JMLR: Workshop and Conference Proceedings 30:1–16, 2015 ACML 2015

# Continuous Target Shift Adaptation in Supervised Learning

**Tuan Duong Nguyen** NTDUONG268@GMAIL.COM
*Tokyo Institute of Technology, Tokyo, 152-8550, Japan.*

**Marthinus Christoffel du Plessis** CHRISTO@MS.K.U-TOKYO.AC.JP
**Masashi Sugiyama** SUGI@K.U-TOKYO.AC.JP
*The University of Tokyo, Tokyo, 113-0033, Japan.*

## Abstract

Supervised learning in machine learning concerns inferring an underlying relation between covariate $\boldsymbol{x}$ and target $y$ based on training covariate-target data. It is traditionally assumed that training data and test data, on which the generalization performance of a learning algorithm is measured, follow the same probability distribution. However, this standard assumption is often violated in many real-world applications such as computer vision, natural language processing, robot control, or survey design, due to intrinsic non-stationarity of the environment or inevitable sample selection bias. This situation is called *dataset shift* and has attracted a great deal of attention recently. In the paper, we consider supervised learning problems under the *target shift* scenario, where the target marginal distribution $p(y)$ changes between the training and testing phases, while the target-conditioned covariate distribution $p(\boldsymbol{x}|y)$ remains unchanged. Although various methods for mitigating target shift in classification (a.k.a. *class prior change*) have been developed so far, few methods can be applied to continuous targets. In this paper, we propose methods for continuous target shift adaptation in regression and conditional density estimation. More specifically, our contribution is a novel importance weight estimator for continuous targets. Through experiments, the usefulness of the proposed method is demonstrated.

**Keywords:** Target shift, importance weighting, $L^2$-distance

## 1. Introduction

The objective of supervised learning is to infer the underlying relation between covariate (input, predictor) $\boldsymbol{x}$ and target (output, label) $y$ from a training dataset consisting of paired covariate-target data. It is traditionally assumed that the training data and test data, on which the generalization performance of a learning algorithm is measured, follow the same probability distribution (Hastie et al., 2009). However, this standard assumption is often violated, i.e., the training data distribution $p_{\mathrm{tr}}(\boldsymbol{x}, y)$ is different from the test data distribution $p_{\mathrm{te}}(\boldsymbol{x}, y)$ in many real-world applications such as natural language processing, speech recognition, robot control, computer vision or survey data analysis, due to inevitable sample selection bias or intrinsic non-stationarity of the environment (Heckman, 1979; Quinonero-Candela et al., 2009; Sugiyama and Kawanabe, 2012). For instance, in natural language processing, part-of-speech taggers, parsers, and text classifiers are trained on a number of annotated training sets, but might be applied to texts from different genres or styles; in social surveillance, the survey sample might not be representative of the whole population of survey targets due to the biased nature of the sampling procedure.

Such a situation is called *dataset shift* and has attracted a great deal of attention recently (Quinonero-Candela et al., 2009).

If the datasets differ arbitrarily between training and test datasets, learning would not be possible. Therefore, in order to enable learning, it is necessary to make a reasonable assumption on the relation between the test and training distributions. In this paper, we address the situation called *target shift* (Zhang et al., 2013), in which the target-marginal distributions $p(y)$ differ between training and test data, while the target-conditional covariate distribution $p(\boldsymbol{x}|y)$ remains unchanged, i.e., $p_{\mathrm{tr}}(y) \neq p_{\mathrm{te}}(y)$ but $p_{\mathrm{tr}}(\boldsymbol{x}|y) = p_{\mathrm{te}}(\boldsymbol{x}|y)$. Note that under this target shift situation, the covariate-marginal distribution $p(\boldsymbol{x})$ and covariate-conditional distribution $p(y|\boldsymbol{x})$ are generally different between training and testing phases due to the shift in $p(y)$. Several methods have been proposed to handle target shift for categorical target $y$, which is also known as *class-prior change* (du Plessis and Sugiyama, 2012; Zhang et al., 2013; Iyer et al., 2014). However, few methods can be applied to continuous target $y$. Therefore, the main focus of this paper will be on the continuous target shift situation, which is often encountered in practice; for instance, prior probability shift (Storkey, 2009), anti-causal regression (Schölkopf et al., 2012), endogenous stratified sampling in econometrics (Manski and Lerman, 1977), or sample selection bias in social surveys (Heckman, 1979).

Motivated by the idea of importance sampling for covariate shift adaptation (Shimodaira, 2000; Sugiyama and Kawanabe, 2012), a similar instance reweighting technique can be employed to handle the continuous target shift situation. Thus, the key technical challenge is how to estimate the importance weight $p_{\mathrm{te}}(y)/p_{\mathrm{tr}}(y)$ for continuous target $y$. In this paper, we propose a novel estimator of the importance weight $p_{\mathrm{te}}(y)/p_{\mathrm{tr}}(y)$ under a semi-supervised setting, where labeled training data and unlabeled test data are given. Moreover, we demonstrate its usefulness in two supervised learning tasks: regression and conditional density estimation under continuous target shift.

In Section 2, we formulate the problem of supervised learning under target shift and show that regression and conditional density estimation can be solved via importance weighting. In Section 3, we propose a novel importance weight estimator and discuss its relations with related work in Section 4. We evaluate the performance of the proposed method through experiments in Section 5 and give a conclusion in Section 6.

## 2. Supervised Learning under Target Shift

In this section, we consider a supervised learning setting under target shift, and provide importance weighted adaptation methods for supervised learning tasks under continuous target shift.

### 2.1. Problem Formulation

Supervised learning problems are mainly concerned with estimating an unknown relation between the covariate (input) and target (output) from a set of labeled training samples. The covariate-target relation is denoted as $f(\boldsymbol{x}) : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the domains of the covariate and target. This covariate-target relation should be estimated based on labeled training samples $\mathcal{D}_{\mathrm{tr}} = \{(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^{n}$. The goal of supervised learning is to

find $f$ which generalizes well on unseen test data $\mathcal{D}_{\mathrm{te}} = \{(\boldsymbol{x}'_j, y'_j) \in \mathcal{X} \times \mathcal{Y}\}_{j=1}^{n'}$; for instance, accurately estimating the unknown target values $\{y'_j\}_{j=1}^{n'}$ from $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ in regression.

