principle of maximum entropy [23].

$$\max_{\hat{f}(y|\mathbf{x})} H(Y|\mathbf{X}) \triangleq \mathbb{E}_{f(\mathbf{x})\hat{f}(y|\mathbf{x})}[-\log f(Y|\mathbf{X})] \tag{6}$$

such that: $\mathbb{E}_{\tilde{f}(\mathbf{x})\hat{f}(y|\mathbf{x})}[\Phi(\mathbf{X}, Y)] = \mathbb{E}_{\tilde{f}(\mathbf{x})\tilde{f}(y|\mathbf{x})}[\Phi(\mathbf{X}, Y)]$.

This principle is a foundational method for deriving many familiar exponential family distributions (e.g., Gaussian, exponential, Laplacian, logistic regression, conditional random fields [24]) by incorporating constraints of various known statistics, $\Phi(\cdot)$ [34]. The OLS model, $\hat{y} = a + \mathbf{b}^T\mathbf{x}$, results from robustly minimizing the log-loss in the independent and identically distributed (IID) setting, subject to matching quadratic interaction features: $\Phi(\mathbf{x}, y) = \text{vector}([y\ \mathbf{x}^T\ 1]^T[y\ \mathbf{x}^T\ 1])$.

A recently developed approach for learning under sample selection bias investigates learning probabilistic classifiers when the entropy measure of Eq. (6) is evaluated according to the target distribution, $f_{\text{trg}}(\mathbf{x})$, and the constraints of the maximum entropy optimization are expectations over the source distribution [26]. We build on this approach by extending it to predict continuous-valued variables using a relative entropy (Kullback-Leibler divergence) formulation that avoids distribution degeneracies in this paper.

3 Robust Bias-Aware Regression

3.1 Minimax estimation formulation

A natural loss function to consider is the conditional logloss on the target distribution,

$$\text{logloss}_{f_{\text{trg}}(\mathbf{x})}(f(y|\mathbf{x}), \hat{f}(y|\mathbf{x})) \triangleq \mathbb{E}_{f_{\text{trg}}(\mathbf{x})f(y|\mathbf{x})}[-\log \hat{f}(Y|\mathbf{X})].$$

This conditional logloss measures the amount of expected "surprise" (in bits) to see samples from $f_{\text{trg}}(x)f(y|\mathbf{x})$ when samples are assumed to come from $f_{\text{trg}}(x)\hat{f}(y|\mathbf{x})$ instead [14]. As described in §2.4, when the source and target distribution match (or their differences are ignored), the minimax robust estimation approach with this loss function subject to quadratic interaction feature constraints yields the ordinary least squares solution to regression. Thus, using this loss function as a starting point can be viewed as a natural generalization of standard linear regression methods.

We define the difference in conditional logloss between an estimator $\hat{f}(y|\mathbf{x})$ and a baseline conditional distribution $f_0(y|\mathbf{x})$ on the target data distribution $f_{\text{trg}}(\mathbf{x})f(y|\mathbf{x})$ as the relative loss:

$$\begin{aligned} &\text{rel-loss}_{f_{\text{trg}}(\mathbf{x})}(f(y|\mathbf{x}), \hat{f}(y|\mathbf{x}), f_0(y|\mathbf{x})) \qquad (7)\\ &\triangleq \text{logloss}_{f_{\text{trg}}(\mathbf{x})}(f(y|\mathbf{x}),\hat{f}(y|\mathbf{x})) - \text{logloss}_{f_{\text{trg}}(\mathbf{x})}(f(y|\mathbf{x}),f_0(y|\mathbf{x}))\\ &= \mathbb{E}_{f_{\text{trg}}(\mathbf{x})f(y|\mathbf{x})}\left[-\log \frac{\hat{f}(Y|\mathbf{X})}{f_0(Y|\mathbf{X})}\right]. \end{aligned}$$

It measures the amount of expected relative "surprise" of data from $f_{\text{trg}}(x)f(y|\mathbf{x})$ assumed to come from $f_{\text{trg}}(x)\hat{f}(y|\mathbf{x})$ instead of $f_{\text{trg}}(x)f_0(y|\mathbf{x})$.

We consider the setting in which the conditional distribution $f(y|\mathbf{x})$ is known to satisfy certain statistical properties (denoted by set Ξ):

$$\Xi \triangleq \left\{f(y|\mathbf{x}) \mid \mathbb{E}_{f_{\text{src}}(\mathbf{x})f(y|\mathbf{x})}[\Phi(\mathbf{X}, Y)] = \mathbf{c}\right\}, \qquad (8)$$

where $\mathbf{c} = \frac{1}{n}\sum_{i=1}^n \phi(\mathbf{x}_i, y_i)$ is a vector of statistics measured from training data. We seek the regression model estimator $\hat{f}(y|\mathbf{x})$ that is most robust to the "most surprising" distribution that can arise from covariate shift (formally expressed by Definition 1).

