Robust Learning under Uncertain Test Distributions: Relating Covariate Shift to Model Misspecification

Junfeng Wen[1] JUNFENG.WEN@UALBERTA.CA
Chun-Nam Yu[2] CHUN-NAM.YU@ALCATEL-LUCENT.COM
Russell Greiner[1] RGREINER@UALBERTA.CA
[1]Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8 CANADA
[2]Bell Labs, Alcatel-Lucent, 600 Mountain Avenue, Murray Hill, NJ 07974 USA

Abstract

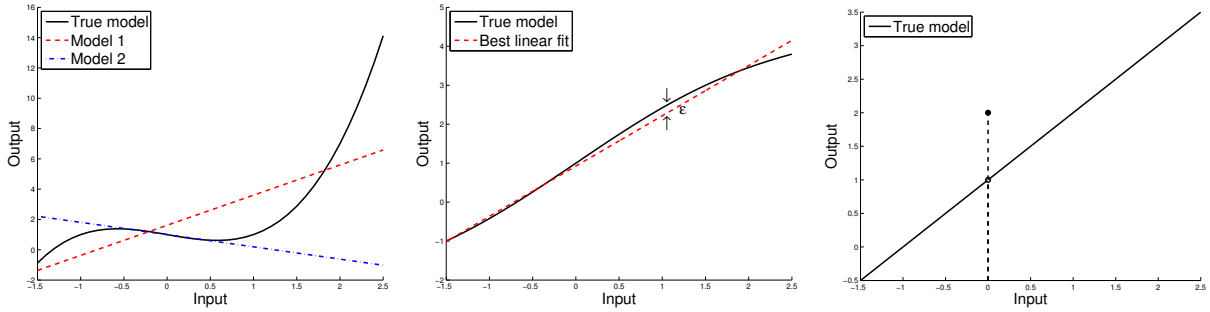
Many learning situations involve learning the conditional distribution $p(y|x)$ when the training instances are drawn from the training distribution $p_{tr}(x)$, even though it will later be used to predict for instances drawn from a different test distribution $p_{te}(x)$. Most current approaches focus on learning how to reweigh the training examples, to make them resemble the test distribution. However, reweighing does not always help, because (we show that) the test error also depends on the correctness of the underlying model class. This paper analyses this situation by viewing the problem of learning under changing distributions as a game between a learner and an adversary. We characterize when such reweighing is needed, and also provide an algorithm, robust covariate shift adjustment (RCSA), that provides relevant weights. Our empirical studies, on UCI datasets and a real-world cancer prognostic prediction dataset, show that our analysis applies, and that our RCSA works effectively.

1. Introduction

Traditional machine learning often explicitly or implicitly assumes that the data used for training a model come from the same distribution as that of the test data. However, this assumption is violated in many real-world applications. For example, biostatisticians often try to collect a large and diverse training set, perhaps for building prognostic predictors for patients with different diseases. When clinicians deploy these predictors, they do not know whether the local test patient population will be even close to that training population. Sometimes we can collect a small sample from the target test population, but in most cases we have nothing more than weak prior knowledge about how the test distribution may shift, such as anticipated changes in gender ratio or age distribution. It is useful to build predictors that are robust against such changes in test distributions.

In this work, we investigate the problem of distribution change under *covariate shift* assumption (Shimodaira, 2000), in which both training and test distributions share the same conditional distribution $p(y|x)$, while their marginal distributions, $p_{tr}(x)$ and $p_{te}(x)$, are different. To correct the shifted distribution, major efforts have been dedicated to importance reweighing (Quionero-Candela et al., 2009; Sugiyama & Kawanabe, 2012). However, reweighing methods will not necessarily improve the performance in test set, as prediction accuracy under covariate shift is also dependent on *model misspecification* (White, 1981). Fig. 1 shows three examples of misspecified models, where we are considering the model class of straight lines of the form $y = ax+b$, for $x \in [-1.5, 2.5]$. In Fig. 1(a), no straight line is a good fit for the cubic curve across the whole interval, but Model 2 fits the curve reasonably well in the small interval $[-0.5, 0.5]$. If training data is spread all over $[-1.5, 2.5]$ while test data concentrates on $[-0.5, 0.5]$, improvement via reweighing could be significant. The situation in Fig. 1(b) is different: although the true model is a curve and not a straight line, the best linear fit is no more than ϵ away from the value of the true model. In this case, no matter what test distributions we see in the interval $[-1.5, 2.5]$, the regression loss of the best linear model will never be more than ϵ from the Bayes optimal loss. In Fig. 1(c), the true model is a straight line except at $x = 0$; perhaps this outlier is a cancer patient whose tumour spontaneously disappeared on its own. Unless the test distribution concentrates most of its mass at $x = 0$, the straight line fit learned from the training data over the interval will still be a very good predictor. Sometimes we can rule out this type of covariate shift through prior knowledge. If such outliers are extremely rare during training time, we would not expect the test population to have many such patients. Reweighing will not help much in cases 1(b) and 1(c).

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.



(a) Large misspecification. (b) Small misspecification. (c) Single point misspecification.

Figure 1. Three different scenarios of model misspecifications.

In this paper, we relate covariate shift to model misspecification and investigate when reweighing can help a learner deal with covariate shift. We introduce a game between a learner and an adversary that performs robust learning. The learner chooses a model $\boldsymbol{\theta}$ from a set Θ to *minimize* the loss, while the adversary chooses a reweighing function $\boldsymbol{\alpha}$ from a set $\mathcal{A}$ to create new test distributions to *maximize* the loss. There are two major contributions in this paper: First, we provide an improved understanding of the relation between covariate shift and model misspecification through this game analysis. If the learner can find a $\boldsymbol{\theta}$ that minimizes the loss against any possible $\boldsymbol{\alpha}$ that the adversary can play, then it is not necessary to perform reweighing against covariate shift scenarios represented by $\mathcal{A}$. Second, we provide a systematic method for checking a model class Θ against different covariate shift scenarios, such as changing gender ratio and age distributions in the prognostic predictor example, to help user decide whether importance reweighing would be beneficial.

For practical use, our method can be used to decide if the model class is sufficient against shifts that are close to a test sample; or robust against a known range of potential shifts if test sample is unavailable. If the model class is insufficient, we can consider different ways to deal with covariate shifts, such as reweighing using unlabelled test samples, or exploring a different model class for the problem.

2. Related Work

Our work is inspired by Grünwald & Dawid (2004), who interpret maximum entropy as a game between an adversary and a learner on minimizing the worst case expected log loss. Teo et al. (2008) and Globerson & Roweis (2006) also consider an adversarial scenario under changing test set conditions, but they are concerned with corruption or deletion of features rather than covariate shift.

Many results on covariate shift correction involve density ratio estimation. Shimodaira (2000) showed that, given covariate shift and model misspecification, reweighing each instance with $p_{te}(x)/p_{tr}(x)$ is asymptotically optimal for log-likelihood estimation, where $p_{tr}(x)$ and $p_{te}(x)$ are assumed to be known or estimated in advance. Sugiyama & Müller (2005) extended this work by proposing an (almost) unbiased estimator for L_2 generalization error. There are several works focusing on minimizing different types of divergence between distributions in the literature (Kanamori et al., 2008; Sugiyama et al., 2008; Yamada et al., 2011). *Kernel mean matching* (KMM) (Huang et al., 2007) reweighs instances to match means in a RKHS (Schölkopf & Smola, 2002). Our work and some other approaches (Pan et al., 2009) adapt the idea of matching means of the datasets to correct shifted distribution, but we extend their approaches from a two-step optimization to a game framework that jointly learns a model and weights with covariate shift correction. Some other approaches (Zadrozny, 2004; Bickel et al., 2007; 2009; Storkey & Sugiyama, 2007) consider different generative models for special cases of covariate shift mechanisms.

Besides these approaches, there are many other works focusing on the theoretical analysis of statistical learning bounds for covariate shift. Ben-David et al. (2007) gave a bound on L_1 generalization error given the presence of mismatched distributions. Analyses on other forms of error were also introduced in the literature (Shimodaira, 2000; Sugiyama & Müller, 2005; Cortes et al., 2010). However, most of these analyses neglect the effect of model misspecification (White, 1981). Apart from Shimodaira (2000) who pointed out a link between covariate shift and model misspecification with some quantitative evidence, and Huang et al. (2007) who observed that simpler models tend to benefit more from density ratio correction, few have addressed the question of determining when reweighing helps versus when it is not needed. In this paper, we show the relationship between covariate shift and model misspecification.

3. Learning Under Uncertain Test Distributions as a Game

Suppose we are given an i.i.d. (independent and identically distributed) training sample $\{(x_1, y_1), \cdots, (x_n, y_n)\}$

drawn from a joint distribution $p_{tr}(x, y)$, and know that the test distribution $p_{te}(x, y)$ is the same as $p_{tr}(x, y)$. The most common and well-established method to learn a prediction function $f : \mathcal{X} \mapsto \mathcal{Y}$ is through solving the following empirical risk minimization (ERM) problem:

$$\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(f_{\boldsymbol{\theta}}(x_i), y_i) + \lambda\,\Omega(\boldsymbol{\theta}), \tag{1}$$

where the prediction function $f_{\boldsymbol{\theta}}(\cdot)$ is parametrized by a vector $\boldsymbol{\theta}$, $l(\cdot,\cdot)$ is a loss function, $\Omega(\cdot)$ is a regularizer on $\boldsymbol{\theta}$ to control overfitting and $\lambda \in \mathbb{R}$ is regularization parameter.