We assume that training data and test data are drawn independently and identically (i.i.d.) from underlying joint distributions with probability densities $p_{\mathrm{tr}}(\boldsymbol{x}, y)$ and $p_{\mathrm{te}}(\boldsymbol{x}, y)$, respectively. It is commonly assumed in standard supervised learning that $p_{\mathrm{tr}}(\boldsymbol{x}, y) = p_{\mathrm{te}}(\boldsymbol{x}, y)$ (Hastie et al., 2009). In the paper, however, we consider the supervised learning problems in a more practical setting where the training data and test data have different distributions, i.e., $p_{\mathrm{te}}(\boldsymbol{x}, y) \neq p_{\mathrm{tr}}(\boldsymbol{x}, y)$. In this scenario, we need to make appropriate assumptions on the relatedness between training and test distributions, otherwise nothing about the test domain can be predicted from training data. In this paper, we assume the *target shift* (TarS) assumption (Zhang et al., 2013):

$$(\mathcal{A}_1) : p_{\mathrm{te}}(\boldsymbol{x}|y) = p_{\mathrm{tr}}(\boldsymbol{x}|y) \quad \text{but} \quad p_{\mathrm{te}}(y) \neq p_{\mathrm{tr}}(y). \tag{1}$$

As a result of the Bayes rule, we can confirm that a shift in the target-marginal distribution $p(y)$ causes changes in data joint distributions as well as the covariate-target relation in general.

### 2.2. Ordinary Empirical Risk Minimization

Let us consider a parametric model $g(\boldsymbol{x}; \boldsymbol{\theta})$ for the relation function $f$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^b$ for some $b > 0$. Note that the following discussion can be applied to non-parametric models as well. Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ = [0, +\infty)$ be a loss function, where $L(g(\boldsymbol{x}; \boldsymbol{\theta}), y)$ measures the discrepancy between true target value $y$ at input point $\boldsymbol{x}$ and its estimate $g(\boldsymbol{x}; \boldsymbol{\theta})$. The risk of an estimator $g(\boldsymbol{x}; \boldsymbol{\theta})$ over test data (a.k.a. the generalization error) is given as

$$R(\boldsymbol{\theta}) = \iint L(g(\boldsymbol{x}; \boldsymbol{\theta}), y) p_{\mathrm{te}}(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x} \mathrm{d}y.$$

Although $p_{\mathrm{te}}(\boldsymbol{x}, y)$ is often unknown in practice, the generalization error can be approximated by the *empirical error* calculated from training data:

$$\widehat{R}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} L(g(\boldsymbol{x}_i; \boldsymbol{\theta}), y_i).$$

When there is no shift in the data distribution, i.e., $p_{\mathrm{te}}(\boldsymbol{x}, y) = p_{\mathrm{tr}}(\boldsymbol{x}, y)$, a standard method called *empirical risk minimization* (ERM) (Vapnik, 1998), in which the empirical error is minimized, can be employed to yield a consistent estimator of the relation $f$.

### 2.3. importance weighted Empirical Risk Minimization

However, due to a difference in distributions, standard ERM is not consistent in the target shift situation. In the following, we show that a common *importance weighting* technique (Sugiyama and Kawanabe, 2012) can be employed to compensate for the shift in distributions. More specifically, the generalization error can be expressed as follows:

$$R(\boldsymbol{\theta}) = \iint L(g(\boldsymbol{x}; \boldsymbol{\theta}), y) p_{\mathrm{te}}(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x} \mathrm{d}y = \iint \left( \frac{p_{\mathrm{te}}(y)}{p_{\mathrm{tr}}(y)} \right) L(g(\boldsymbol{x}; \boldsymbol{\theta}), y) p_{\mathrm{tr}}(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x} \mathrm{d}y.$$

Here the last equality comes from:

$$p_{\mathrm{te}}(\boldsymbol{x}, y) = p_{\mathrm{tr}}(\boldsymbol{x}|y)p_{\mathrm{te}}(y) = \left(\frac{p_{\mathrm{te}}(y)}{p_{\mathrm{tr}}(y)}\right) p_{\mathrm{tr}}(\boldsymbol{x}, y), \tag{2}$$

which is due to the target shift assumption (1). This gives rise the following assumption:

$$(\mathcal{A}_2) : w(y) = \frac{p_{\mathrm{te}}(y)}{p_{\mathrm{tr}}(y)} < \infty \text{ for all } y. \tag{3}$$

Using the importance weight $w(y) = p_{\mathrm{te}}(y)/p_{\mathrm{tr}}(y)$, we can approximate the generalization error with the following importance weighted empirical error:

$$\widehat{R}_w(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} w(y_i)L(g(\boldsymbol{x}_i; \boldsymbol{\theta}), y_i).$$

Note that

$$\begin{aligned}\mathop{\mathbb{E}}_{(\boldsymbol{x}_i, y_i)\sim p_{\mathrm{tr}}(\boldsymbol{x}, y)} [w(y_i)L(g(\boldsymbol{x}_i; \boldsymbol{\theta}), y_i)] &= \iint L(g(\boldsymbol{x}; \boldsymbol{\theta}), y)w(y)p_{\mathrm{tr}}(\boldsymbol{x}, y)\mathrm{d}\boldsymbol{x}\mathrm{d}y \\ &= \iint L(g(\boldsymbol{x}; \boldsymbol{\theta}), y)p_{\mathrm{te}}(\boldsymbol{x}, y)\mathrm{d}\boldsymbol{x}\mathrm{d}y = R(\boldsymbol{\theta}),\end{aligned}$$

where the last equality comes from (2). Therefore, the importance weighted empirical error may be used to estimate the generalization error.

Since many methods for conditional density estimation and regression can be expressed in terms of loss functions, importance weighting can be readily applied to adapt these methods to target shift. Below we discuss importance weighted least-squares conditional density estimation (IWLSCDE) and importance weighted least-squares regression (IWLS).

### 2.4. Importance Weighted Conditional Density Estimation

We consider the problem of estimating the conditional density $p(y|\boldsymbol{x})$ under the target shift situation, where $\boldsymbol{x} \in \mathbb{R}^d$ denotes $d$-dimensional input (covariate) and $y \in \mathbb{R}$ denotes output (target). Note that we consider one dimensional output here for simplicity, but the technique discussed below can be applied to multi-dimensional output as well.

Suppose that we are given labeled training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ drawn independently from a joint probability distribution with density $p_{\mathrm{tr}}(\boldsymbol{x}, y)$, and unlabeled test data $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ drawn independently from another probability distribution with density $p_{\mathrm{te}}(\boldsymbol{x}) = \int p_{\mathrm{te}}(\boldsymbol{x}, y)\mathrm{d}y$. We consider the target shift scenario, i.e.,

$$p_{\mathrm{te}}(y) \neq p_{\mathrm{tr}}(y), \text{ but } p_{\mathrm{te}}(\boldsymbol{x}|y) = p_{\mathrm{tr}}(\boldsymbol{x}|y).$$

Note that the input-conditional densities are generally different in this situation, i.e., $p_{\mathrm{te}}(y|\boldsymbol{x}) \neq p_{\mathrm{tr}}(y|\boldsymbol{x})$. Thus, naive conditional density estimation from training data will not generalize well to test data since they follow a different distribution. Our goal is to estimate a conditional density $p_{\mathrm{te}}(y|\boldsymbol{x})$ for test data following $p_{\mathrm{te}}(\boldsymbol{x}, y)$.