Definition 1. *The* **robust bias-aware regression estimator**, *$\hat{f}(y|\mathbf{x})$, is the saddle point solution of the following minimax optimization:*

$$\min_{\hat{f}(y|\mathbf{x})} \max_{f(y|\mathbf{x})\in\Xi} \textit{rel-loss}_{f_{trg}(\mathbf{x})}(f(y|\mathbf{x}), \hat{f}(y|\mathbf{x}), f_0(y|\mathbf{x})).$$

This minimax optimization can be interpreted as a two-player game in which the estimator first chooses $\hat{f}(y|\mathbf{x})$ to minimize relative loss, and then the adversarial evaluation player chooses $f(y|\mathbf{x})$ to maximize relative loss. Note that both conditional probabilities are also constrained to the conditional probability simplex, denoted Δ: $\forall\mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}, f(y|\mathbf{x}) \geq 0;\quad \forall\mathbf{x} \in \mathcal{X}, \int_y f(y|\mathbf{x})\, dy = 1$.

Theorem 1. *The solution of the minimax relative logloss optimization (Def. 1) minimizes the target distribution conditional Kullback-Leibler divergence, $D_{f_{trg}(\mathbf{x}),\hat{f}(y|\mathbf{x})}\left(\hat{f}(Y|\mathbf{X})||f_o(Y|\mathbf{X})\right) \triangleq \mathbb{E}_{f_{trg}(\mathbf{X})\hat{f}(Y|\mathbf{X})}\left[\log \frac{\hat{f}(Y|\mathbf{X})}{f_0(Y|\mathbf{X})}\right]$, subject to matching statistics, $\mathbf{c}$, on the source distribution:*

$$\min_{\hat{f}(y|\mathbf{x})\in\Delta} D_{f_{trg}(\mathbf{x}),\hat{f}(y|\mathbf{x})}\left(\hat{f}(Y|\mathbf{X})||f_o(Y|\mathbf{X})\right) \qquad (9)$$
$$\textit{such that: } \mathbb{E}_{f_{src}(\mathbf{x})\hat{f}(y|\mathbf{x})}[\Phi(\mathbf{X}, Y)] = \mathbf{c}.$$

We note that the objective function of this optimization is convex and the constraints are each affine. Thus, standard tools from convex optimization can be employed to obtain the solution to the constrained optimization. We establish the parametric solution to the optimization problem in Theorem 2.

Theorem 2. *The robust bias-aware regression for target distribution $f_{trg}(\mathbf{x})$ estimated using constraints from the source distribution $f_{src}(\mathbf{x})$ takes the form:*

$$\hat{f}_\theta(y|\mathbf{x}) \propto f_o(y|\mathbf{x})e^{-\frac{f_{src}(\mathbf{x})}{f_{trg}(\mathbf{x})}\theta^T\Phi(\mathbf{x},y)}, \qquad (10)$$

with parameters obtained via target distribution maximum conditional log likelihood estimation:

$$\theta = \arg\max_\theta \mathbb{E}_{f_{trg}(\mathbf{x})f(y|\mathbf{x})}\left[\log \hat{f}_\theta(Y|X)\right]. \qquad (11)$$

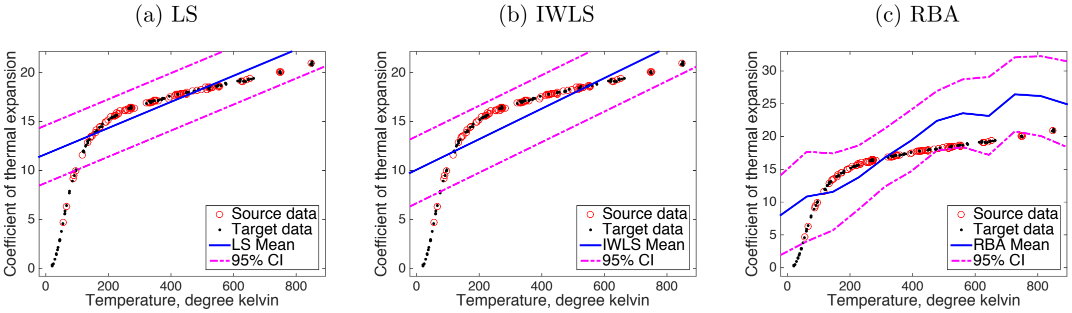
(a) LS (b) IWLS (c) RBA

Figure 2: The conditional mean (solid blue line) and 95% confidence interval (CI, dashed dot magenta line) of least squares linear regression (LS), importance weighted least squares (IWLS) and robust bias-aware regression (RBA) via KL-divergence learned via 90 biased source samples (red cross) and evaluated on 118 target datapoints (black point) of the Hahn1 dataset.

The distribution's certainty is moderated by the density ratio, $f_{\text{src}}(\mathbf{x})/f_{\text{trg}}(\mathbf{x})$. Note that it is the inverse of the importance weight, $f_{\text{trg}}(\mathbf{x})/f_{\text{src}}(\mathbf{x})$ from Eq. (3). As the density ratio $f_{\text{src}}(\mathbf{x})/f_{\text{trg}}(\mathbf{x})$ goes to zero, the estimator $\hat{f}(y|\mathbf{x})$ converges to the baseline conditional probability $f_0(y|\mathbf{x})$. As the ratio goes towards infinity, the estimator converges to a deterministic point estimate. The behavior of the robust approach is shown in Figure 2. Where source data is sparse, its uncertainty increases substantially to provide a more uncertain—and in this case, better—fit to the non-linearity of the underlying dataset. In contrast, least squares (LS) and importance-weighted least squares (IWLS) estimate their extrapolative uncertainty based only on the (reweighted) source data, yielding over-confident predictions for the target distribution.