When there is covariate shift, the feature distribution $p_{te}(x)$ is different from $p_{tr}(x)$ but the conditional distribution $p(y|x)$ representing the classification/regression rule remains the same. In this scenario, one of the most common approach to correct for the effect of covariate shift is to reweigh the training instances in the ERM problem to reflect their true proportions on the test set:

$$\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} w(x_i)\, l(f_{\boldsymbol{\theta}}(x_i), y_i) + \lambda\,\Omega(\boldsymbol{\theta}), \tag{2}$$

where $w(x_i)$ is a reweighing function that approximates the density ratio $p_{te}(x_i)/p_{tr}(x_i)$. There are different methods for estimating the density ratio $w(x)$ using unlabelled test data (Quionero-Candela et al., 2009; Sugiyama & Kawanabe, 2012). This suggests viewing the learning problem as a two-step estimation problem, where the density ratio $w(x)$ is estimated first before estimating $\boldsymbol{\theta}$.

Interestingly, in econometrics the phenomenon of covariate shift has been used to detect model misspecification. White (1981) considered the problem of detecting model misspecification in non-linear least squares regression for models $y_i = f_{\boldsymbol{\theta}}(x_i)+\epsilon_i$, where ϵ_i is i.i.d. noise. He showed that under certain assumptions, when there is no misspecification, the objective and solution $\boldsymbol{\theta}^\star$ of the problem

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} (y_i - f_{\boldsymbol{\theta}}(x_i))^2$$

converge to the same limits as the reweighed problem:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} w_i (y_i - f_{\boldsymbol{\theta}}(x_i))^2,$$

for any fixed set of non-negative weights w_i such that $\sum_i w_i = 1$. He then derived several misspecification tests based on asymptotic approximation of the difference of the solutions. The key idea in his work is to detect model misspecification by creating his own covariate shift w_i, so that correct inference on the effects of different variables in regression can be performed.

In this paper, we explore White's main insight further by modelling the reweighing functions $w(x_i)$ as an adversary in a game against the learner. Instead of detecting model misspecification, we want to tell whether density ratio correction is needed under a set of potential distribution shifts. The rest of this section will introduce our game formulation. Section 4 will then explain how it can be used to detect whether density ratio correction is needed or not.

We tie the two problems of density ratio estimation and learning a predictor together through the robust Bayes framework (Grünwald & Dawid, 2004). The learner tries to *minimize* the loss by selecting a model $\boldsymbol{\theta} \in \Theta$, while the adversary tries to *maximize* the loss by selecting a reweighing function $w \in \mathcal{W}$. Formally, we model the learning problem as a (regularized) minimax game:

$$\min_{\boldsymbol{\theta}\in\Theta} \max_{w\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} w(x_i)\, l(f_{\boldsymbol{\theta}}(x_i), y_i) + \lambda\,\Omega(\boldsymbol{\theta}). \tag{3}$$

The learner can be seen as minimizing the worst case loss over the set of test distributions $\mathcal{W}$ produced by the adversary. The definition of the strategy set $\mathcal{W}$ used by the adversary is important in our approach, as it determines the extent to which any model misspecification can be exploited by the adversary to increase the loss. Depending on the application scenario, it can be defined using our prior knowledge on how the test distributions could change, or based on unlabelled test data if they are available.

To refine this formulation, we assume the reweighing functions $w(x)$ are linearly parametrized:

$$w_{\boldsymbol{\alpha}}(x) = \sum_{j=1}^{k} \alpha_j k_j(x), \tag{4}$$

where $\boldsymbol{\alpha}$ contains the mixing coefficients and $k_j(x)$ are non-negative basis functions. For example, each $k_j(x)$ could be a non-negative kernel function, say, the Gaussian kernel

$$K(b_j, x) = \exp\left(-||b_j - x||^2/2\sigma^2\right) \tag{5}$$

with basis b_j, or it could be $I_j(x)$, the indicator function for the jth disjoint group of the data, representing groups from different genders, age ranges, or k-means clusters, etc. It could be viewed as the conditional probability $p(x|j)$ of observing x given class j in a mixture model. As for $\boldsymbol{\alpha}$, it is generally constrained to lie in some compact subspace $\mathcal{A}$ of the non-negative quadrant of Euclidean space. This linear formulation is flexible enough to capture many different types of uncertainties in the test distributions, and yet simple enough to be solved efficiently as a convex optimization problem. Therefore, we consider uncertain test distributions and optimize the following minimax game:

$$\begin{aligned} &\min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{\alpha}\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^{n} w_{\boldsymbol{\alpha}}(x_i)\, l(f_{\boldsymbol{\theta}}(x_i), y_i) + \lambda\,\Omega(\boldsymbol{\theta}) \\ &\text{s.t.}\ \frac{1}{n}\sum_{i=1}^{n} w_{\boldsymbol{\alpha}}(x_i) = 1, \qquad 0 \leqslant \alpha_j \leqslant B. \end{aligned} \tag{6}$$

The sum-to-one normalization constraint ensures that $w_{\boldsymbol{\alpha}}(x)$ behaves like a Radon-Nikodym derivative that properly reweighs the training distribution to a potential test distribution (Shimodaira, 2000; Sugiyama et al., 2008). The

bound $B \in \mathbb{R}$ on the parameters α_j ensure that the reweighing function $w_{\boldsymbol{\alpha}}(x)$ is bounded, which naturally controls the capacity of the adversary. In this formulation, the strategy set[1] $\mathcal{A}_n$ of the adversary is the intersection of a hypercube and an affine subspace:

$$\mathcal{A}_n = \left\{ \boldsymbol{\alpha} \middle| \frac{1}{n}\sum_{i=1}^{n} w_{\boldsymbol{\alpha}}(x_i) = 1, 0 \leqslant \alpha_j \leqslant B \right\}, \quad (7)$$

which is closed and convex. For the games defined above between the learner and the adversary, a minimax solution $(\boldsymbol{\theta}^*, \boldsymbol{\alpha}^*)$ exists (Rockafellar, 1996, Corollary 37.3.2).

3.1. Solving the Training Problem

We first define the *empirical adversarial loss* as

$$L_{\mathcal{A}_n}(\boldsymbol{\theta}) = \max_{\boldsymbol{\alpha} \in \mathcal{A}_n} \frac{1}{n}\sum_{i=1}^{n} w_{\boldsymbol{\alpha}}(x_i)\, l(f_{\boldsymbol{\theta}}(x_i), y_i). \quad (8)$$

The training problem in Eq. (6) can be solved efficiently for loss functions $l(f_{\boldsymbol{\theta}}(\cdot), \cdot)$ that are convex in $\boldsymbol{\theta}$. Notice Eq. (8) is a *convex* function in $\boldsymbol{\theta}$ if $l(f_{\boldsymbol{\theta}}(\cdot), \cdot)$ is convex in $\boldsymbol{\theta}$, as we are taking the maximum over a set of convex functions. A subgradient of $L_{\mathcal{A}_n}(\boldsymbol{\theta}')$ at a point $\boldsymbol{\theta}'$ is:

$$\frac{\partial}{\partial \boldsymbol{\theta}} L_{\mathcal{A}_n}(\boldsymbol{\theta}') = \frac{1}{n}\sum_{i=1}^{n} w_{\boldsymbol{\alpha}'}(x_i) \frac{\partial}{\partial \boldsymbol{\theta}} l(f_{\boldsymbol{\theta}'}(x_i), y_i), \quad (9)$$

where $\boldsymbol{\alpha}'$ is the solution of the problem with $\boldsymbol{\theta}'$ fixed:

$$\boldsymbol{\alpha}' = \operatorname*{argmax}_{\boldsymbol{\alpha} \in \mathcal{A}_n} \frac{1}{n}\sum_{i=1}^{n} w_{\boldsymbol{\alpha}}(x_i)\, l(f_{\boldsymbol{\theta}'}(x_i), y_i). \quad (10)$$

Since the strategy set $\mathcal{A}_n$ is linearly constrained and the objective is also linear, we can easily use linear programming to solve for $\boldsymbol{\alpha}'$ in Eq. (10). Knowing how to compute the subgradient, we can treat the robust training problem as a convex ERM problem with the adversarial loss. The optimization problem can be solved efficiently with subgradient methods or bundle methods (Kiwiel, 1990).

3.2. Incorporating Unlabelled Test Data

If unlabelled test data $\{x_{n+1}, \ldots, x_{n+m}\}$ are available, we would require the reweighing functions $w_{\boldsymbol{\alpha}}(x)$ used by the adversary to produce test distributions that are close to the unlabelled data, especially when covariate shift occurs. In this case we can further restrict the strategy set $\mathcal{A}_n$ of the adversary via moment matching constraints (MMC):

$$\begin{aligned} \min_{\boldsymbol{\theta} \in \Theta} \max_{\boldsymbol{\alpha} \in \mathbb{R}^k} & \frac{1}{n}\sum_{i=1}^{n} w_{\boldsymbol{\alpha}}(x_i)\, l(f_{\boldsymbol{\theta}}(x_i), y_i) + \lambda \Omega(\boldsymbol{\theta}) \\ \text{s.t.} & \frac{1}{n}\sum_{i=1}^{n} w_{\boldsymbol{\alpha}}(x_i) = 1, \qquad 0 \leqslant \alpha_j \leqslant B \\ & \frac{1}{n}\sum_{i=1}^{n} w_{\boldsymbol{\alpha}}(x_i) \boldsymbol{\phi}(x_i) = \frac{1}{m}\sum_{i=n+1}^{n+m} \boldsymbol{\phi}(x_i), \end{aligned} \quad (11)$$

[1] We use the subscript n to denote its dependence on the sample $\{x_1, \cdots, x_n\}$.

where $\boldsymbol{\phi}(\cdot)$ are feature functions similar to those used in maximum entropy models (Berger et al., 1996). Let $K_n \boldsymbol{\alpha} = \bar{\boldsymbol{\phi}}_{te}$ represent the linear constraint of Eq. (11), then the strategy set of the adversary becomes the closed convex set:

$$\mathcal{A}_n^{\text{MMC}} = \left\{ \boldsymbol{\alpha} \middle| \frac{1}{n}\sum_{i=1}^{n} w_{\boldsymbol{\alpha}}(x_i) = 1, 0 \leqslant \alpha_j \leqslant B, K_n \boldsymbol{\alpha} = \bar{\boldsymbol{\phi}}_{te} \right\}.$$

In practice, it might not be feasible to satisfy all the moment matching constraints. It is also unwise to enforce these as hard constraints, as the small test sample might not be representative of the true test distribution. We can incorporate an L_1 or L_2 penalty on the constraint violations similar to Altun & Smola (2006) while retaining convexity of the optimization problem. The details are not shown here due to space constraints. We refer to Eq. (11) as robust covariate shift adjustment (RCSA).