Here, we consider a variation of *least-squares conditional density estimation* (Sugiyama et al., 2010) called *importance weighted least-squares conditional density estimation* (IWLSCDE), which employs the importance for instance reweighting.

The conditional density $p_{\mathrm{te}}(y|\boldsymbol{x})$ is modeled by a non-parametric Gaussian kernel model:

$$r(\boldsymbol{x}, y) = \sum_{l=1}^{n} \theta_l \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_l\|^2}{2\sigma^2}\right) \exp\left(-\frac{(y - y_l)^2}{2\sigma^2}\right).$$

If $n$ is too large, only a subset of training samples may be used as Gaussian centers. The parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ is determined by minimizing the following squared error:

$$\begin{aligned}
&\frac{1}{2} \iint \left(r(\boldsymbol{x}, y) - p_{\mathrm{te}}(y|\boldsymbol{x})\right)^2 p_{\mathrm{te}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}y \\
&= \frac{1}{2} \int \left(\int r(\boldsymbol{x}, y)^2 \mathrm{d}y\right) p_{\mathrm{te}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \iint r(\boldsymbol{x}, y) w(y) p_{\mathrm{tr}}(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x}\mathrm{d}y + C' \\
&= \frac{1}{2} \iint \left(\int r(\boldsymbol{x}, y)^2 \mathrm{d}y\right) w(y) p_{\mathrm{tr}}(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x}\mathrm{d}y - \iint r(\boldsymbol{x}, y) w(y) p_{\mathrm{tr}}(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x}\mathrm{d}y + C', \qquad (4)
\end{aligned}$$

where $C'$ is a constant and thus can be ignored and $w(y) = p_{\mathrm{te}}(y)/p_{\mathrm{tr}}(y)$. Note that the above derivation follows from (2).

Replacing the expectations by the sample averages and including an $\ell_2$-regularizer, we arrive at the following optimization problem:

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\frac{1}{2}\boldsymbol{\theta}^\top \widehat{\boldsymbol{H}} \boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\theta} + \frac{\lambda}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta}\right] = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}},$$

where $\lambda \geq 0$ is the regularization parameter, $\boldsymbol{I}$ denotes the identity matrix, and

$$\begin{aligned}
\widehat{H}_{l,l'} &= \frac{\sigma\sqrt{\pi}}{n} \sum_{i=1}^{n} w(y_i) \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_l\|^2 + \|\boldsymbol{x}_i - \boldsymbol{x}_{l'}\|^2}{2\sigma^2}\right) \exp\left(-\frac{(y_l - y_{l'})^2}{4\sigma^2}\right), \\
\widehat{h}_l &= \frac{1}{n} \sum_{i=1}^{n} w(y_i) \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_l\|^2}{2\sigma^2}\right) \exp\left(-\frac{(y_i - y_l)^2}{2\sigma^2}\right).
\end{aligned}$$

A similar derivation of $\widehat{\boldsymbol{H}}$ and $\widehat{\boldsymbol{h}}$ can be found in Sugiyama et al. (2010), and is therefore omitted for brevity. Hyper-parameters for the above model can be selected by using importance weighted cross validation (Sugiyama et al., 2007; Sugiyama and Kawanabe, 2012).

### 2.5. Importance Weighted Regression

We consider the semi-supervised regression problem under target shift. More specifically, given training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ drawn independently from a joint probability distribution with density $p_{\mathrm{tr}}(\boldsymbol{x}, y)$ and unlabeled test data $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ drawn independently from another probability distribution with density $p_{\mathrm{te}}(\boldsymbol{x}) = \int p_{\mathrm{te}}(\boldsymbol{x}, y)\mathrm{d}y$, we learn the conditional expectation:

$$f(\boldsymbol{x}) = \int y p_{\mathrm{te}}(y|\boldsymbol{x})\mathrm{d}y.$$

The conditional expectation $f(\boldsymbol{x})$ is modeled as

$$g(\boldsymbol{x}) = \sum_{l=1}^{n} \eta_l \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_l\|^2}{2\sigma_{\mathrm{x}}^2}\right).$$

The parameter $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$ is determined by the *importance weighted least-squares* (IWLS) (Sugiyama and Kawanabe, 2012):

$$\begin{aligned}
&\frac{1}{2}\int (g(\boldsymbol{x}) - f(\boldsymbol{x}))^2 p_{\mathrm{te}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\int g(\boldsymbol{x})^2 p_{\mathrm{te}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int g(\boldsymbol{x}) \left[\int y p_{\mathrm{te}}(y|\boldsymbol{x})\mathrm{d}y\right] p_{\mathrm{te}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + C'' \\
&= \frac{1}{2}\iint g(\boldsymbol{x})^2 w(y) p_{\mathrm{tr}}(\boldsymbol{x}, y)\mathrm{d}\boldsymbol{x}\mathrm{d}y - \iint y g(\boldsymbol{x}) w(y) p_{\mathrm{tr}}(\boldsymbol{x}, y)\mathrm{d}\boldsymbol{x}\mathrm{d}y + C'',
\end{aligned}$$

where $C''$ is a constant and thus can be ignored. Note that the last equality follows from (2). Replacing the expectations by the sample averages and including an $\ell_2$-regularizer, we arrive at the following optimization problem:

$$\widehat{\boldsymbol{\eta}} = \operatorname*{argmin}_{\boldsymbol{\eta}} \left[\frac{1}{2}\boldsymbol{\eta}^\top \widehat{\boldsymbol{G}}\boldsymbol{\eta} - \widehat{\boldsymbol{g}}^\top \boldsymbol{\eta} + \frac{\gamma}{2}\boldsymbol{\eta}^\top \boldsymbol{\eta}\right] = (\widehat{\boldsymbol{G}} + \gamma \boldsymbol{I})^{-1}\widehat{\boldsymbol{g}},$$

where $\gamma \geq 0$ is the regularization parameter and

$$\widehat{G}_{l,l'} = \frac{1}{n}\sum_{i=1}^{n} w(y_i) \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_l\|^2 + \|\boldsymbol{x}_i - \boldsymbol{x}_{l'}\|^2}{2\sigma_{\mathrm{x}}^2}\right), \quad \widehat{g}_l = \frac{1}{n}\sum_{i=1}^{n} y_i w(y_i) \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_l\|^2}{2\sigma_{\mathrm{x}}^2}\right).$$

The Gaussian kernel width $\sigma_{\mathrm{x}}$ and the regularization parameter $\gamma$ can be chosen by importance weighted cross validation (Sugiyama et al., 2007).