3.2 Parameter Estimation

Optimizing Eq. (11) appears to be difficult because samples from $f_{\text{trg}}(\mathbf{x})f(y|\mathbf{x})$ are unavailable to estimate the target distribution log likelihood. However, the form of $f_\theta(y|\mathbf{x})$ prescribed by the robust bias-aware approach (Theorem 2), enables the gradient of the target likelihood function to be computed efficiently using the source distribution (Theorem 3).

Theorem 3. *The gradient of the conditional log likelihood estimation takes the following form:*

$$\nabla_\theta \mathbb{E}_{f_{trg}(\mathbf{x})f(y|\mathbf{x})}\left[\log \hat{f}_\theta(Y|X)\right] \\ = \mathbb{E}_{f_{src}(\mathbf{x})\hat{f}_\theta(y|\mathbf{x})}\left[\Phi(X,Y)\right] - \mathbf{c}. \tag{12}$$

Unfortunately, computing the needed expectations can be difficult when arbitrary feature functions $\Phi(\cdot,\cdot)$, are employed. By restricting consideration to quadratic feature functions and conjugate baseline conditional distributions $f_0(y|\mathbf{x})$, a distribution $\hat{f}(y|\mathbf{x})$ with tractable normalization and expectation computations is obtained (Corollary 1). As noted in §2.4, this set of quadratic features is also the basis for standard least squares linear regression under a robust logarithmic loss formulation. Thus, the representational power in terms of features is equivalent to OLS and IWLS variants (§2.2).

Corollary 1. *If the base distribution is conditional Gaussian, $f_o(y|\mathbf{x}) = N(\mu_o, \sigma_o^2)$, and the feature function has a quadratic form, $\Phi(\mathbf{x}, y) = [y\ \mathbf{x}^T\ 1]^T[y\ \mathbf{x}^T\ 1]$ (where θ is the vectorized matrix $\mathbf{M}$):*

$$\theta^T vector(\Phi(\mathbf{x},y)) = \underbrace{\mathbf{M}\cdot\begin{bmatrix}y\\ \mathbf{x}\\ 1\end{bmatrix}\begin{bmatrix}y\\ \mathbf{x}\\ 1\end{bmatrix}^T}_{matrix\ dot\ product} = \begin{bmatrix}y\\ \mathbf{x}\\ 1\end{bmatrix}^T \mathbf{M}\begin{bmatrix}y\\ \mathbf{x}\\ 1\end{bmatrix},$$

then the robust bias-aware regression is also a conditional Gaussian distribution:

$$\hat{f}_{\mathbf{M}}(y|\mathbf{x}) \sim N(\mu(\mathbf{x},\mathbf{M}), \sigma^2(\mathbf{x},\mathbf{M})), \tag{13}$$

$$\textit{where: } \mathbf{M} = \begin{bmatrix}\mathbf{M}_{(y,y)} & \mathbf{M}_{(y,\mathbf{x}1)}\\ \mathbf{M}_{(\mathbf{x}1,y)} & \mathbf{M}_{(1,1)}\end{bmatrix}$$

$$\mu(\mathbf{x},\mathbf{M}) = \left(2\frac{f_{src}(\mathbf{x})}{f_{trg}(\mathbf{x})}\mathbf{M}_{(y,y)} + \frac{1}{\sigma_o^2}\right)^{-1} \\ \left(-2\frac{f_{src}(\mathbf{x})}{f_{trg}(\mathbf{x})}\mathbf{M}_{(y,\mathbf{x}1)}\begin{bmatrix}\mathbf{x}\\ 1\end{bmatrix} + \frac{1}{\sigma_o^2}\mu_o\right) \tag{14}$$

$$\sigma^2(\mathbf{x},\mathbf{M}) = \left(2\frac{f_{src}(\mathbf{x})}{f_{trg}(\mathbf{x})}\mathbf{M}_{(y,y)} + \frac{1}{\sigma_o^2}\right)^{-1}. \tag{15}$$

Correspondingly, by Theorem 2, model parameters are selected by maximizing the target log likelihood:

$$\mathbf{M} = \underset{\mathbf{M}}{\operatorname{argmax}}\, \mathbb{E}_{f_{\text{trg}}(\mathbf{x})f(y|\mathbf{x})}\left[\log \hat{f}_{\mathbf{M}}(Y|X)\right], \tag{16}$$

$$\text{with gradient: } \mathbb{E}_{f_{\text{src}}(\mathbf{x})\hat{f}_{\mathbf{M}}(y|\mathbf{x})}\left[\Phi(X,Y)\right] - \mathbf{c}. \tag{17}$$

Base Distribution The base distribution plays an important role in our robust regression approach. A simple and straightforward way to set up the base distribution is to assume that it is a Gaussian distribution $N(\mu_o, {\sigma_o}^2)$ with mean and variance estimated from the range $[y_{\min}, y_{\max}]$ of y's of the source dataset D_{src}:

$$\mu_o = \frac{y_{\min} + y_{\max}}{2}, \quad \sigma_o^2 = \left(\frac{y_{\max} - \mu_o}{2}\right)^2. \tag{18}$$

Hence all of the y's of the source dataset are located within the 95% confidence of the base distribution.