4. Relating Covariate Shift to Model Misspecification

This section relates covariate shift to model misspecification and describes a procedure for testing whether correcting for covariate shift could be needed, assuming the test distribution comes from the strategy set $\mathcal{A}$ of the adversary. We will also state and discuss several theoretical results to justify our test. Their proofs are in Appendix A. We start with a definition:

Definition 1 (Pointwise Domination)**.** *A parameter $\boldsymbol{\theta}^\star$ is said to* pointwisely dominate *all $\boldsymbol{\theta}' \in \Theta$ over the loss function $l(\cdot, \cdot)$ if, for all $x \in \mathcal{X}$ and for all $\boldsymbol{\theta}' \in \Theta$,*

$$\int l(f_{\boldsymbol{\theta}^\star}(x), y) p(y|x) dy \leqslant \int l(f_{\boldsymbol{\theta}'}(x), y) p(y|x) dy. \quad (12)$$

This condition means there is a single $\boldsymbol{\theta}^\star$ that pointwisely minimizes the loss l for any $x \in \mathcal{X}$. It is easy to see that this pointwise domination condition is implied by the traditional definition of correctly specified model class when $l(\cdot, \cdot)$ is the log loss, $-\log p_{\boldsymbol{\theta}}(y|x)$. With log loss, the pointwise domination condition then becomes:

$$-\int p(y|x) \log p_{\boldsymbol{\theta}^\star}(y|x) dy \leqslant -\int p(y|x) \log p_{\boldsymbol{\theta}'}(y|x) dy.$$

This inequality always holds because $p_{\boldsymbol{\theta}^\star}(y|x) = p(y|x)$ minimizes the entropy on the left hand side. Therefore, a correctly specified model always implies the existence of a pointwise dominator $\boldsymbol{\theta}^\star$. However, the converse is not always true, as the underlying model class Θ might be too weak (e.g., if Θ contains only a single model $\boldsymbol{\theta}$).

Note that pointwise domination condition does not depend on the marginal distribution $p(x)$. If we can find such a

pointwise dominator, then the test distribution can be arbitrarily shifted without damaging the performance of pointwise dominator $\boldsymbol{\theta}^\star$. However, this condition is too stringent and is almost never true on real data. This motivates us to consider the game formulation in Section 3: Instead of requiring $\boldsymbol{\theta}$ to minimize the loss at every single point x, we require $\boldsymbol{\theta}$ to minimize the loss against every reweighing function $w_{\boldsymbol{\alpha}}(x)$ that the adversary can play. We define

$$\mathcal{A}_\infty = \left\{ \boldsymbol{\alpha} \in \mathcal{A}^S \middle| \int w_{\boldsymbol{\alpha}}(x)\, dF(x,y) = 1 \right\},$$

where $\mathcal{A}^S$ is the support of the strategy set of the adversary that does not depend on training samples (e.g., the hypercube $0 \leqslant \alpha_j \leqslant B$), and $F(x,y)$ is the joint training distribution of (x,y). Now we define dominant strategy:

Definition 2 (Dominant Strategy)**.** *We say that $\boldsymbol{\theta}^\dagger \in \Theta$ is a* dominant strategy *for the learner if, for all $\boldsymbol{\alpha} \in \mathcal{A}_\infty$, for all $\boldsymbol{\theta}' \in \Theta$,*

$$\int w_{\boldsymbol{\alpha}}(x) l(f_{\boldsymbol{\theta}^\dagger}(x), y) dF(x,y) \leqslant \int w_{\boldsymbol{\alpha}}(x) l(f_{\boldsymbol{\theta}'}(x), y) dF(x,y).$$

It is easy to show that the pointwise domination condition implies the existence of dominant strategy.

Theorem 3. *Suppose a pointwise dominator $\boldsymbol{\theta}^\star$ exists, then $\boldsymbol{\theta}^\star$ is also a dominant strategy for the learner, against any bounded adversarial set $\mathcal{A}$.*

The existence of a dominant strategy of the learner is the key criterion in deciding whether density ratio correction is necessary. If such a dominator $\boldsymbol{\theta}^\dagger$ exists, then it gives lower or equal loss compared to other models $\boldsymbol{\theta}'$, no matter which reweighing function $w_{\boldsymbol{\alpha}}(x)$ is used. Thus if one can find $\boldsymbol{\theta}^\dagger$, no density ratio correction is needed in expectation, since $\boldsymbol{\theta}^\dagger$ is asymptotically optimal as long as the training and test distributions come from the given adversarial set. However, if no such dominator exists, then for any model $\boldsymbol{\theta}$, there exists another model $\boldsymbol{\theta}'$ and a reweighing function $w_{\boldsymbol{\alpha}'}(x)$ such that $\boldsymbol{\theta}'$ has strictly lower loss than $\boldsymbol{\theta}$ on $w_{\boldsymbol{\alpha}'}(x)$. This means that a reweighing $w_{\boldsymbol{\alpha}'}(x)$ and its corresponding model $\boldsymbol{\theta}'$ are preferable. As a result, density ratio correction will be helpful if the test set is drawn from $w_{\boldsymbol{\alpha}'}(x)$ while the training set is not, provided that we can estimate $w_{\boldsymbol{\alpha}'}(x)$ accurately.

How do we know if such a dominant strategy exists for a game between Θ and $\mathcal{A}$? The robust solution of the game can help us in finding out. Let $\bar{\boldsymbol{\theta}}$ be the solution of the unweighed loss minimization problem

$$\bar{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \int l(f_{\boldsymbol{\theta}}(x), y)\, dF(x,y), \tag{13}$$

and $\widehat{\boldsymbol{\theta}}$ be the solution of the reweighed adversarial loss minimization problem

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \max_{\boldsymbol{\alpha} \in \mathcal{A}_\infty} \int w_{\boldsymbol{\alpha}}(x) l(f_{\boldsymbol{\theta}}(x), y)\, dF(x,y). \tag{14}$$

Our main observation is that, if a dominant strategy $\boldsymbol{\theta}^\dagger$ exists, then under suitable assumptions on the adversary, the unweighed solution $\bar{\boldsymbol{\theta}}$ is also a dominant strategy.

Theorem 4. *Suppose the reweighing function $w_{\boldsymbol{\alpha}}(x)$ is linear in $\boldsymbol{\alpha}$, and the constant reweighing $\boldsymbol{\alpha}_0$ with $w_{\boldsymbol{\alpha}_0}(x) = 1$ is in the relative interior of $\mathcal{A}_\infty$. If a dominant strategy $\boldsymbol{\theta}^\dagger$ of the learner exists, then the unweighed solution $\bar{\boldsymbol{\theta}}$ is also a dominant strategy for the learner.*

Thm. 4 suggests a way to test for the existence of dominant strategy. Consider the *adversarial loss*

$$L_{\mathcal{A}_\infty}(\boldsymbol{\theta}) = \max_{\boldsymbol{\alpha} \in \mathcal{A}_\infty} \int w_{\boldsymbol{\alpha}}(x) l(f_{\boldsymbol{\theta}}(x), y)\, dF(x,y).$$

By definition, any dominant strategy $\boldsymbol{\theta}^\dagger$ minimizes the adversarial loss, so Thm. 4 implies that the unweighed solution $\bar{\boldsymbol{\theta}}$ will also minimize the adversarial loss. Therefore, by comparing the value of the minimax solution $L_{\mathcal{A}_\infty}(\widehat{\boldsymbol{\theta}})$ (which by definition minimizes the adversarial loss) against $L_{\mathcal{A}_\infty}(\bar{\boldsymbol{\theta}})$, we can tell if a dominant strategy exists. If they are not equal, then we are certain that no such dominant strategy exists, and density ratio correction can be helpful, depending on the choice of training and test distributions from $\mathcal{A}$. On the other hand, if they are equal, we cannot conclude that a dominant strategy exists, as it is possible that the reweighed adversarial distribution matches the uniform unweighed distribution arbitrarily closely. However, such examples are rather contrived and we never encountered such a situation in any of our experiments.

The final question left is how to compare $L_{\mathcal{A}_\infty}(\widehat{\boldsymbol{\theta}})$ and $L_{\mathcal{A}_\infty}(\bar{\boldsymbol{\theta}})$ in practice with a finite sample. Let $\widehat{\boldsymbol{\theta}}_n$ be a solution of the robust game:

$$\widehat{\boldsymbol{\theta}}_n = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \max_{\boldsymbol{\alpha} \in \mathcal{A}_n} \frac{1}{n} \sum_{i=1}^n w_{\boldsymbol{\alpha}}(x_i) l(f_{\boldsymbol{\theta}}(x_i), y_i), \tag{15}$$

and let $\bar{\boldsymbol{\theta}}_n$ be a solution of the unweighed ERM problem:

$$\bar{\boldsymbol{\theta}}_n = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^n l(f_{\boldsymbol{\theta}}(x_i), y_i). \tag{16}$$

Our convergence results below states that, instead of $L_{\mathcal{A}_\infty}(\widehat{\boldsymbol{\theta}})$ and $L_{\mathcal{A}_\infty}(\bar{\boldsymbol{\theta}})$, we can compare the empirical adversarial losses $L_{\mathcal{A}_n}(\widehat{\boldsymbol{\theta}}_n)$ and $L_{\mathcal{A}_n}(\bar{\boldsymbol{\theta}}_n)$.