## 3. Importance Weight Estimation by Distribution Matching

As shown in the previous section, the target shift adaptation techniques require the values of the importance weight at the training points $\{w(y_i)\}_{i=1}^n$, which are usually unknown in practice. Moreover, few methods have been proposed to estimate the importance weight under *continuous* target shift situations. In this section, we propose a novel method to estimate the importance weight for continuous target $y$.

### 3.1. Problem Formulation

In practice, the weighting function $w(y)$ is unknown. Given labeled training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and unlabeled test data $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ drawn independently from the probability distributions with densities $p_{\mathrm{tr}}(\boldsymbol{x}, y)$ and $p_{\mathrm{te}}(\boldsymbol{x}) = \int p_{\mathrm{te}}(\boldsymbol{x}, y)\mathrm{d}y$ respectively, our aim is to estimate the importance weight function, $w(y) = p_{\mathrm{te}}(y)/p_{\mathrm{tr}}(y)$. Note that labeled data from the test distribution $\{y'_j\}_{j=1}^{n'}$ are not available under the setting, which makes the estimation problem non-trivial.

Denote a model for the unknown importance weight $w(y)$ as $s(y)$. We can then model the test input density as

$$q_s(\boldsymbol{x}) = \int s(y) p_{\mathrm{tr}}(\boldsymbol{x}, y) \mathrm{d}y,$$

where notation $q_s(\boldsymbol{x})$ emphasizes that it depends on $s(y)$.

The importance weight $w(y)$ can be estimated by matching $q_s(\boldsymbol{x})$ to the test distribution $p_{\mathrm{te}}(\boldsymbol{x})$ under a discrepancy measure $\mathbb{D}$. This is justified by the target shift assumption ($\mathcal{A}_1$) and the definition of the importance weight $w(y)$:

$$\begin{aligned} p_{\mathrm{te}}(\boldsymbol{x}) &= \int p_{\mathrm{te}}(\boldsymbol{x}, y) \mathrm{d}y = \int p_{\mathrm{te}}(\boldsymbol{x}|y) p_{\mathrm{te}}(y) \mathrm{d}y \\ &= \int p_{\mathrm{tr}}(\boldsymbol{x}|y) p_{\mathrm{tr}}(y) w(y) \mathrm{d}y = \int w(y) p_{\mathrm{tr}}(\boldsymbol{x}, y) \mathrm{d}y. \end{aligned}$$

The optimization problem of minimizing the discrepancy $\mathbb{D}$ between $q_s(\boldsymbol{x})$ and $p_{\mathrm{te}}(\boldsymbol{x})$ is:

$$\begin{aligned} \widehat{w} \quad &= \underset{s}{\operatorname{argmin}} \, \mathbb{D}\left(q_s(\boldsymbol{x}), p_{\mathrm{te}}(\boldsymbol{x})\right) \\ &\quad \text{subject to } s(y) \geq 0 \text{ for all } y \in \mathcal{Y} \text{ and } \int p_{\mathrm{tr}}(y) s(y) \mathrm{d}y = 1, \end{aligned} \tag{5}$$

where the constraints are for non-negativity and normalization.

Ratio-based divergences, e.g., the Kullback-Leibler (KL) divergence, and difference-based distances, e.g., the $L^2$-distance are common discrepancy measures in machine learning and statistics. Among them, $L^2$-distance is a proper distance measure, always bounded as long as each density is bounded and thus stable (Sugiyama et al., 2013a). More importantly, it can be accurately and analytically approximated in a computationally efficient and numerically stable manner via direct density-difference estimation (Sugiyama et al., 2013b). Therefore, $L^2$-distance would be a suitable discrepancy measure for the current problem of estimating the importance weight. For the $L^2$-distance, the optimization problem becomes

$$\begin{aligned} \widehat{w} \quad &= \underset{s}{\operatorname{argmin}} \, L^2\left(q_s(\boldsymbol{x}), p_{\mathrm{te}}(\boldsymbol{x})\right) \\ &\quad \text{subject to } s(y) \geq 0 \text{ for all } y \text{ and } \int p_{\mathrm{tr}}(y) s(y) \mathrm{d}y = 1, \end{aligned} \tag{6}$$

where

$$L^2(q_s(\boldsymbol{x}), p_{\mathrm{te}}(\boldsymbol{x})) = \frac{1}{2} \int \left(q_s(\boldsymbol{x}) - p_{\mathrm{te}}(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}.$$

This optimization problem is convex in $s$ and attains a minimum at $q_s(\boldsymbol{x}) = p_{\mathrm{te}}(\boldsymbol{x})$.

To take advantage of the fact that the above objective function is convex in $s(y)$, we use a linear-in-parameter model:

$$s(y) = \sum_{l=1}^{b} \alpha_l \phi_l(y) = \boldsymbol{\alpha}^\top \boldsymbol{\phi}(y), \tag{7}$$

where $b$ is the number of parameters, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_b)^\top$ are parameters to be learned from data, $\top$ denotes the transpose, and $\boldsymbol{\phi}(y) = (\phi_1(y), \ldots, \phi_b(y))^\top$ are basis functions.

Note that $b$ and $\{\phi_l(y)\}_{l=1}^b$ can depend on the samples $\{y_l\}_{l=1}^n$ and thus non-parametric kernel models can also be represented by (7). In practice, we use the non-parameteric Gaussian kernel model, $\phi_l(y) = \exp\left(-\left(y - y_l\right)^2 / (2\kappa_{\mathrm{y}}^2)\right)$, for $b = n$.

### 3.2. $L^2$-distance Estimation

Since the $L^2$-distance above contains unknown densities $p_{\mathrm{tr}}(\boldsymbol{x}, y)$ and $p_{\mathrm{te}}(\boldsymbol{x})$, it cannot be directly computed. Here, following Sugiyama et al. (2013b), let us model the density difference $q_s(\boldsymbol{x}) - p_{\mathrm{te}}(\boldsymbol{x})$ by a non-parametric Gaussian kernel model:

$$t(\boldsymbol{x}) = \sum_{l=1}^{n+n'} \beta_l \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_l\|^2}{2\kappa_{\mathrm{x}}^2}\right). \tag{8}$$

The parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{n+n'})^\top$ is determined by minimizing the squared error:

$$\frac{1}{2}\int (t(\boldsymbol{x}) - (q_s(\boldsymbol{x}) - p_{\mathrm{te}}(\boldsymbol{x})))^2 \mathrm{d}\boldsymbol{x} = \frac{1}{2}\int t(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} - \int t(\boldsymbol{x}) q_s(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int t(\boldsymbol{x}) p_{\mathrm{te}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + C,$$

where $C$ is a constant and thus can be ignored. Replacing the expectations by the sample averages and including an $\ell_2$-regularizer, we arrive at the following optimization problem:

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \left[\frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{U}\boldsymbol{\beta} - \boldsymbol{\beta}^\top(\widehat{\boldsymbol{V}}\boldsymbol{\alpha} - \widehat{\boldsymbol{u}}) + \frac{\delta}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right] = (\boldsymbol{U} + \delta \boldsymbol{I})^{-1}(\widehat{\boldsymbol{V}}\boldsymbol{\alpha} - \widehat{\boldsymbol{u}}),$$

where $\delta \geq 0$ is the regularization parameter and

$$U_{l,l'} = (\pi\kappa_{\mathrm{x}}^2)^{d/2} \exp\left(-\frac{\|\boldsymbol{x}_l - \boldsymbol{x}_{l'}\|^2}{4\kappa_{\mathrm{x}}^2}\right), \quad \widehat{u}_l = \frac{1}{n'}\sum_{j=1}^{n'} \exp\left(-\frac{\left\|\boldsymbol{x}'_j - \boldsymbol{x}_l\right\|^2}{2\kappa_{\mathrm{x}}^2}\right),$$

$$\widehat{V}_{l,l'} = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_l\|^2}{2\kappa_{\mathrm{x}}^2}\right) \exp\left(-\frac{(y_i - y_{l'})^2}{2\kappa_{\mathrm{y}}^2}\right).$$

With the solution $\widehat{\boldsymbol{\beta}}$, an $L^2$-distance estimator can be obtained as

$$\widehat{L}^2\left(q(\boldsymbol{x}), p_{\mathrm{te}}(\boldsymbol{x})\right) = J(\boldsymbol{\alpha}) + \frac{1}{2}\widehat{\boldsymbol{u}}^\top(\boldsymbol{U} + \delta\boldsymbol{I})^{-1}\widehat{\boldsymbol{u}},$$

where

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^\top\widehat{\boldsymbol{V}}^\top(\boldsymbol{U} + \delta\boldsymbol{I})^{-1}\widehat{\boldsymbol{V}}\boldsymbol{\alpha} - \widehat{\boldsymbol{u}}^\top(\boldsymbol{U} + \delta\boldsymbol{I})^{-1}\widehat{\boldsymbol{V}}\boldsymbol{\alpha}.$$

The Gaussian kernel width $\kappa_{\mathrm{x}}$ and the regularization parameter $\delta$ can be chosen by cross validation (CV) with respect to the squared error.

### 3.3. $L^2$-distance based Importance Weight Estimation

Finally, the parameter $\boldsymbol{\alpha}$ in the importance weight model (7) is learned by minimizing the above $L^2$-distance estimator with an $\ell_2$-regularizer:

$$\widehat{\boldsymbol{\alpha}} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \left[J(\boldsymbol{\alpha}) + \frac{\rho}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}\right]$$
$$\text{subject to} \quad \boldsymbol{\alpha} \geq \boldsymbol{0} \text{ and } \boldsymbol{\alpha}^\top\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\phi}(y_i)\right) = 1,$$

where

$$\boldsymbol{\phi}(y) = \left(\exp\left(-(y-y_1)^2/(2\kappa_{\mathrm{y}}^2)\right), \ldots, \exp\left(-(y-y_n)^2/(2\kappa_{\mathrm{y}}^2)\right)\right)^\top.$$

When all the hyper-parameters $\kappa_{\mathrm{x}}, \delta, \kappa_{\mathrm{y}}$ and $\rho$ included in the objective function above are fixed in advance, the optimization problem becomes a linearly constrained quadratic program and thus can be solved effectively by any off-the-shelf solver. However, it is our aim to objectively optimize all the hyper-parameters based on data. For this purpose, we employ a nested CV procedure as described below.

In the outer CV loop, we iterate over a list of candidate hyper-parameter pairs $\{(\kappa_{\mathrm{y}}, \rho)\}$. For each pair $(\kappa_{\mathrm{y}}, \rho)$, we perform the following iterative algorithm, where $\boldsymbol{\alpha}^{(k)}$ denotes the solution at the $k^{\mathrm{th}}$ step.

1. Start with an initial solution $\boldsymbol{\alpha}^{(0)}$.
2. Given $(\kappa_{\mathrm{y}}, \rho)$, the current solution $\boldsymbol{\alpha}^{(k)}$ and a list of candidate pairs $\{(\kappa_{\mathrm{x}}, \delta)\}$, conduct the inner CV loop to select $\kappa_{\mathrm{x}}$ and $\delta$ in the $L^2$-distance estimator.
3. Given the current solution $\boldsymbol{\alpha}^{(k)}$ and selected hyper-parameters, solve the following quadratically constrained quadratic program (QCQP):

$$\boldsymbol{\alpha}^{(k+1)} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \left[J(\boldsymbol{\alpha}) + \frac{\rho}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}\right]$$
$$\text{subject to } \boldsymbol{\alpha} \geq \mathbf{0},\ \boldsymbol{\alpha}^\top\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\phi}(y_i)\right) = 1 \text{ and } \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(k)}\|^2 \leq \epsilon,$$

   where $\epsilon$ is a small positive step size.
4. Continue Step 2 - Step 3 until convergence.

Instead of iterating until convergence, we may preset the maximum number of iterations to control the running time of the algorithm.

The Gaussian kernel width $\kappa_{\mathrm{y}}$ and the regularization parameter $\rho$ are chosen by the outer CV loop with respect to the estimated $L^2$-distance. We employ the off-the-shelf Gurobi solver[1] to solve the QCQP problem in the latter experiments. We call the proposed method *$L^2$-distance based importance weight estimation* (L2IWE).

## 4. Discussion

In this section, we discuss the relations between the proposed method and related work.

### 4.1. A Kernel Mean Matching Approach

We first give a brief review of a recently proposed method (Zhang et al., 2013), called TarS for estimating the importance weight under the target shift situation.

In machine learning problems, it is desirable to avoid density estimation when comparing distributions. The idea of distribution matching under maximum mean discrepancy (MMD) in a reproducing kernel Hilbert space (RKHS) is to avoid density estimation by matching

1. The solver is available from http://www.gurobi.com.

two distributions based on their kernel mean embedding (Gretton et al., 2006). The kernel mean embedding of the covariate density $p(\boldsymbol{x})$ is a point in the RKHS $\mathcal{H}$ defined as

$$\mu_{\boldsymbol{X}} = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[\psi(\boldsymbol{x})],$$

where $\psi(\boldsymbol{x})$, is a feature map with the corresponding kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \psi(\boldsymbol{x}), \psi(\boldsymbol{x}') \rangle_{\mathcal{H}}$, and $< \cdot, \cdot >_{\mathcal{H}}$ denotes the inner product in $\mathcal{H}$. In order to estimate the importance weight, we match the test covariate density $p_{\mathrm{te}}(\boldsymbol{x})$ with the density

$$q(\boldsymbol{x}) = \int p_{\mathrm{tr}}(\boldsymbol{x}|y) p_{\mathrm{tr}}(y) w(y) \mathrm{d}y.$$

The TarS method employs a Gaussian kernel, i.e., $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/(2\sigma^2)\right)$, where $\sigma$ is the kernel width. This results in an empirical version that is a quadratic program (Zhang et al., 2013, Eq. (5)).