Optimization and Regularization Due to the convexity [10] of the robust formulation (Theorem 1), convergence to a global optimum is guaranteed using standard gradient-based methods. We solve the optimization problem with the quadratic feature function defined in Corollary 1 by its dual form (Eq. (16)). Computing the gradient (Eq. (17)) requires taking the expectation over the source distribution, which is challenging. Instead, we use the empirical expectation over the source dataset, which approximates the real expectation in the constraints:

$$\mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\hat{f}_{\mathbf{M}}(y|\mathbf{x})}\left[\Phi(X,Y)\right] \approx \mathbb{E}_{f_{\text{src}}\hat{f}_{\mathbf{M}}(y|\mathbf{x}_i)}\left[\Phi(\mathbf{x}_i,Y)\right].$$

Additionally, only estimates of source distribution statistics, $\tilde{\mathbf{c}} \triangleq \mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\tilde{f}(y|\mathbf{x})}\left[\Phi(X,Y)\right] = \frac{1}{N}\sum_{i=1}\Phi(\mathbf{x}_i, y_i)$, from sample data are used to constrain the estimated conditional probability distribution, $\hat{f}(y|\mathbf{x})$, rather than exact source distribution statistics, $\mathbf{c} = \mathbb{E}_{f_{\text{src}}(\mathbf{x})f(y|\mathbf{x})}\left[\Phi(X,Y)\right]$. This introduces finite sample error that should be accounted for in the parameter estimation. One way to accomplish this is by relaxing the source distribution constraints to incorporate some slack, ϵ: $||\mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\hat{f}(y|\mathbf{x})}[\Phi(X,Y)] - \tilde{\mathbf{c}}|| \le \epsilon$. Primal relaxed constraints correspond with regularization of the conditional log likelihood maximization in the dual [1]. Following Theorem 3, the gradient under ℓ_2-regularization is:

$$\begin{aligned} &\mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\hat{f}(y|\mathbf{x})}[\Phi(X,Y)] - \tilde{\mathbf{c}} - \lambda\mathbf{M} \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\hat{f}_{\mathbf{M}}(y|\mathbf{x}_i)}\left[\Phi(\mathbf{x}_i,Y)\right] - \tilde{\mathbf{c}} - \lambda\mathbf{M}, \end{aligned} \tag{19}$$

where λ is the regularization weight.

Algorithm 1 shows the batch gradient descent method for robust bias-aware regression.

Algorithm 1 Batch gradient descent method for robust bias-aware regression

```
Input: source dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, source/target distributions $f_{\text{src}}(\mathbf{x})$; $f_{\text{trg}}(\mathbf{x})$, feature function $\Phi(\mathbf{x}, y)$, statistics $\tilde{\mathbf{c}}$, learning rate $\gamma$, convergence threshold $\tau$ and regularization weight $\lambda$.
Output: Model parameters $\mathbf{M}$
  initialize $\mathbf{M}$
  repeat
    for each source data example $i$ do
      $\mu(\mathbf{x}_i, \mathbf{M}) \leftarrow \left(2\frac{f_{\text{src}}(\mathbf{x}_i)}{f_{\text{trg}}(\mathbf{x}_i)}\mathbf{M}_{(y,y)} + \frac{1}{\sigma_o^2}\right)^{-1} \left(-2\frac{f_{\text{src}}(\mathbf{x}_i)}{f_{\text{trg}}(\mathbf{x}_i)}\mathbf{M}_{(y,\mathbf{x}1)}\begin{bmatrix}\mathbf{x}_i \\ 1\end{bmatrix} + \frac{1}{\sigma_o^2}\mu_o\right)$
      $\sigma^2(\mathbf{x}_i, \mathbf{M}) \leftarrow \left(2\frac{f_{\text{src}}(\mathbf{x}_i)}{f_{\text{trg}}(\mathbf{x}_i)}\mathbf{M}_{(y,y)} + \frac{1}{\sigma_o^2}\right)^{-1}$
    end for
    $\nabla\mathcal{L} \leftarrow \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\hat{f}_{\mathbf{M}}(y|\mathbf{x}_i)}\left[\Phi(\mathbf{x}_i, Y)\right] - \tilde{\mathbf{c}} - \lambda\mathbf{M}$
    $\mathbf{M} \leftarrow \mathbf{M} - \gamma\nabla\mathcal{L}$
  until $||\nabla\mathcal{L}|| \le \tau$
  return $\mathbf{M}$
```

3.3 Relation with other methods

Equivalence with Importance Weighting Robust-bias aware regression (RBA) is a general framework that reduces to maximizing the relative target conditional entropy under certain constraints. Due to a lack of target data, constraints are usually chosen to match feature expectations under the source distribution. However, it could be helpful to manipulate the constraints and incorporate feature matching under a different distribution based on some side information, corresponding to different assumptions. We assume the statistics provided by expert knowledge or computed under a generalized distribution are $\mathbf{c}'$. When we incorporate the strong assumption that the feature expectation under the target distribution is equivalent to the expectation of reweighted features on source data, RBA is equivalent to IWLS (Theorem 4).