Theorem 5. *Suppose the support $\mathcal{A}^S$ for $\boldsymbol{\alpha}$ and Θ for $\boldsymbol{\theta}$ are each closed, convex, and bounded. Suppose also $w_{\boldsymbol{\alpha}}(x)$ and $l(f_{\boldsymbol{\theta}}(x), y)$ are bounded continuous functions in $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ for each (x,y) pair. If the set satisfying the normalization constraint $\{\boldsymbol{\alpha} \in \mathcal{A}^S | \int w_{\boldsymbol{\alpha}}(x) dF(x,y) = 1\}$ is non-empty in the relative interior of $\mathcal{A}^S$, then we have, for all $\boldsymbol{\theta} \in \Theta$,*

$$L_{\mathcal{A}_n}(\boldsymbol{\theta}) \to L_{\mathcal{A}_\infty}(\boldsymbol{\theta})$$

in probability, i.e., for all $\epsilon, \delta > 0$, we can find $m \in \mathbb{N}$ such that for all $n \geqslant m$, we have

$$|L_{\mathcal{A}_n}(\boldsymbol{\theta}) - L_{\mathcal{A}_\infty}(\boldsymbol{\theta})| < \epsilon$$

with probability at least $1-\delta$.

For simplicity, we do not consider the moment matching constraints $K_n\boldsymbol{\alpha} = \bar{\phi}_{te}$, but this can be handled in the proof with techniques similar to the normalization constraint. We use t-test with cross-validation to compare these quantities in the experiments.

The chain of implications can be summarized as follows:

No model misspecification for Θ
$\Rightarrow$Pointwise dominator exists for Θ
$\Rightarrow$Dominant strategy against any bounded adversary $\mathcal{A}$ exists

We can see that "no model misspecification" is a very strong condition, as it requires a dominator against any bounded adversary $\mathcal{A}$, including pathologically spiky test distributions with tall spikes and small support. Also, there is an *implicit* assumption in using density ratio correction that covariate shifts on the test set are not represented by arbitrarily complex functions. Otherwise estimation of density ratio cannot take place and there is no way to correct covariate shift. Therefore, instead of focusing on density ratio correction alone, we look at another way of dealing with covariate shift, by performing model checking against a set of restricted changes in test distributions represented by the adversary $\mathcal{A}$. In the next section we will provide empirical evaluations of this particular approach.

Our analysis is different from Shimodaira (2000). We do not start with a strong condition like "no model misspecification", which is equivalent to requiring the learner to have a dominant strategy against *any* adversaries. Instead, given weak prior knowledge about how the distributions can shift, we provide a method that can determine whether reweighing can be helpful. This analysis is more practical than deciding whether reweighing is needed based on the strong notion of no model misspecification, as almost all models are misspecified on real datasets.

5. Empirical Studies

5.1. Experiments on Toy Datasets

We first present two toy examples to show the effectiveness of our test. We construct a linear model, $f_1(x) = x+1+\epsilon$, and a non-linear (cubic) model, $f_2(x) = x^3 - x + 1 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1^2)$ is additive Gaussian noise (adapted from Shimodaira (2000)). For both, we learn a linear regressor $f_{\boldsymbol{\theta}}(x) = \theta_1 \cdot x + \theta_0$ from data to minimize squared loss $l(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) = ||\boldsymbol{\theta}^T\mathbf{x}_i - y_i||^2$ with L_2 regularizer on $\boldsymbol{\theta}$: $\Omega(\boldsymbol{\theta}) = \frac{1}{2}||\boldsymbol{\theta}||_2^2$.

To show how to detect whether a dominant strategy exists with various adversarial sets $\mathcal{A}$, we generate 500 data points uniformly in the interval $[-1.5, 2]$, which we partition into training and test sets via 10-fold cross validation. To construct reasonable adversaries, we use Eq.(4) with Gaussian kernel as our reweighing function. As we mentioned earlier, the adversarial set is determined by prior knowledge of how the test distribution might change. In this toy example, we use a large range of σ, based on the average distance from an instance to its $\frac{n}{c}$-nearest neighbours, where n is the number of training points and $c \in \{2, 4, 8, 16, \cdots\}$. The smaller σ is, the more powerful the adversary can be, i.e., the more possible test distributions it can generate. The bases, b_j, are chosen to be the training points. B is set to be 5, a bound that is rarely reached in practice due to the normalization constraint. Therefore, this bound does not significantly limit the adversary's power, as it allows the adversary to put as much importance on a single kernel as it wants. We tune the parameter λ via 10-fold cross validation.[2] Figure 2(a) shows that $L_{\mathcal{A}_n}(\widehat{\boldsymbol{\theta}}_n)$ and $L_{\mathcal{A}_n}(\bar{\boldsymbol{\theta}}_n)$ (mean and one standard deviation as error bar) are very close for all σ in the linear example, indicating that the adversary cannot exploit the weakness of linear learner. Figure 2(b) shows that, for the non-linear example, even with moderate σ, there is a noticeable difference between $L_{\mathcal{A}_n}(\widehat{\boldsymbol{\theta}}_n)$ against $L_{\mathcal{A}_n}(\bar{\boldsymbol{\theta}}_n)$, strongly suggesting that no dominant strategy exists in this case, which suggests that covariate shift correction is necessary if the test distribution is shifted here. The experiments showing covariate shift scenarios are in Appendix B due to space limits.

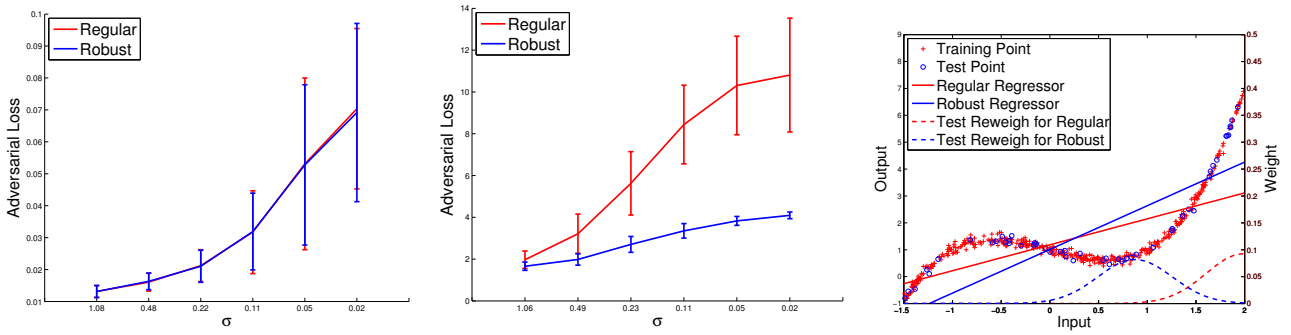
To see how the adversary creates different empirical adversarial losses in a non-linear example, we fix the σ to the average distance from an instance to its $\frac{n}{5}$-nearest neighbour and illustrate a concrete trial in Figure 2(c). It is obvious that the adversary puts more weights at the test points where the loss of the classifier learned from training data is large. Our robust formulation incorporates the adversary and prevents any point from having too large a loss. As a result, the adversary cannot undermine the robust learner severely, which leads to the gap of the empirical adversarial losses of robust and regular learners in Figure 2(b).

5.2. Experiments on Real-world Datasets

This section presents the experimental results on real world datasets to show how our formulation determines whether dominant strategy exists against some adversaries and if so, how to correct such covariate shifts. We investigate both regression problems using squared loss, and classification problems using hinge loss. A linear model is learned from the dataset unless otherwise specified.

We obtain some classification datasets from UCI reposi-

[2]Here, as there is no covariate shift, we just use simple cross validation. Whenever test distribution is shifted in the experiment, parameters are tuned via importance weighted cross validation (Sugiyama et al., 2007).



(a) Linear example. (b) Cubic example. (c) Adversarial reweighing.

Figure 2. Toy examples. Adversarial test losses are shown in Figures 2(a) and 2(b), where the x-axis shows the values of σ. Figure 2(c) provides a non-linear example to show how the adversary attacks the regressors by reweighing the test points, with output on the left y-axis and weight on the right y-axis.

Table 1. Dataset Summary

DATASET	SIZE	DIM	TYPE
AUSTRALIAN	690	14	CLASSIFICATION
BREAST_CANCER	683	10	CLASSIFICATION
GERMAN_NUMER	1000	24	CLASSIFICATION
HEART	270	13	CLASSIFICATION
IONOSPHERE	351	34	CLASSIFICATION
LIVER_DISORDER	345	6	CLASSIFICATION
SONAR	208	60	CLASSIFICATION
SPLICE	1000	60	CLASSIFICATION
AUTO-MPG	392	6	REGRESSION
CANCER	1523	40	REGRESSION

tory[3]. All are binary classification problems. For regression task, we use `Auto-mpg` dataset, which admits a natural covariate shift scenario, as it contains data collected from 3 different cities. We also have a set of cancer patient survival time data provided by our medical collaborators, containing 1523 uncensored patients with 40 features, including gender, stage of cancer, and various measurements obtained at the time of diagnosis. Table 1 shows the summary of the datasets we used in the experiments.