The performance of MMD-based TarS depends on the choice of three tuning hyper-parameters (one of which is the kernel width $\sigma$). Since estimates of the importance weight are available only at $\{y_i\}_{i=1}^n$, there is no reliable way to automatically optimize these hyperparameters, for instance, by cross validation (CV) (Tsuboi et al., 2009; Sugiyama and Kawanabe, 2012). In a similar way to Kanamori et al. (2013), TarS can be extended to its inductive variant for out-of-sample prediction, i.e., the entire importance weight function is estimated. Such a procedure allows two of the hyper-parameters to be estimated via CV.

However, the Gaussian kernel width $\sigma$ may not be appropriately determined by CV since changing the value of $\sigma$ implies changing the RKHS as well as the error metric defined using the corresponding RKHS norm. Since the objective values for different norms are not comparable, CV cannot be performed to appropriately optimize $\sigma$. Although a popular heuristic to choose $\sigma$ is to use the median distance between samples as the Gaussian width $\sigma$ (Song et al., 2007), there seems to be no strong justification for this heuristic.

### 4.2. Categorical Target Shift Adaptation

Various methods have been proposed to handle the target shift situation for categorical targets, also known as *class prior change* (du Plessis and Sugiyama, 2012; Zhang et al., 2013; Iyer et al., 2014). We discuss relations between our proposed method and some state-of-the-art methods for categorical target shift.

As discussed above, the difficulties in hyper-parameter tuning make these MMD-based methods (Zhang et al., 2013; Iyer et al., 2014) less useful in practice. Meanwhile, du Plessis and Sugiyama (2014) estimated the test target distribution $p_{\mathrm{te}}(y)$ by matching distributions under some divergence measures, e.g., Pearson divergence. For categorical target $y \in \{1, 2, \ldots, c\}$, the training target distribution $p_{\mathrm{tr}}(y = t)$ can be naively estimated from training data $\{y_i\}_{i=1}^n$ as $\widehat{p}_{\mathrm{tr}}(y = t) = n_t/n$, where $n_t$ is the number of training instances whose target values are equal to $t$. Therefore, it suffices to estimate only the test target distribution $p_{\mathrm{te}}(y)$ instead of the importance weight function $w(y)$ as a whole. In du Plessis and Sugiyama (2014), the test target distribution is modeled with a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^c$, where $\theta_t = p_{\mathrm{te}}(y = t)$, and the parameter $\boldsymbol{\theta}$ is learned by distribution

matching as

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}^\top \mathbf{1}=1, \boldsymbol{\theta} \geq 0}{\operatorname{argmin}} \; \mathbb{D}\left( p_{\mathrm{te}}(\boldsymbol{x}), \sum_{t=1}^{c} \theta_t \; p_{\mathrm{tr}}(\boldsymbol{x}|y=t) \right). \tag{9}$$

Our proposed method thus can be regarded as a natural extension of du Plessis and Sugiyama (2014) for the case of continuous targets. In the case of a categorical $y$, we can obtain samples from $p_{\mathrm{tr}}(\boldsymbol{x}|y)$ by selecting the $y$-labeled samples in the labeled training dataset. However, when $y$ is continuous, it is not possible to sample directly from $p(\boldsymbol{x}|y)$. This makes the extension from the categorical to continuous cases non-trivial.

## 5. Experiments

In this section, we experimentally evaluate the performance of the proposed L2IWE method, particularly in comparison with TarS (Zhang et al., 2013)[2].

### 5.1. Illustrative Examples

Let us consider the following one-dimensional toy problem in the target shift setting:

$$p_{\mathrm{tr}}(y) = 0.4\, \mathcal{N}(y; 1, 1.5^2) + 0.6\, \mathcal{N}(y; 2.5, 0.5^2), \quad p_{\mathrm{te}}(y) = \mathcal{N}(y; 2.5, 0.5^2),$$

where $\mathcal{N}(y; \mu, \sigma^2)$ denotes the Gaussian density with mean $\mu$ and variance $\sigma^2$ with respect to $y$. We first draw $n = 300$ samples $\{y_i\}_{i=1}^{n}$ from $p_{\mathrm{tr}}(y)$ and $n' = 300$ samples $\{y'_j\}_{j=1}^{n'}$ from $p_{\mathrm{te}}(y)$. Then training inputs $\{\boldsymbol{x}_i\}_{i=1}^{n}$ are generated as $x_i = y_i + 3 + \epsilon_i$, where the noise $\{\epsilon_i\}_{i=1}^{n}$ is independently drawn following $\mathcal{N}(\epsilon; 0, 1.5^2)$. Test inputs $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ are also generated from $\{y'_j\}_{j=1}^{n'}$ in the same way.

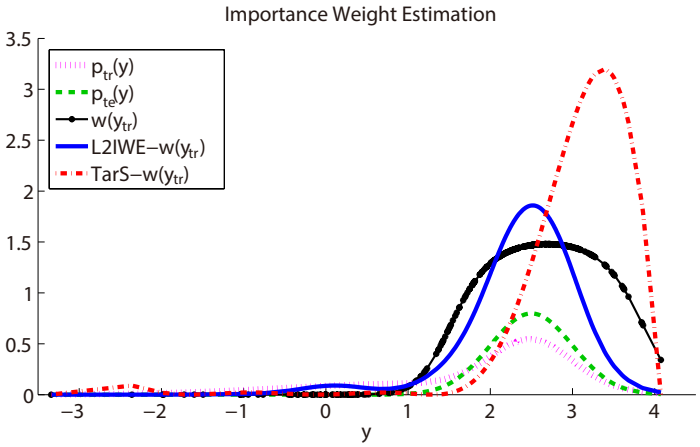
Given labeled training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ and unlabeled test data $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$, the task is to estimate the importance weight values at training target points $\{w(y_i)\}_{i=1}^{n}$. We run the experiments 100 times for the proposed L2IWE method and the TarS method, and evaluate the quality of the importance weight estimates $\{\widehat{w}(y_i)\}_{i=1}^{n}$ by the *normalized mean squared error* (NMSE):

$$\mathrm{NMSE} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\widehat{w}(y_i)}{\sum_{i'=1}^{n} \widehat{w}(y_{i'})} - \frac{w(y_i)}{\sum_{i'=1}^{n} w(y_{i'})} \right)^2.$$