Theorem 4. *When a target-distribution-based constraint,* $\mathbb{E}_{f_{trg}(\mathbf{x})\hat{f}(y|\mathbf{x})}[\Phi(X,Y)] = \tilde{\mathbf{c}}' \triangleq \mathbb{E}_{\tilde{f}_{src}(\mathbf{x})\tilde{f}(y|\mathbf{x})}\left[\frac{f_{trg}(X)}{f_{src}(X)}\Phi(X,Y)\right]$ *is applied, and the feature function takes the quadratic form of Corollary 1, RBA regression is equivalent to IWLS regression.*

This indicates that IWLS is a special case under the general robust bias-aware regression framework.

Difference from Bayesian linear regression

Figure 2(c) may appear similar to a Bayesian treatment of the regression problem (e.g., a Gaussian process). We establish a key difference between our approach and the comparable Bayesian technique, Bayesian linear regression model (BLR), here. By letting $\theta^T = [\mathbf{b}^T a]$ and $\hat{y}_\theta(\mathbf{x}) = [\mathbf{x}^T 1]\theta$, the linear regression model becomes:

$$f(y|\mathbf{x}) = \hat{y}_\theta(\mathbf{x}) + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2). \tag{20}$$

If we treat σ^2 as a known constant, BLR [7] assumes a prior distribution for θ. The multivariate Gaussian conjugate prior, $\theta \sim N(\mu, \Sigma)$, is commonly applied. Given source data $(\mathbf{x}_i, y_i)_{i=1}^N$, the posterior distribution of θ is inferred by Bayes' rule:

$$f(\theta|\mathbf{X}, \mathbf{y}) = N(\mathbf{A}^{-1}(\Sigma^{-1}\mu + \sigma^{-2}\mathbf{X}^T\mathbf{y}), \mathbf{A}^{-1}),$$

where $\mathbf{X}$ is a $N \times (d+1)$ design matrix with each row $[\mathbf{x}_i^T 1], \mathbf{y} = [y_1, y_2, ..., y_n]^T$ and $\mathbf{A} = \sigma^{-2}\mathbf{X}^T\mathbf{X} + \Sigma^{-1}$. The predictive distribution $f(y|\mathbf{x}_t, \mathbf{X}, \mathbf{y})$ for target datapoint $\mathbf{x}_t$ is given by averaging the output of all possible linear models with respect to the posterior, yielding a conditional Gaussian distribution $N(\mu(\mathbf{x}_t), \Sigma(\mathbf{x}_t))$:

$$\mu(\mathbf{x}_t) = [\mathbf{x}_t^T 1]\mathbf{A}^{-1}(\Sigma^{-1}\mu + \sigma^{-2}\mathbf{X}^T\mathbf{y}) \qquad (21)$$

$$\Sigma(\mathbf{x}_t) = [\mathbf{x}_t^T 1]\mathbf{A}^{-1}[\mathbf{x}_t^T 1]^T + \sigma^2 \qquad (22)$$

Thus, as the amount of source data increases, BLR converges to the OLS regression model [7], minimizing source distribution squared loss rather than target distribution squared loss. In contrast, RBA provides nonlinear prediction means that robustly minimize target logloss with uncertainty in its distribution that is moderated by the density ratio.

Difference from RBA via differential entropy ($\mathbf{RBA_{DE}}$) A straightforward extension of RBA for continuous-valued variables is to build RBA via conditional differential entropy $H(Y|\mathbf{X})$. Following the similar procedure built for RBA via KL-divergence in section §3, the resulting conditional Gaussian distribution of RBA_{DE} has the following form:

$$\mu(\mathbf{x}, \mathbf{M}) = -(\mathbf{M}_{(y,y)})^{-1}(\mathbf{M}_{(y,\mathbf{x}1)}\begin{bmatrix}\mathbf{x}\\1\end{bmatrix}) \qquad (23)$$

$$\sigma^2(\mathbf{x}, \mathbf{M}) = \left(2\frac{f_{\text{src}}(\mathbf{x})}{f_{\text{trg}}(\mathbf{x})}\mathbf{M}_{(y,y)}\right)^{-1}. \qquad (24)$$

We note that when the density ration $f_{\text{src}}(\mathbf{x})/\mathbf{f}_{\text{trg}}(\mathbf{x})$ goes to zero, the predictor will give very large logloss due to predictions produced that have very large uncertainty (variance). In contrast, our proposed RBA via KL-divergence (RBA_{KLD}) converges to the base distribution $f_o(y|\mathbf{x})$. A base distribution provides an upper bound of logloss of RBA_{KLD}.