5.2.1. DOMINANT STRATEGY DETECTION

We first detect the existence of dominant strategy as we did in the toy example. To construct reasonable adversaries, Gaussian kernel is applied to Eq.(4), setting σ to be the average distance from an instance to its $\frac{n}{5}$-nearest neighbour, the bases b_j to be the training points and B to be 5. [4]

Figure 3(a) shows the experimental results. `Auto-mpg12` explores when the training data comes from city 1 and test data is from city 2, while `Auto-mpg13` explores when training data comes from city 1 and test data comes from city 3. Here we focus on the empirical adversarial losses of robust versus regular models. A significant difference indicates that there is no dominant strategy and thus, the linear model is vulnerable to our reweighing adversary. For classification datasets and the cancer dataset, we apply 10-fold cross validation to obtain training and test sets. For `Auto-mpg`, we fix the test set and apply 10-fold cross validation to obtain training set. Figure 3(a) presents these losses over the datasets (mean and one standard deviation as error bar). t-test at significance level 0.05 indicates that two losses are significantly different for the `Liver_disorders` and `Auto-mpg` datasets. As a result, the linear model is vulnerable for these sets.

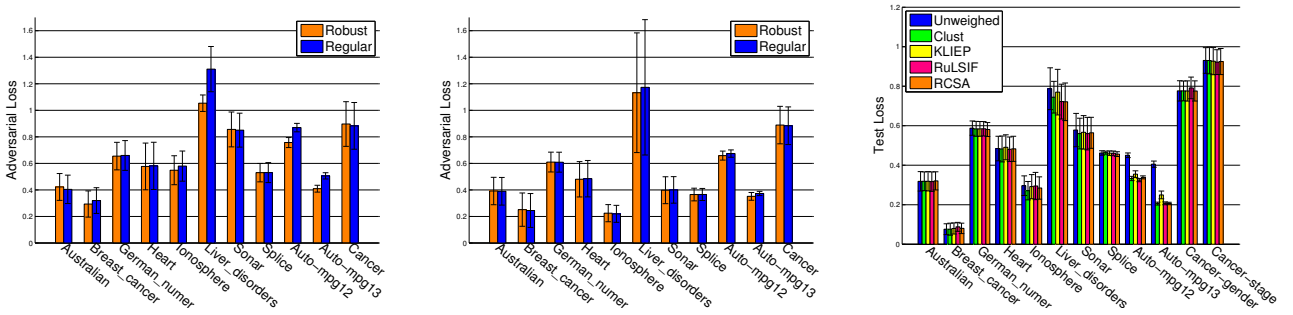
To further substantiate the incapability of the linear model, we attempted to detect dominant strategy for a Gaussian model set Θ (i.e., changing from linear kernel to Gaussian kernel where the learner use internal cross-validation to chose the kernel width). Results are shown in Figure 3(b). Compared to Figure 3(a), the gap of empirical adversarial losses between robust and regular models shrinks significantly in Figure 3(b). Our result indicates that t-test no longer claims a significant difference between these losses, suggesting that the adversary cannot severely undermine the performance of regular learning. Therefore, model revision can be a good alternative to performing covariate shift correction.

5.2.2. REWEIGHING ALGORITHM FOR COVARIATE SHIFT SCENARIOS

As previously mentioned, the reweighing mechanism could improve the performance if the model is vulnerable to the reweighing adversary. For the covariate shift correction task, we set the test points as the reference bases b_j of the weight function (Eq. (4)), because they are more informa-

[3] http://archive.ics.uci.edu/ml/index.html

[4] We allow the user to set up other adversaries, as the appropriate adversary depends on user's belief about how the test distribution may change. We use this medium power adversary to differentiate between datasets under linear models. If the adversary is too weak, no correction is needed for all datasets. If the adversary is too strong, all datasets require correction, as on real data there is no "correct" model. We omit these less interesting cases due to space limits.



(a) Linear learner. (b) Gaussian learner. (c) Reweighed linear model performance.

Figure 3. Experimental results for dominant strategy detection and covariate shift correction. Figure 3(a) and Figure 3(b) show the adversarial test losses of robust and regular learners.

tive than training points about the test distribution, as suggested by Sugiyama et al. (2008). The reweighing set (σ and B) is chosen as in Section 5.2.1.

To create covariate shift scenarios in the classification datasets, we apply the following scheme to obtain shifted test set. We first randomly pick 75% of the whole set for robust training (6) to attain a model $\boldsymbol{\theta}$. Then we evaluate the empirical adversarial loss (8) of this model on the 25% hold-out test set, and record all the weights of these test points. The probability that a test instance x will remain in the test set is $\min\left(1, \frac{w_{\boldsymbol{\alpha}}(x)}{1/m}\right)$, where m is the number of test points at the moment (25% of the set). About 10% of the whole dataset remain after filtering; these instances will serve as the final test set with shifted distribution. The intuition is: we want some test points with large errors even for robust learner. These points are more likely to undermine any model, meaning covariate shift will have more significant impact on the performance. The procedure is performed 10 times, leading to the average test losses reported in Figure 3(c). As `Auto-mpg` has a natural covariate shift scenario, we do not artificially partition the dataset. We applied 10-fold cross validation to obtain the training set. We consider two shifted scenarios in cancer survival time prediction:

1. `Gender` split. The dataset contains about 60% male and 40% female patients. In gender split, we randomly take 20% of the male and 80% of the female patients into training set, while the rest goes to test set. That is, the training set is dominated by male patients while the test set is dominated by female patients.

2. Cancer `stage` split. Approximately 70% of the dataset are of stage-4. In cancer stage split, we randomly take 20% of stage-1-to-3 and 80% of stage-4 patients to training set, while the rest goes to test set. The training set is dominated by stage-4 patients while the test set is dominated by stage-1-to-3 patients.

Figure 3(c) compares the test losses of RCSA with the regular unweighed learning algorithm, the clustering-based reweighing algorithm (Cortes et al., 2008), KLIEP (Sugiyama et al., 2008) and RuLSIF (Yamada et al., 2011). Recall that our analysis in Section 5.2.1 shows that linear model is insufficient for the `Liver_disorders` and `Auto-mpg` datasets, which suggests that reweighing may help. This is confirmed in Figure 3(c): by putting more weights on the training instances that are similar to test instances, the reweighing algorithms can produce models with smaller test losses for these datasets. Although our robust game formulation is mainly designed to detect dominant strategy, our RCSA algorithm can correct shifted distribution using moment matching constraints described in Section 3.2. As shown in Figure 3(c), our method performs on par with state-of-the-art algorithms when covariate shift correction is required. For the datasets that appear linear (i.e., where the linear model performs relatively well), we found that the reweighing algorithms did not significantly reduce the test losses. In some cases, reweighing actually increased the test losses due to the presence of noise.

6. Conclusions

We have provided a method for determining if covariate shift correction is needed, given a pre-defined set of potential changes in the test distribution. This is useful for ensuring the learned predictor will still perform well when there are uncertainties about the test distribution in the deployment environment, such as changes in gender ratio and case mix in the cancer prognostic predictor example. It can also be used to decide if a model class revision of Θ is necessary. Experimental results show that our detection test is effective on several UCI datasets and a real-world cancer patient dataset. This analysis shows the importance of studying the interaction of covariate shift and model misspecification, because the final test set error depends on both factors.

References

Altun, Y. and Smola, A. Unifying divergence minimization and statistical inference via convex duality. In *Learning theory*, pp. 139–153. Springer, 2006.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 19, pp. 137, 2007.

Berger, A., Pietra, V., and Pietra, S. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning for differing training and test distributions. In *International Conference on Machine Learning*, pp. 81–88, 2007.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning under covariate shift. *The Journal of Machine Learning Research*, 10:2137–2155, 2009.

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. Sample selection bias correction theory. *Algorithmic Learning Theory*, 5254:38–53, 2008.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. *Advances in Neural Information Processing Systems*, 23:442–450, 2010.

Globerson, Amir and Roweis, Sam. Nightmare at test time: robust learning by feature deletion. In *International Conference on Machine learning*, pp. 353–360. ACM, 2006.

Grünwald, P. D. and Dawid, A. P. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M., and Schölkopf, B. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, volume 19, pp. 601, 2007.

Kanamori, T., Hido, S., and Sugiyama, M. Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection. In *Advances in Neural Information Processing Systems*, 2008.

Kiwiel, K. C. Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical Programming*, 46(1):105–122, 1990.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. Domain adaptation via transfer component analysis. In *International Joint Conference on Artifical Intelligence*, pp. 1187–1192, 2009.

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

Rockafellar, R. T. *Convex analysis*. Princeton University Press, 1996.

Schölkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization and beyond*. The MIT Press, 2002.

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

Storkey, A. J. and Sugiyama, M. Mixture regression for covariate shift. In *Advances in Neural Information Processing Systems*, 2007.

Sugiyama, M. and Kawanabe, M. *Machine learning in non-stationary environments: introduction to covariate shift adaptation*. The MIT Press, 2012.

Sugiyama, M. and Müller, K. Model selection under covariate shift. In *Artificial Neural Networks: Formal Models and Their Applications*, pp. 235–240. Springer, 2005.

Sugiyama, M., Krauledat, M., and Müller, K. Covariate shift adaptation by importance weighted cross validation. *Journal of Maching Learning Research*, 8:985–1005, 2007.

Sugiyama, M., Nakajima, S., Kashima, H., Von Buenau, P., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, 2008.

Teo, C. H., Globerson, A., Roweis, S., and Smola, A. Convex learning with invariances. In *Advances in Neural Information Processing Systems*, 2008.

White, H. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, 76(374):419–433, 1981.

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., and Sugiyama, M. Relative density-ratio estimation for robust distribution comparison. In *Advances in Neural Information Processing Systems*, pp. 594–602, 2011.

Zadrozny, B. Learning and evaluating classifiers under sample selection bias. In *International Conference on Machine Learning*, pp. 114. ACM, 2004.