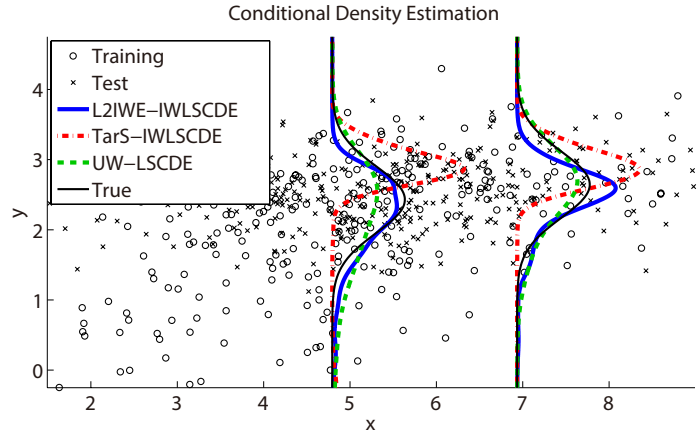
The average (and standard error) of NMSE over 100 runs for L2IWE and TarS are $3.02(\pm 0.3) \times 10^6$ and $4.71(\pm 0.58) \times 10^6$. According to a t-test with significance level 5%, L2IWE gives a significantly more accurate estimate of the importance weight than TarS.

The average computation time of L2IWE and TarS (over 100 runs) is 8929.7s and 60.56s respectively[3]. L2IWE is slower than TarS in two orders of magnitude, which is due to the cross validation of $(\kappa_{\mathrm{x}}, \delta)$ in $L^2$-distance estimation. We note that, without cross validation, the computation time of L2IWE is in the same order as that of TarS. Figure 1(*a*) shows an exemplary example of importance weight estimation by L2IWE and TarS.

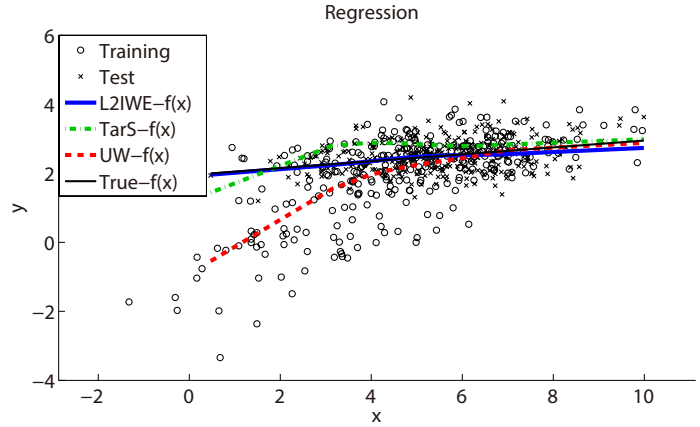
2. The program code is available at http://people.tuebingen.mpg.de/kzhang/Code-TarS.zip. The hyper-parameters are chosen according to the supplementary material available at http://jmlr.org/proceedings/papers/v28/zhang13d-supp.pdf.
3. Simulations were performed on a PC with an Intel Xeon E5-2450 2.1GHz processor.




(*a*) An illustrative example of importance weight estimation for toy data.




(*b*) Target shift adaptation in conditional density estimation for toy data.




(*c*) Target shift adaptation in regression for toy data.

Figure 1: Illustrative examples

Table 1: Average and standard error of MSE in conditional density estimation (CDE) and regression (REG) over 100 trials for toy data. Bold face indicates significant difference by t-test with significance level 5%. Note that MSE for CDE ignores a positive constant so the MSE value can be negative.

| | L2IWE | TarS | Unweighted |
|---|---|---|---|
| CDE | $\mathbf{-0.28 \pm 0.02}$ | $-0.25 \pm 0.04$ | $-0.25 \pm 0.01$ |
| REG | $\mathbf{0.28 \pm 0.09}$ | $0.31 \pm 0.07$ | $0.50 \pm 0.07$ |

We further apply the estimated importance weight values $\{\widehat{w}_i\}_{i=1}^n$ to target shift adaptation in conditional density estimation and regression. Examples of target shift adaptation are illustrated in Figure 1($b$) and Figure 1($c$), and the average and the standard error of MSE in conditional density estimation and regression over 100 trials are summarized in Table 1[4], where a prefix "UW-" or "Unweighted" means that no importance weight is used. The results show that the proposed method is promising.

### 5.2. Target Shift Adaptation for Benchmark Datasets

Next, we evaluate the performance of importance weight estimators on benchmark datasets which are collected from mldata.org[5] and DELVE[6].

Each dataset consists of input/output pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$. We set the number of training samples and test samples to $n = 300$ and $n' = 300$, respectively. We normalize all the output samples into [0,1] and select the test samples $\{(\boldsymbol{x}'_j, y'_j)\}_{j=1}^{n'}$ from the data pool through the following biased sampling scheme. We first randomly shuffle the data pool at the beginning of each trial of the experiments. A sample $(\boldsymbol{x}_k, y_k)$ is then randomly chosen from the pool and accepted if $y_k \in [a, b]$ for $0 < a < b < 1$. We remove the sample from the pool and repeat this procedure until we accept $n'$ samples. We choose the training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ uniformly from the rest. Note that we only use labeled training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and unlabeled test samples $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ for training conditional density estimators and regressors. The test output values $\{y'_j\}_{j=1}^{n'}$ are used only for evaluating the test performance.

The average and the standard error of MSE in conditional density estimation and regression over 100 trials for the benchmark datasets are summarized in Table 2 and Table 3, respectively. They show that the proposed method is still promising for the benchmark datasets.

For the conditional density estimation task, Table 2 shows that the proposed method L2IWE outperforms TarS and Unweighted on 7 out of 10 datasets. Meanwhile, Table 3 shows that L2IWE performs better than TarS and Unweighted on 8 out of 10 datasets.

4. In the case of regression, we computed MSE in a standard way for test data. However, in the case of conditional density estimation, we ignored a positive constant term which cannot be computed only from test data; see Eq. (4)) for details. For this reason, MSE for conditional density estimation often takes a negative value.
5. http://mldata.org/
6. http://www.cs.utoronto.ca/~delve/

Table 2: Average and standard error of MSE (without the positive constant) in conditional density estimation over 100 trials for benchmark datasets. Bold face indicates significant difference by t-test with significance level 5%.