4 Experiments

4.1 Datasets

We employ publicly available regression datasets from the UCI repository [3] to evaluate our approach. The number of examples and features, and a basic description of the output of each dataset are listed in Table 1. We refer to Appendix B for more detailed information on each dataset.

Table 1: Datasets for empirical evaluation

Dataset	#Examples	#Features	Output
Airfoil	1503	5	sound pressure
Concrete	1030	8	strength
Housing	506	14	value of home
Music	1059	66	latitude
Crime	1994	127	crime rate
Parkinsons	5725	16	UPDRS score
WineQuality	6497	11	quality score
IndoorLocation	21048	529	latitude

Table 2: Experimental settings

Dataset	#Source	#Target	Bias Setting
Airfoil	150-751	752	synthetic
Concrete	100-515	515	synthetic
Housing	75-253	253	synthetic
Music	160-529	530	synthetic
Parkinsons	1430-2862	2863	synthetic
Crime	40-278	1716-1954	different state
WineQuality	4898	1599	different color
Parkinsons	1877	1839	different age
IndoorLocation	9371	10566	different floor

4.2 Constructing dataset with bias

We consider experimental settings with both synthetically created bias and naturally occurring bias. The amounts of source and target data, and the type of bias for each experiment are listed in Table 2.

Synthetically biased data: We first evaluate our approach on datasets with artificially created bias between the source and target distributions. This allows us to show the generalizability of the proposed method on a larger number of controlled experiments. Given a dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^n$, we create biased source D_{src} and target D_{trg} datasets using the following sampling procedure:

1. Split D randomly and evenly into two disjoint datasets D_1 and D_2.
2. Compute the sample mean $\bar{\mathbf{x}}$ and sample covariance Q of D_1.
3. Construct a Gaussian distribution $N(\bar{\mathbf{x}}, Q)$ and sample a data $\mathbf{x}_{seed}$ to construct another Gaussian distribution $N(\mathbf{x}_{seed}, \alpha Q)$ with bias weight $\alpha \in [0, 1]$. We employ $\alpha = 0.3$ in our experiments.
4. Sample from D_1 in proportion to $N(\mathbf{x}_{seed}, \alpha Q)$ to get the biased source dataset D_{src}.
5. Let D_2 be the target dataset D_{trg}.

Our sampling procedure is motivated by the purpose of learning from a biased source dataset and comparing prediction performance on an unbiased target dataset.

Naturally Biased Data: To highlight the benefits of our method in practice, we also conduct experiments under naturally occurring bias. The crime dataset is separated into pairs of source and target data where the source only contains data for a single state and the target contains data for all other states combined. We average the results of training with each state as the source and the other states as the target. The Parkinsons dataset is separated by different age ranges. The source data includes all samples of subjects whose age is below 59 while the target dataset holds samples with subjects ranging in age from 60 to 69. For the wine quality dataset, we train our model on white wine samples as the source dataset, and then use red wine samples as the target dataset. Finally, we evaluate our approach on a large, high-dimensional dataset: the IndoorLocation dataset. We use data collected from low floors as source data and data collected from high floors as target data.

4.3 Density Estimation

Multivariate Gaussian Kernel density estimation [7] is a popular density estimation method that converges to the true probability density of samples asymptotically. Unfortunately it suffers from the curse of dimensionality and is not reliable in high-dimensional problems. Importance weighted methods have been proposed where logistic regression is used to directly form the estimations [31]. By applying Bayes theorem, the importance ratio can be written as:

$$\frac{f_{\text{trg}}(\mathbf{x})}{f_{\text{src}}(\mathbf{x})} = \frac{n_{src}}{n_{trg}} \exp(\hat{\omega}^T \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}),$$

where ω is selected to maximize the (regularized) likelihood of source and target datapoints, which need not be labeled. We refer the reader to previous work [28, 11, 6, 31] for a full description of the derivation for this approach.

4.4 Comparison Approaches

We compare our approach, robust-bias aware regression via KL-divergence (**RBA$_{\mathbf{KLD}}$**), with six other regression methods, which each produce (conditional) Gaussian estimates. We evaluate the empirical logloss of each model on target data D_{trg}: $-\frac{1}{n_{trg}} \sum_{i=1}^{n_{trg}} \log f(y_i|\mu(\mathbf{x}_i), \sigma^2(\mathbf{x}_i))$. The methods are:

The baseline (**BS**) Gaussian distribution $Y \sim N(\mu_o, \sigma_o^2)$ is independent of input $\mathbf{x}$ with mean and variance estimated from source data (18).

Robust bias-aware regression using differential entropy (**RBA$_{\mathbf{DE}}$**) yields the conditional Gaussian distribution $Y|X \sim N(\mu(\mathbf{x}, \mathbf{M}), \sigma^2(\mathbf{x}, \mathbf{M}))$ with mean and variance defined by Equation (23) and (24) respectively, and $\mathbf{M}$ estimated via, e.g., gradient descent.