Appendix A

Proof of Theorem 3

Proof. By the definition of a pointwise dominator

$$\int l(f_{\boldsymbol{\theta}^\star}(x), y)dF(y \mid x) - \int l(f_{\boldsymbol{\theta}'}(x), y)dF(y \mid x) \leqslant 0$$

for all $\boldsymbol{\theta}' \in \Theta$. Given any bounded adversarial set $\mathcal{A}$, for any $\boldsymbol{\alpha} \in \mathcal{A}$, $w_{\boldsymbol{\alpha}}(x)$ is a non-negative function of x. Therefore integrating with respect to $dF(x)$ gives

$$\int w_{\boldsymbol{\alpha}}(x)\left[\int l(f_{\boldsymbol{\theta}^\star}(x), y)dF(y \mid x) - \int l(f_{\boldsymbol{\theta}'}(x), y)dF(y \mid x)\right]dF(x) \leqslant 0$$

$$\int w_{\boldsymbol{\alpha}}(x) l(f_{\boldsymbol{\theta}^\star}(x), y)dF(x, y) \leqslant \int w_{\boldsymbol{\alpha}}(x) l(f_{\boldsymbol{\theta}'}(x), y)dF(x, y).$$

Thus $\boldsymbol{\theta}^\star$ is also a dominant strategy against the adversarial set $\mathcal{A}$. □

Proof of Theorem 4

Proof. We use $\boldsymbol{h}(\boldsymbol{\theta})$ from Eq. (19) to denote the cost vector for expected adversarial loss, with the extra argument $\boldsymbol{\theta}$ to emphasize its dependence on $\boldsymbol{\theta}$. As $\boldsymbol{\theta}^\dagger$ is a dominant strategy, we have

$$\boldsymbol{h}(\boldsymbol{\theta}^\dagger)^T\boldsymbol{\alpha} \leqslant \boldsymbol{h}(\bar{\boldsymbol{\theta}})^T\boldsymbol{\alpha} \Rightarrow (\boldsymbol{h}(\boldsymbol{\theta}^\dagger) - \boldsymbol{h}(\bar{\boldsymbol{\theta}}))^T\boldsymbol{\alpha} \leqslant 0 \quad (17)$$

for all $\boldsymbol{\alpha} \in \mathcal{A}_\infty$. By definition, $\bar{\boldsymbol{\theta}}$ minimizes the adversarial loss for the constant unweighed strategy $\boldsymbol{\alpha}_0$ of the adversary, so we have

$$(\boldsymbol{h}(\boldsymbol{\theta}^\dagger) - \boldsymbol{h}(\bar{\boldsymbol{\theta}}))^T\boldsymbol{\alpha}_0 = 0. \quad (18)$$

Let $\boldsymbol{\alpha}' \in \mathcal{A}_\infty$. As $\boldsymbol{\alpha}_0$ is in the relative interior of $\mathcal{A}_\infty$ and $\mathcal{A}_\infty$ is convex, there exists $\epsilon > 0$ such that

$$\boldsymbol{\alpha}'' = \boldsymbol{\alpha}' + (1 + \epsilon)(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}')$$

is in $\mathcal{A}_\infty$. Now by Eq. (17) and (18), we have three colinear points such that

$$(\boldsymbol{h}(\boldsymbol{\theta}^\dagger) - \boldsymbol{h}(\bar{\boldsymbol{\theta}}))^T\boldsymbol{\alpha}' \leqslant 0$$
$$(\boldsymbol{h}(\boldsymbol{\theta}^\dagger) - \boldsymbol{h}(\bar{\boldsymbol{\theta}}))^T\boldsymbol{\alpha}_0 = 0$$
$$(\boldsymbol{h}(\boldsymbol{\theta}^\dagger) - \boldsymbol{h}(\bar{\boldsymbol{\theta}}))^T\boldsymbol{\alpha}'' \leqslant 0.$$

So $(\boldsymbol{h}(\boldsymbol{\theta}^\dagger) - \boldsymbol{h}(\bar{\boldsymbol{\theta}}))^T\boldsymbol{\alpha}$ must be identically 0 on the interval $[\boldsymbol{\alpha}', \boldsymbol{\alpha}'']$, as it is a linear function in $\boldsymbol{\alpha}$.

This shows $\boldsymbol{h}(\bar{\boldsymbol{\theta}})^T\boldsymbol{\alpha}' = \boldsymbol{h}(\boldsymbol{\theta}^\dagger)^T\boldsymbol{\alpha}'$. As $\boldsymbol{\alpha}'$ is arbitrary, the unweighed solution $\bar{\boldsymbol{\theta}}$ is also a dominant strategy for the learner Θ. □

Proof of Theorem 5

Notice the reweighed loss is linear in $\boldsymbol{\alpha}$ for fixed $\boldsymbol{\theta}$:

$$\frac{1}{n}\sum_{i=1}^n w_{\boldsymbol{\alpha}}(x_i) l(f_{\boldsymbol{\theta}}(x_i), y_i) = \sum_{j=1}^k \alpha_j \frac{1}{n}\sum_{i=1}^n k_j(x_i) l(f_{\boldsymbol{\theta}}(x_i), y_i)$$
$$= \boldsymbol{h}_n^T\boldsymbol{\alpha},$$

where

$$(\boldsymbol{h}_n)_j = \frac{1}{n}\sum_{i=1}^n k_j(x_i) l(f_{\boldsymbol{\theta}}(x_i), y_i).$$

Therefore we can write $L_{\mathcal{A}_n}(\boldsymbol{\theta})$ as:

$$L_{\mathcal{A}_n}(\boldsymbol{\theta}) = \max_{\boldsymbol{\alpha} \in \mathcal{A}_n} \boldsymbol{h}_n^T\boldsymbol{\alpha}$$

Similarly, define the corresponding cost vector $\boldsymbol{h}$ for the expected adversarial loss such that

$$(\boldsymbol{h})_j = \int k_j(x)\, l_{\boldsymbol{\theta}}(f(x), y)dF(x, y), \quad (19)$$

and we have

$$L_{\mathcal{A}_\infty}(\boldsymbol{\theta}) = \max_{\boldsymbol{\alpha} \in \mathcal{A}_\infty} \boldsymbol{h}^T\boldsymbol{\alpha}$$

Similarly, define for the normalization constraint:

$$\frac{1}{n}\sum_{i=1}^n w_{\boldsymbol{\alpha}}(x_i) = \sum_{j=1}^k \alpha_j \frac{1}{n}\sum_{i=1}^n k_j(x_i) = \boldsymbol{g}_n^T\boldsymbol{\alpha},$$

where

$$(\boldsymbol{g}_n)_j = \frac{1}{n}\sum_{i=1}^n k_j(x_i).$$

Define the corresponding constraint vector $\boldsymbol{g}$ such that

$$(\boldsymbol{g})_j = \int k_j(x)\, dF(x, y).$$

Translating into the new notations, we want to prove

$$\left|\max_{\boldsymbol{\alpha} \in \mathcal{A}^S : \boldsymbol{g}^T\boldsymbol{\alpha} = 1} \boldsymbol{h}^T\boldsymbol{\alpha} - \max_{\boldsymbol{\alpha} \in \mathcal{A}^S : \boldsymbol{g}_n^T\boldsymbol{\alpha} = 1} \boldsymbol{h}_n^T\boldsymbol{\alpha}\right| < \epsilon$$

with probability at least $1 - \delta$, for all sufficiently large n.

To prove the result we need two lemmas, whose proofs appear after the main proof. The first lemma states that the sample cost vector $\boldsymbol{h}_n$ converges in the infinite limit to $\boldsymbol{h}$. The second lemma states that near the feasible solutions of $\boldsymbol{g}^T\boldsymbol{\alpha} = 1$, there are feasible solutions of finite sample constraint $\boldsymbol{g}_n^T\boldsymbol{\alpha} = 1$ for large n, and also vice versa.

Lemma 6. *Assume the basis $k_j(x)$ for the reweighing function $w(x)$ are bounded above by B_k, and the loss function l bounded above by B_l. We then have*

$$Pr(\|\boldsymbol{h}_n - \boldsymbol{h}\|_2 \geqslant \epsilon) \leqslant 2k \exp\left(-\frac{2n\epsilon^2}{B_k^2 B_l^2 k^2}\right).$$

Lemma 7. *Suppose $\epsilon, \delta > 0$ are given and let $\boldsymbol{\alpha}^* \in \mathcal{A}_\infty$. Then there exists $m \in \mathbb{N}$ such that, for all $n \geqslant m$, with probability at least $1 - \delta$, we can find $\boldsymbol{\alpha}_n \in \mathcal{A}_n$ such that*

$$\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}_n\| \leqslant \epsilon$$

Similarly, suppose $\epsilon, \delta > 0$ are given. Then there exists $m \in \mathbb{N}$ such that for all $n \geqslant m$, for any $\boldsymbol{\alpha}_n \in \mathcal{A}_n$, with probability at least $1 - \delta$, we can find $\boldsymbol{\alpha}^* \in \mathcal{A}_\infty$ such that

$$\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}^*\| \leqslant \epsilon.$$

PROOF OF MAIN THEOREM

By Lemma 6, there exists $n_1 \in \mathbb{N}$ such that for all $n \geqslant n_1$, $\|\boldsymbol{h}-\boldsymbol{h}_n\| \leqslant \epsilon/(2B_\alpha)$ with probability $1-\delta/3$. [condition 1]

Let $\boldsymbol{h}^T\boldsymbol{\alpha}^* = \max_{\boldsymbol{\alpha}\in\mathcal{A}_\infty} \boldsymbol{h}^T\boldsymbol{\alpha}$. By Lemma 7, there exists $n_2 \in \mathbb{N}$ such that for all $n \geqslant n_2$, we can find $\boldsymbol{\alpha}'_n \in A_n$ with $\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}'_n\| \leqslant \epsilon/(2\|\boldsymbol{h}\|)$ with probability $1-\delta/3$ [condition 2]. Conditions 1 and 2 give

