Double-Weighting for Covariate Shift Adaptation

José I. Segovia-Martín [1] **Santiago Mazuelas** [1 2] **Anqi Liu** [3]

Abstract

Supervised learning is often affected by a covariate shift in which the marginal distributions of instances (covariates x) of training and testing samples $\mathrm{p_{tr}}(x)$ and $\mathrm{p_{te}}(x)$ are different but the label conditionals coincide. Existing approaches address such covariate shift by either using the ratio $\mathrm{p_{te}}(x)/\mathrm{p_{tr}}(x)$ to weight training samples (reweighted methods) or using the ratio $\mathrm{p_{tr}}(x)/\mathrm{p_{te}}(x)$ to weight testing samples (robust methods). However, the performance of such approaches can be poor under support mismatch or when the above ratios take large values. We propose a minimax risk classification (MRC) approach for covariate shift adaptation that avoids such limitations by weighting both training and testing samples. In addition, we develop effective techniques that obtain both sets of weights and generalize the conventional kernel mean matching method. We provide novel generalization bounds for our method that show a significant increase in the effective sample size compared with reweighted methods. The proposed method also achieves enhanced classification performance in both synthetic and empirical experiments.

1. Introduction

Most supervised learning methods assume that training and testing samples are drawn i.i.d. from the same underlying distribution. However, practical scenarios are often affected by a covariate shift in which the marginal distributions of instances (covariates x) of training and testing samples $\mathrm{p_{tr}}(x)$ and $\mathrm{p_{te}}(x)$ are different (see e.g., (Sugiyama & Kawanabe, 2012; Quiñonero-Candela et al., 2008)), while the conditional label distribution stays the same. In such scenarios, conventional supervised classification methods, like empirical risk minimization, can perform poorly because the empirical risk is approximating the training expected risk, rather than the test expected risk.

Most of the existing methods for covariate shift adaptation are based on the reweighted approach (Sugiyama & Kawanabe, 2012; Quiñonero-Candela et al., 2008; Cortes et al., 2008; Huang et al., 2006). These methods weight loss functions at training using the ratio $\mathrm{p_{te}}(x)/\mathrm{p_{tr}}(x)$ so that training samples more likely in the test distribution are assigned higher weights (see Fig. 1), increasing their relevance at training. Such ratios can be estimated by using training and testing instances (Tsuboi et al., 2009; Yamada et al., 2011; Liu et al., 2013). Reweighted methods are designed for situations where the support of $\mathrm{p_{tr}}$ contains that of $\mathrm{p_{te}}$. However, even if such condition is satisfied, reweighted methods may achieve poor performances if the ratio $\mathrm{p_{te}}(x)/\mathrm{p_{tr}}(x)$ take large values at certain training samples, leading to inaccurate estimations of expected losses (see e.g., (Cortes & Mohri, 2014; Reddi et al., 2015)). Such problems can be alleviated by flattening the above ratio (Shimodaira, 2000; Yamada et al., 2011), by utilizing a regularization term based on the unweighted solution (Reddi et al., 2015), and by directly estimating weights for training samples through kernel mean matching (KMM) methods (Gretton et al., 2008; Huang et al., 2006).

Robust methods for covariate shift adaptation (Liu & Ziebart, 2014; 2017; Chen et al., 2016) are derived from a distributionally robust learning framework, where the feature expectation matching constraints are obtained from training samples but the adversarial risk is defined on the test distribution. Such methods weight feature functions at testing using the ratio $\mathrm{p_{tr}}(x)/\mathrm{p_{te}}(x)$ (see Fig. 1). The resulting parametric form produces less confident predictions when testing samples are less likely in the training distribution. Robust methods are designed for situations where the support of $\mathrm{p_{te}}$ contains that of $\mathrm{p_{tr}}$. However, even if such condition is satisfied, robust methods may achieve poor performances if the ratio $\mathrm{p_{tr}}(x)/\mathrm{p_{te}}(x)$ take large values at certain testing samples, leading to overconfident classification rules.

[1]Basque Center for Applied Mathematics (BCAM), Bilbao, Spain [2]IKERBASQUE-Basque Foundation for Science [3]CS department, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, USA. Correspondence to: José I. Segovia-Martín <jsegovia@bcamath.org>, Santiago Mazuelas <smazuelas@bcamath.org>, Anqi Liu <aliu@cs.jhu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