| Data | MSE/test error ± std.error | | |
|---|---|---|---|
| | L2IWE | TarS | Unweighted |
| puma8nh | $\mathbf{-0.85 \pm 0.01}$ | $-0.77 \pm 0.01$ | $\mathbf{-0.84 \pm 0.01}$ |
| kin8nm | $\mathbf{-1.09 \pm 0.01}$ | $-1.00 \pm 0.01$ | $-0.94 \pm 0.01$ |
| kin8nh | $\mathbf{-1.07 \pm 0.01}$ | $-0.96 \pm 0.02$ | $-0.96 \pm 0.02$ |
| kin8fm | $\mathbf{-1.61 \pm 0.01}$ | $\mathbf{-1.62 \pm 0.02}$ | $-1.50 \pm 0.01$ |
| kin8fh | $\mathbf{-1.42 \pm 0.02}$ | $\mathbf{-1.44 \pm 0.02}$ | $-1.24 \pm 0.02$ |
| CA Housing | $\mathbf{-1.16 \pm 0.02}$ | $-0.75 \pm 0.08$ | $-0.94 \pm 0.01$ |
| elevators | $\mathbf{-3.08 \pm 0.04}$ | $-1.05 \pm 0.12$ | $\mathbf{-3.11 \pm 0.03}$ |
| delta ailerons | $\mathbf{-5.13 \pm 0.05}$ | $-4.77 \pm 0.06$ | $-4.91 \pm 0.05$ |
| abalone | $\mathbf{-2.10 \pm 0.02}$ | $-1.99 \pm 0.02$ | $\mathbf{-2.08 \pm 0.02}$ |
| housing | $-1.30 \pm 0.02$ | $\mathbf{-1.66 \pm 0.02}$ | $-1.07 \pm 0.02$ |

Table 3: Average and standard error of MSE in regression over 100 trials for benchmark datasets. All values are multiplied with $10^2$. Bold face indicates significant difference by t-test with significance level 5%.

| Data | MSE/test error ± std.error | | |
|---|---|---|---|
| | L2IWE | TarS | Unweighted |
| puma8nh | $\mathbf{2.81 \pm 0.03}$ | $2.98 \pm 0.04$ | $3.05 \pm 0.03$ |
| kin8nm | $\mathbf{0.98 \pm 0.01}$ | $1.19 \pm 0.02$ | $1.27 \pm 0.02$ |
| kin8nh | $\mathbf{1.06 \pm 0.01}$ | $1.28 \pm 0.03$ | $1.31 \pm 0.02$ |
| kin8fm | $\mathbf{0.07 \pm 0.00}$ | $0.07 \pm 0.00$ | $\mathbf{0.07 \pm 0.00}$ |
| kin8fh | $\mathbf{0.45 \pm 0.00}$ | $0.50 \pm 0.01$ | $0.50 \pm 0.00$ |
| CA Housing | $\mathbf{2.77 \pm 1.00}$ | $\mathbf{6.87 \pm 4.30}$ | $\mathbf{9.76 \pm 3.72}$ |
| elevators | $\mathbf{0.15 \pm 0.00}$ | $3.54 \pm 0.34$ | $0.28 \pm 0.01$ |
| delta ailerons | $\mathbf{0.08 \pm 0.00}$ | $0.14 \pm 0.01$ | $0.12 \pm 0.00$ |
| abalone | $\mathbf{0.37 \pm 0.00}$ | $0.54 \pm 0.02$ | $0.43 \pm 0.01$ |
| housing | $\mathbf{0.52 \pm 0.01}$ | $0.68 \pm 0.03$ | $1.21 \pm 0.02$ |

Overall, we conclude that the proposed L2IWE is a useful and promising method for target shift adaptation.

## 6. Conclusion

We considered the supervised learning problems under target shift (Zhang et al., 2013), where the target-marginal distributions $p(y)$ differ between training and testing phases, yet the target-conditional distribution $p(\boldsymbol{x}|y)$ remains the same. This dataset shift scenario is often encountered in practice; for instance, prior probability shift (Storkey, 2009),

anti-causal regression (Schölkopf et al., 2012), endogenous stratified sampling in econometrics (Manski and Lerman, 1977), or sample selection bias in econometrics and sociology (Heckman, 1979). Although various methods have been proposed for supervised learning under the categorical target shift (a.k.a. *class prior change*) (du Plessis and Sugiyama, 2014; Iyer et al., 2014; Zhang et al., 2013), few of them can be applied to the continuous target shift. Therefore, the main goal of this paper was to handle the continuous target shift in supervised learning. The key part of the proposed approach is a novel importance weight estimation procedure, called L2IWE. Utilizing labeled training data and unlabeled test data, L2IWE estimates the importance weight function by direct distribution matching under $L^2$-distance. Compared with the state-of-the-art method (Zhang et al., 2013), L2IWE is equipped with an automatic model selection procedure, and thus is practically more useful. The experiments showed that our proposed method achieves better performance for the target shift adaptation in conditional density estimation and regression tasks.

## Acknowledgments

The authors are grateful to reviewers for helpful comments. DTN was supported by the MEXT scholarship. MCdP was supported by the JST CREST program and MS was supported by KAKENHI 25700022.

## References

M. C. du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In *Proceedings of 29th International Conference on Machine Learning (ICML2012)*, pages 823–830, Edinburgh, Scotland, Jun. 26–Jul. 1 2012.

M. C. du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119, 2014.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2006.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, NY, USA, 2009.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

A. Iyer, S. Nath, and S. Sawaragi. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *Proceedings of 31st International Conference on Machine Learning (ICML)*, pages 530–538, Beijing, China, 2014.

T. Kanamori, T. Suzuki, and M. Sugiyama. Computational complexity of kernel-based density-ratio estimation: A condition number analysis. *Machine Learning*, 90(3):431–460, 2013.

C. Manski and S. Lerman. The estimation of choice probabilities from choice-based samples. *Econometrica*, 45:1977–1988, 1977.

J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence (eds.). *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, USA, 2009.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

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

L. Song, A. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo. Supervised feature selection via dependence estimation. In *Proceedings of 24th International Conference on Machine Learning*, pages 823–830, 2007.

A. Storkey. When training and test sets are different: characterizing learning transfer. In J. C. Quinonero, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, editors, *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, 2009.

M. Sugiyama and M. Kawanabe. *Machine Learning in Non-Stationary Environments*. MIT Press, Cambridge, MA, USA, 2012.

M. Sugiyama, M. Kraudelat, and K. R. Muller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

M. Sugiyama, I. Takeuchi, T. Kanamori, T. Suzuki, H. Hachiya, and D. Okanohara. Least-squares conditional density estimation. *IEICE Transactions on Information and Systems*, E93-D(3):583–594, 2010.

M. Sugiyama, S. Liu, M. C. du Plessis, M. Yamanaka, M. Yamada, T. Suzuki, and T. Kanamori. Direct divergence approximation between probability distributions and its applications in machine learning. *Journal of Computing Science and Engineering*, 7 (2):99–111, 2013a.

M. Sugiyama, T. Suzuki, T. Kanamori, M. C. du Plessis, S. Liu, and I. Takeuchi. Density-difference estimation. *Neural Computation*, 25(10):2734–2775, 2013b.

Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, 1998.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *Proceedings of 30th International Conference on Machine Learning (ICML)*, pages 819–827, Atlanta, Georgia, USA, 2013.