Bayesian linear regression (**BLR**) provides the conditional Gaussian distribution $Y|X \sim N(\mu(\mathbf{x}), \Sigma(\mathbf{x}))$ with mean and variance specified by Equation (21) and (22) respectively. In practice, it is challenging to get the prior information for a Bayesian approach. Here we assume the prior distribution of θ follows a Gaussian distribution $N(\mathbf{0}, \mathbf{I})$ where $\mathbf{I}$ is the identity matrix. The variance σ^2 of noise ϵ (Equation (20)) is the same as the optimal estimator of variance of LS.

LS, IWLS and **BAIWLS** share the common formulation of the conditional Gaussian distribution $Y|X \sim N(\mathbf{b}^T\mathbf{x} + a, \sigma^2)$. The linear parameters $(\mathbf{b}, a)$ of LS are estimated by equation (2) that is to maximize the log-likelihood of source data which also gives the optimal estimator of σ^2 [7]. Being different from LS, the linear parameters $(\mathbf{b}, a)$ and the variance σ^2 of adaptive importance weighted least squares (AIWLS) are estimated by maximizing adaptive reweighted log-likelihood from Eq. (5), which is equivalent to minimizing adaptive reweighted residual error (Equation (5)) when estimating $(\mathbf{b}, a)$. The argument is very similar to the case of LS. The linear parameters $(\mathbf{b}, a)$ and the variance σ^2 of IWLS are given by Equation (5) with $\gamma = 1$. For BAIWLS, we consider the minimum empirical logloss over target dataset D_{trg} achieved by choosing the optimal flattening parameter γ from $\Upsilon = \{0.1, 0.2, ..., 0.8, 0.9\}$.

4.5 Results

As shown in Figure 3, RBA$_{\text{KLD}}$ has the smallest average empirical log loss, and it provides better performance than the other methods on almost all experiment settings. The result varies as the amount of synthetic bias source data increases, which highlights the motivation of RBA$_{\text{KLD}}$. When there is only a small amount of source data that is significantly biased from the target data, the performance of LS, IWLS and BAIWLS are much worse than RBA$_{\text{KLD}}$ of which the performance guided by baseline is still sound. When more of the source data is given, and the covariate shift still exists, RBA$_{\text{KLD}}$ makes significant improvement from the baseline and still performs better than nearly all other methods even if they have better performance than the baseline in some cases. When the amount of source data gets close to the target so that the degree of covariate shift is small, the performance of RBA$_{\text{KLD}}$ converges to the performance of the other linear regression methods.

As for BAIWLS, we apply a whole range of flattening parameter γ $(0.1, 0.2, ..., 0.8, 0.9)$ in each experiment and choose the one which gives the minimum empiri-

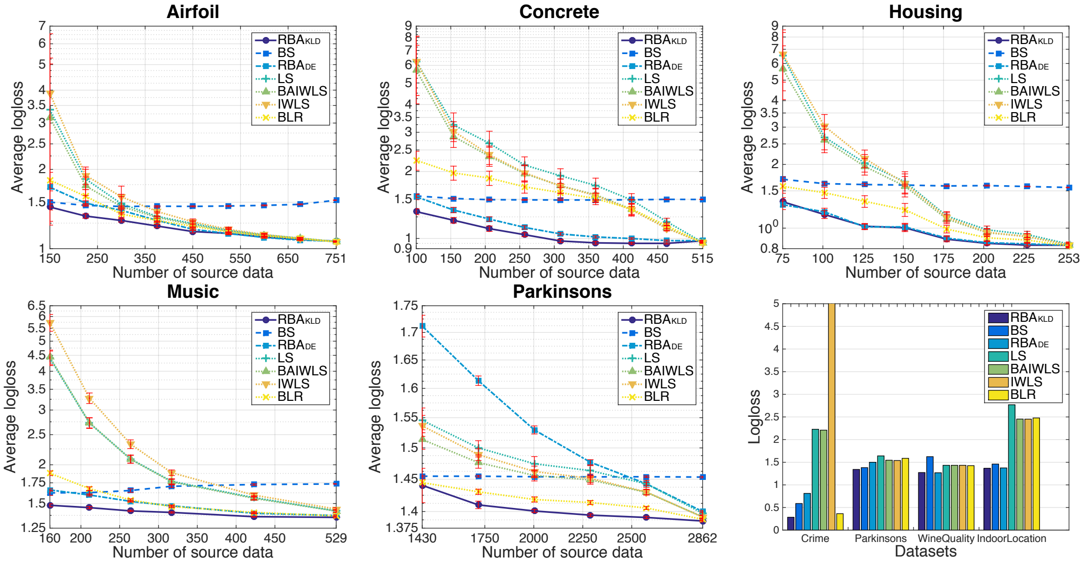

Figure 3: Five plots of the average empirical logloss of seven methods for target datasets with 95% confidence interval with amounts of source data. A bar figure showing empirical log loss on four natural bias datasets.

cal logloss each time. Even though this unrealistically generous choice of the flattening parameter is given to BAIWLS, it turns out that the flattening parameter does not improve much with respect to LS and IWLS.