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}\in\mathcal{A}_\infty} \boldsymbol{h}^T\boldsymbol{\alpha} - \max_{\boldsymbol{\alpha}\in\mathcal{A}_n} \boldsymbol{h}_n^T\boldsymbol{\alpha} \\
&\leqslant \boldsymbol{h}^T\boldsymbol{\alpha}^* - \boldsymbol{h}_n^T\boldsymbol{\alpha}'_n \\
&= \boldsymbol{h}^T\boldsymbol{\alpha}^* - \boldsymbol{h}^T\boldsymbol{\alpha}'_n + \boldsymbol{h}^T\boldsymbol{\alpha}'_n - \boldsymbol{h}_n^T\boldsymbol{\alpha}'_n \\
&= \boldsymbol{h}^T(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}'_n) + (\boldsymbol{h}-\boldsymbol{h}_n)^T\boldsymbol{\alpha}'_n \\
&\leqslant \|\boldsymbol{h}\|\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}'_n\| + \|\boldsymbol{h}-\boldsymbol{h}_n\|\|\boldsymbol{\alpha}'_n\| \\
&\leqslant \|\boldsymbol{h}\|\frac{\epsilon}{2\|\boldsymbol{h}\|} + \frac{\epsilon}{2B_\alpha}B_\alpha = \epsilon
\end{aligned}$$

Similarly, let $\boldsymbol{h}_n^T\boldsymbol{\alpha}_n^* = \max_{\boldsymbol{\alpha}\in\mathcal{A}_n} \boldsymbol{h}_n^T\boldsymbol{\alpha}$. By Lemma 7, there exists $n_3 \in \mathbb{N}$ such that for each $n \geqslant n_3$, we can find $\boldsymbol{\alpha}' \in A_\infty$ with $\|\boldsymbol{\alpha}_n^* - \boldsymbol{\alpha}'\| \leqslant \epsilon/2\|\boldsymbol{h}\|$ with probability $1-\delta/3$ [condition 3]. Conditions 1 and 3 give

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}\in\mathcal{A}_n} \boldsymbol{h}_n^T\boldsymbol{\alpha} - \max_{\boldsymbol{\alpha}\in\mathcal{A}_\infty} \boldsymbol{h}^T\boldsymbol{\alpha} \\
&\leqslant \boldsymbol{h}_n^T\boldsymbol{\alpha}_n^* - \boldsymbol{h}^T\boldsymbol{\alpha}' \\
&\leqslant \boldsymbol{h}_n^T\boldsymbol{\alpha}_n^* - \boldsymbol{h}^T\boldsymbol{\alpha}_n^* + \boldsymbol{h}^T\boldsymbol{\alpha}_n^* - \boldsymbol{h}^T\boldsymbol{\alpha}' \\
&= (\boldsymbol{h}_n-\boldsymbol{h})^T\boldsymbol{\alpha}_n^* + \boldsymbol{h}^T(\boldsymbol{\alpha}_n^* - \boldsymbol{\alpha}') \\
&\leqslant \|\boldsymbol{h}_n-\boldsymbol{h}\|\|\boldsymbol{\alpha}_n^*\| + \|\boldsymbol{h}\|\|\boldsymbol{\alpha}_n^* - \boldsymbol{\alpha}'\| \\
&\leqslant \frac{\epsilon}{2B_\alpha}B_\alpha + \|\boldsymbol{h}\|\frac{\epsilon}{2\|\boldsymbol{h}\|} = \epsilon
\end{aligned}$$

Therefore when $n \geqslant \max\{n_1, n_2, n_3\}$, with probability at least $1-\delta$ (by union bound), we have

$$|\max_{\boldsymbol{\alpha}\in\mathcal{A}_n} \boldsymbol{h}_n^T\boldsymbol{\alpha} - \max_{\boldsymbol{\alpha}\in\mathcal{A}_\infty} \boldsymbol{h}^T\boldsymbol{\alpha}| \leqslant \epsilon \qquad \square$$

PROOF OF LEMMA 6

By Hoeffding's inequality, we have

$$Pr(|(\boldsymbol{h}_n)_j - (\boldsymbol{h})_j| > \frac{\epsilon}{k}) \leqslant 2\exp\left(-\frac{2n\epsilon^2}{B_k^2B_l^2k^2}\right).$$

By union bound, we have

$$Pr(\|\boldsymbol{h}_n - \boldsymbol{h}\|_1 \geqslant \epsilon) \leqslant 2k\exp\left(-\frac{2n\epsilon^2}{B_k^2B_l^2k^2}\right).$$

As $\|\boldsymbol{h}_n - \boldsymbol{h}\|_2 \leqslant \|\boldsymbol{h}_n - \boldsymbol{h}\|_1$, we have

$$Pr(\|\boldsymbol{h}_n - \boldsymbol{h}\|_2 \geqslant \epsilon) \leqslant 2k\exp\left(-\frac{2n\epsilon^2}{B_k^2B_l^2k^2}\right). \qquad \square$$

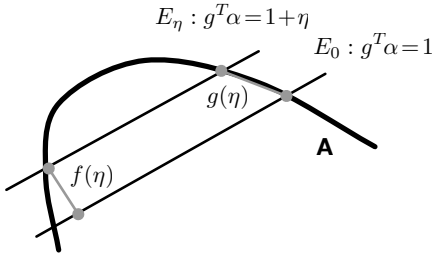


Figure 4. Definition of $f(\eta)$ and $g(\eta)$

PROOF OF LEMMA 7

Using Hoeffding's inequality and union bound (similar to the proof of Lemma 6), we have

$$Pr(\|\boldsymbol{g}_n - \boldsymbol{g}\|_2 \geqslant \epsilon) \leqslant 2k\exp\left(-\frac{2n\epsilon^2}{B_k^2k^2}\right). \qquad (20)$$

Define

$$E_\eta = \{\boldsymbol{\alpha} \in \mathcal{A}^S | \boldsymbol{g}^T\boldsymbol{\alpha} = 1+\eta\}$$

for each $\eta \in \mathbb{R}$. This is the set of subspaces parallel to $\boldsymbol{g}^T\boldsymbol{\alpha} = 1$ (E_0). Define also $f(\eta)$ as the maximum distance of any points in E_η to E_0, and $g(\eta)$ as the maximum distance of any points in E_0 to E_η (see Fig. 4), i.e.,

$$\begin{aligned}
f(\eta) &= \max_{\boldsymbol{\alpha}\in E_\eta} \min_{\boldsymbol{\alpha}'\in E_0} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|, \\
g(\eta) &= \max_{\boldsymbol{\alpha}\in E_0} \min_{\boldsymbol{\alpha}'\in E_\eta} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|.
\end{aligned}$$

Suppose $\epsilon, \delta > 0$ are given. Using Lemma 8 below, $f(\eta) \to 0$ as $\eta \to 0$, so we can find $\eta_0 > 0$ such that $f(\eta) < \epsilon$ whenever $|\eta| < \eta_0$. From Eq. (20), we can find $m \in \mathbb{N}$ such that for all $n \geqslant m$, $\|\boldsymbol{g}_n - \boldsymbol{g}\| < \eta_0/B_\alpha$ with probability at least $1-\delta$.

Letting $\boldsymbol{\alpha}_n \in \mathcal{A}_n$ for $n \geqslant m$, we have

$$|\boldsymbol{g}^T\boldsymbol{\alpha}_n - 1| = |\boldsymbol{g}^T\boldsymbol{\alpha}_n - \boldsymbol{g}_n^T\boldsymbol{\alpha}_n| \leqslant \|\boldsymbol{g}-\boldsymbol{g}_n\|\|\boldsymbol{\alpha}_n\| \leqslant \frac{\eta_0}{B_\alpha}B_\alpha = \eta_0 \qquad (21)$$

with probability at least $1-\delta$. Hence the subspace $\boldsymbol{g}_n^T\boldsymbol{\alpha} = 1$, i.e. $\mathcal{A}_n$, lies between E_{η_0} and $E_{-\eta_0}$ with probability $1-\delta$. Specifically for a fixed $\boldsymbol{\alpha}_n \in \mathcal{A}_n$, it lies on E_η for some η with $|\eta| < \eta_0$. Therefore

$$\min_{\boldsymbol{\alpha}'\in E_0} \|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}'\| \leqslant \max_{\boldsymbol{\alpha}\in E_\eta} \min_{\boldsymbol{\alpha}'\in E_0} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\| = f(\eta) \leqslant \epsilon$$

with probability $1-\delta$.

For the second part, let $\boldsymbol{\alpha}^* \in \mathcal{A}_\infty (= E_0)$, and $\epsilon, \delta > 0$ be given. Using Lemma 8 below, $g(\eta) \to 0$ as $\eta \to 0$, so we can find $\eta_0 > 0$ such that $g(\eta) < \epsilon$ whenever $|\eta| < \eta_0$.

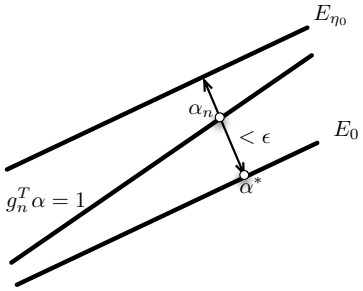


Figure 5. Illustration for proof of Lemma 7

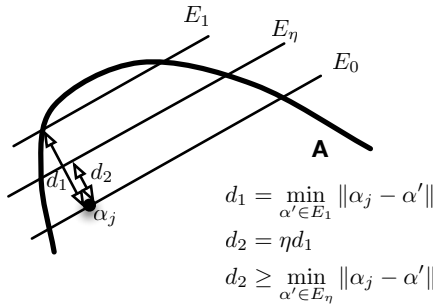


Figure 6. Illustration for the proof of Lemma 8

By Eq. (21) above, we can find $m \in \mathbb{N}$ such that $\mathcal{A}_n$ lies entirely between $E_{-\eta_0}$ and E_{η_0} with probability at least $1 - \delta$. By definition

$$\min_{\boldsymbol{\alpha}' \in E_{\eta_0}} \|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}'\| \leqslant g(\eta_0) \leqslant \epsilon.$$

Let $\boldsymbol{\alpha}_{\eta_0}$ be a point on E_{η_0} minimizing the distance to $\boldsymbol{\alpha}^*$, then the line joining $\boldsymbol{\alpha}_{\eta_0}$ and $\boldsymbol{\alpha}^*$ has to intersect with the subspace $\boldsymbol{g}_n^T \boldsymbol{\alpha} = 1$ at some $\boldsymbol{\alpha}_n$ (see Fig. 5). This holds for all $n \geqslant m$ and we have $\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}^*\| \leqslant \epsilon$. The same argument applies to the case when $\boldsymbol{g}_n^T \boldsymbol{\alpha} = 1$ lies between E_0 and $E_{-\eta_0}$. Thus

$$\min_{\boldsymbol{\alpha}' \in \mathcal{A}_n} \|\boldsymbol{\alpha}' - \boldsymbol{\alpha}^*\| \leqslant \epsilon$$

for all $n \geqslant m$, with probability at least $1 - \delta$.

Lemma 8.

$$f(\eta) = \max_{\boldsymbol{\alpha} \in E_\eta} \min_{\boldsymbol{\alpha}' \in E_0} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|$$
$$g(\eta) = \max_{\boldsymbol{\alpha} \in E_0} \min_{\boldsymbol{\alpha}' \in E_\eta} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|$$

converge to 0 as $\eta \to 0$.

Proof. We want to show $f(\eta) \to 0$ as $\eta \to 0$. If not, then there exists $f_0 > 0$ and a sequence $\{\eta_t\}_{t=1}^{\infty}$ with $\eta_t \to 0$, such that $f(\eta_t) \geqslant f_0$ infinitely often. We collect all those indices t_n such that $f(\eta_{t_n}) \geqslant f_0$, and form a new sequence $\mu_n = \eta_{t_n}$. Let

$$\boldsymbol{\alpha}_n = \operatorname*{argmax}_{\boldsymbol{\alpha} \in E_{\mu_n}} \min_{\boldsymbol{\alpha}' \in E_0} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|.$$

As $\boldsymbol{\alpha}_n$ lies in a compact set $\mathcal{A}^S$, there exist a convergent subsequence, say $\boldsymbol{\beta}_n$. Let the subsequence $\boldsymbol{\beta}_n$ converge to some $\boldsymbol{\beta}$, and by continuity we know $\boldsymbol{g}^T \boldsymbol{\beta} = 1$, so $\boldsymbol{\beta} \in E_0$.

The function

$$s(\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha}' \in E_0} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|$$

is a continuous function in $\boldsymbol{\alpha}$ (minimum of a bivariate continuous function over a compact set).

We have $s(\boldsymbol{\beta}_n) \geqslant f_0$ and $\boldsymbol{\beta}_n \to \boldsymbol{\beta}$, so $s(\boldsymbol{\beta}_n)$ converges to some $f_0' \geqslant f_0$ as s is continuous. However, since $\boldsymbol{\beta} \in E_0$, we have $s(\boldsymbol{\beta}) = 0$. This creates a contradiction and therefore $f(\eta) \to 0$.

Next we want to show $g(\eta) \to 0$ as $\eta \to 0$. Given $\gamma > 0$, as E_0 is compact, we can cover E_0 with at most k balls of radius $\gamma/2$ for some finite k. We label the centres of these balls as $\boldsymbol{\alpha}_j$, $1 \leqslant j \leqslant k$.

We consider the case where $\eta > 0$. The case for $\eta < 0$ is symmetric. By the assumption of the theorem the set $\{\boldsymbol{\alpha} \in \mathcal{A}^S \mid \boldsymbol{g}^T \boldsymbol{\alpha} = 1\}$ is non-empty in the relative interior of $\mathcal{A}^S$. So there exists $\eta > 0$ such that E_η is non-empty. Without loss of generality, assume E_1 non-empty (can rescale with any positive constant other than 1), define

$$d_j = \min_{\boldsymbol{\alpha}' \in E_1} \|\boldsymbol{\alpha}_j - \boldsymbol{\alpha}'\|.$$

By convexity (see Fig. 6), for $0 < \eta \leqslant 1$,

$$\min_{\boldsymbol{\alpha}' \in E_\eta} \|\boldsymbol{\alpha}_j - \boldsymbol{\alpha}'\| \leqslant \eta \min_{\boldsymbol{\alpha}' \in E_1} \|\boldsymbol{\alpha}_j - \boldsymbol{\alpha}'\| = \eta d_j$$

For any $\boldsymbol{\alpha} \in E_0$, it lies within one of the k balls, say $\boldsymbol{\alpha}_j$. We have

$$\begin{aligned}\min_{\boldsymbol{\alpha}' \in E_\eta} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\| &\leqslant \min_{\boldsymbol{\alpha}' \in E_\eta} [\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_j\| + \|\boldsymbol{\alpha}_j - \boldsymbol{\alpha}'\|] \\ &= \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_j\| + \min_{\boldsymbol{\alpha}' \in E_\eta} \|\boldsymbol{\alpha}_j - \boldsymbol{\alpha}'\| \\ &\leqslant \frac{\gamma}{2} + \eta d_j\end{aligned}$$

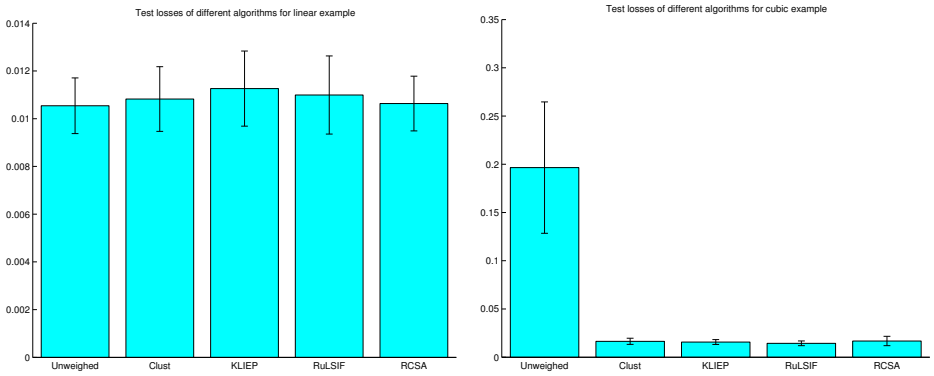
Since the k balls altogether cover E_0, for all $\boldsymbol{\alpha} \in E_0$, when $\eta \leqslant \frac{\gamma}{2 \max_{1 \leqslant j \leqslant k} d_j}$,

$$\min_{\boldsymbol{\alpha}' \in E_\eta} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\| \leqslant \frac{\gamma}{2} + \eta \max_{1 \leqslant j \leqslant k} d_j \leqslant \frac{\gamma}{2} + \frac{\gamma}{2} = \gamma$$

Hence

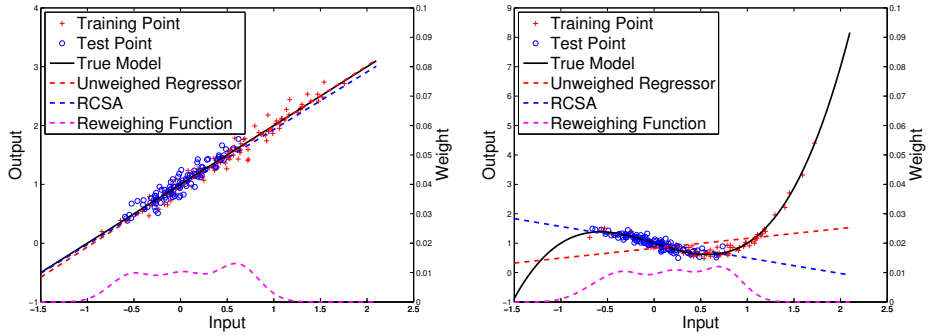
$$\max_{\boldsymbol{\alpha} \in E_0} \min_{\boldsymbol{\alpha}' \in E_\eta} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\| \leqslant \gamma$$

whenever $\eta \leqslant \min(1, \gamma/(2 \max_{1 \leqslant j \leqslant k} d_j))$. The argument for $\eta < 0$ is symmetric. Therefore $g(\eta) \to 0$ as $\eta \to 0$. □



(a) Linear example. (b) Cubic example.

Figure 7. Test losses of different reweighing algorithms for linear and cubic models.



(a) Linear example. (b) Cubic example.

Figure 8. Reweighing for different models. The model class is correctly specified in (a), while is noticeably misspecified in (b).

Appendix B

This appendix shows the experimental results of covariate shift scenario for the synthetic data in Section 5.1. Following the setting in Section 5.1, we generate 100 training points where $x_{tr} \sim \mathcal{N}(0.5, 0.5^2)$ and 100 test points where $x_{te} \sim \mathcal{N}(0, 0.3^2)$. This scenario is adapted from Shimodaira (2000). The performance of both linear model and cubic model are investigated for this covariate shift scenario. Figure 7 reports the test losses (mean and one standard deviation as errorbar) of different reweighing algorithms over 10 trials. Obviously, reweighing is relatively effective when there is no dominant strategy in the underlining model class (Figure 7(b)), compared to the relatively well specified case (Figure 7(a)), where reweighing does not reduce the test loss significantly. To see how reweighing behaves for different model classes, Figure 8 provide one trial of the experiment. Although in both cases, the learning procedure focus on a subset of training points, reweighing is more influential in the cubic example. As the model class is well specified in Figure 8(a), learning from a subset of training points can still recover the global structure of the true model. However, in Figure 8(b), focusing on a subset of training points is more likely to recover local structure of the true model in the test region, and thus the reweighed model performs better in the test region compared to the unweighed model.