BLR has better performance than LS, BAIWLS and IWLS. The reason may be that the prior assumption in BLR compensates for the biased results from learning the source. But, as discussed in §3.3, BLR converges to LS as the amount of source data increases. Shown in the bar figure of Figure 3, the performance of BLR in the experimental settings with natural bias and large number of source data is as worse as LS.

RBA_{DE} has competitive performance in some of the experimental settings. But, it has the worst performance in the experimental setting of Parkinsons with synthetic bias, and it has even worse performance than BS in the experimental settings of Crime and Parkinsons with natural bias. As discussed in §3.3, the reason may be that RBA_{DE} will lead to very large log loss in wide support if the target data is very biased from the source one that the density ratio in (24) goes to zero.

5 Conclusion and Future Work

We proposed a novel minimax approach for regression problems under covariate shift. The minimax approach minimizes relative loss in the worst case, and reduces to minimizing conditional Kullback-Leibler divergence. Due to continuous-valued prediction variables in the regression case, we restrict our constraint to quadratic feature functions and use conditional Gaussian baseline distributions, leading to a conditional Gaussian regression model. We compared the proposed robust method with a range of existing regression models on both synthetically created and natural bias experimental settings of a range of real regression datasets including large size and high dimensional datasets to show its benefit under covariate shift.

A number of important extensions remain as future work. One important generalization is to expand the applicability of our approach to settings where the regression model's output is multivariate. Motivated by benefits observed in the classification setting [25], active learning for regression using this model as the basis for both prediction and label solicitation strategies is also an interesting avenue of future research. We plan to conduct experiments in each of these areas in the future.

Acknowledgments

The research in this paper was supported in part by NSF grant RI-#1526379 and in part by a grant from the Future of Life Institute.

References

[1] Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *Learning Theory*, pages 139–153. Springer Berlin Heidelberg, 2006.

[2] Kaiser Asif, Wei Xing, Sima Behpour, and Brian D. Ziebart. Adversarial cost-sensitive classification. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2015.

[3] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[4] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 19:137, 2007.

[5] Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010.

[6] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the International Conference on Machine Learning*, pages 81–88. ACM, 2007.

[7] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer New York, 2006.

[8] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 129–136, 2008.

[9] John Blitzer, Sham Kakade, and Dean P Foster. Domain adaptation with coupled subspaces. In *International Conference on Artificial Intelligence and Statistics*, pages 173–181, 2011.

[10] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[11] Kuang Fu Cheng and Chih-Kang Chu. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4):583–604, 2004.

[12] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, pages 442–450, 2010.

[13] Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

[14] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley and Sons, 2006.

[15] Hal Daumé, III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006.

[16] Miroslav Dudík, Steven J Phillips, and Robert E Schapire. Correcting sample selection bias in maximum entropy density estimation. In *Advances in neural information processing systems*, pages 323–330, 2005.

[17] Wei Fan, Ian Davidson, Bianca Zadrozny, and Philip S. Yu. An improved categorization of classifier's sensitivity on sample selection bias. In *Proc. of the International Conference on Data Mining*, pages 4–pp, 2005.

[18] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning.* Springer Series in Statistics New York, 2001.

[19] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the International Conference on Machine Learning*, pages 513–520, 2011.

[20] Peter D. Grünwald and A. Phillip Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32:1367–1433, 2004.

[21] T. Hahn. Copper thermal expansion study. *NIST*, 1979. Unpublished.

[22] Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schlkopf. Correcting sample selection bias by unlabeled data. In *Neural Information Processing Systems*, pages 601–608, 2006.

[23] Edwin T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982.

[24] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the International Conference on Machine Learning*, pages 282–289, 2001.

[25] Anqi Liu, Lev Reyzin, and Brian D. Ziebart. Shift-pessimistic active learning using robust bias-aware prediction. In *AAAI Conference on Artificial Intelligence*, pages 2764–2770, 2015.

[26] Anqi Liu and Brian D. Ziebart. Robust classification under sample selection bias. In *Advances in Neural Information Processing Systems*, pages 37–45, 2014.

[27] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective.* MIT Press, 2012.

[28] Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

[29] Sashank Jakkam Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[30] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, October 2000.

[31] Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation.* MIT Press, 2012.

[32] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V. Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, pages 1433–1440, 2008.

[33] Flemming Topsøe. Information theoretical optimization techniques. *Kybernetika*, 15(1):8–27, 1979.

[34] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2), 2008.

[35] Hong Wang, Wei Xing, Kaiser Asif, and Brian D. Ziebart. Adversarial prediction games for multivariate losses. In *Advances in Neural Information Processing Systems*, pages 2710–2718, 2015.

[36] Junfeng Wen, Chun-Nam Yu, and Russ Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *Proc. of the International Conference on Machine Learning*, pages 631–639, 2014.

[37] Yaoliang Yu and Csaba Szepesvári. Analysis of kernel mean matching under covariate shift. In *Proc. of the International Conference on Machine Learning*, pages 607–614, 2012.

[38] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proc. of the International Conference on Machine Learning*, pages 903–910, 2004.

[39] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *Proceedings of the International Conference on Machine Learning*, pages 819–827, 2013.