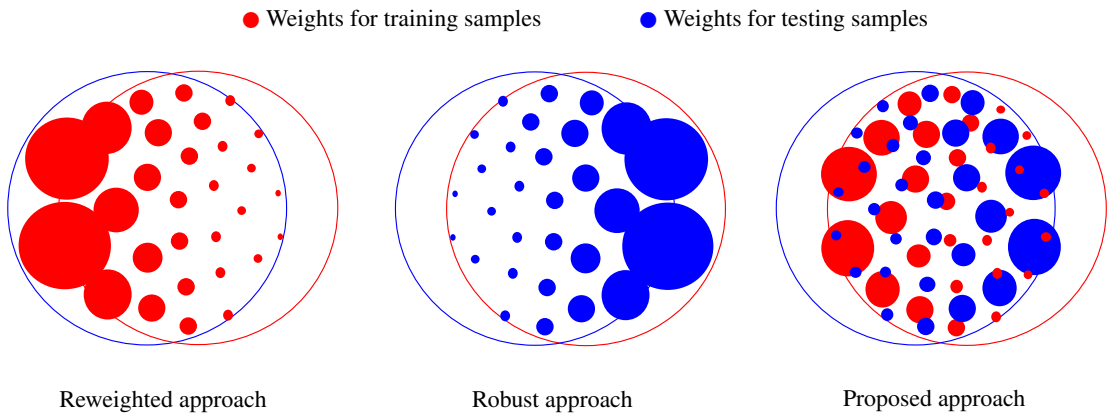

Figure 1. Different approaches for covariate shift adaptation (training and testing instances follow Gaussian distributions with probability mass concentrated in the red and blue circles, resp.). Reweighted methods weight training instances x using the ratio $\beta(x) = \mathrm{p_{te}}(x)/\mathrm{p_{tr}}(x)$ while robust methods weight testing instances x using the ratio $\alpha(x) = \mathrm{p_{tr}}(x)/\mathrm{p_{te}}(x)$. The proposed approach utilizes weights both for training and testing instances and can avoid large weights $\beta(x)$ by reducing the corresponding $\alpha(x)$ and avoid large weights $\alpha(x)$ by reducing the corresponding $\beta(x)$.

In practice, the distributions of training and testing instances can differ in an arbitrary manner (e.g., their supports may not be contained in each other). This paper proposes a learning methodology that can tackle such general covariate shift and addresses the limitations of existing approaches by weighting both training and testing samples (see Fig. 1). In particular, the methods proposed are based on minimax risk classifiers (MRCs) (Mazuelas et al., 2022; 2023) and utilize weighted averages of training samples to estimate expectations of weighted feature functions under the test distribution. Specifically, the main contributions in the paper are as follows.

- We present a learning framework for general covariate shift adaptation based on a double-weighting of both training and testing samples. Our framework encompasses existing approaches for specific choices of weights.
- We propose effective techniques that obtain weights for training and testing samples, and generalize the conventional KMM that only obtains weights for training samples.
- We develop generalization bounds for the proposed methods that show a significant increase in effective sample size compared with reweighted approaches.
- We experimentally assess the performance improvement obtained by the proposed techniques in multiple covariate shift scenarios.

Notations. Calligraphic upper case letters represent sets; bold lower and upper case letters represent vectors and matrices, respectively; for a vector $\mathbf{v}$, $v^{(i)}$ denotes its i-th component, $|\mathbf{v}|$ denotes its component-wise absolute value, and $(\mathbf{v})_+$ denotes its positive part; $\mathbf{1}$ denotes a vector with all components equal to 1; $||\cdot||_1$, $||\cdot||_\infty$, and $||\cdot||_{\mathcal{H}}$ denote the 1-norm, the infinity, and the Hilbert space norm of its argument, respectively; $\preceq$ and $\succeq$ represent vector component-wise inequalities; $N(\mathbf{m}, \mathbf{\Sigma})$ denotes the pdf of a Gaussian r.v. $\mathbf{x}$ with mean $\mathbf{m}$ and covariance matrix $\mathbf{\Sigma}$; and $\mathbb{E}_{\mathrm{p}}\{\cdot\}$ denotes the expectation of its argument w.r.t distribution p.

2. Preliminaries

This section describes the learning setup, the two main existing approaches for covariate shift adaptation, and the framework of MRCs.

Setup. Let $\mathcal{X}$ be the set of instances and $\mathcal{Y}$ the set of labels represented by the set $\{1, \ldots, |\mathcal{Y}|\}$. We denote by $\Delta(\mathcal{X} \times \mathcal{Y})$ the set of probability distributions over $\mathcal{X}$ and $\mathcal{Y}$, and by $\mathrm{T}(\mathcal{X}, \mathcal{Y})$ the set of classification rules. For $\mathrm{h} \in \mathrm{T}(\mathcal{X}, \mathcal{Y})$, we denote by $\mathrm{h}(y|x)$ the probability with which instance $x \in \mathcal{X}$ is classified by label $y \in \mathcal{Y}$. We use the notation $\mathrm{p_{te}}$ for the underlying distribution at test, and $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$ for the set of training samples. The ℓ-risk of a classification rule h is its expected classification loss with respect to the true underlying distribution at test $\mathrm{p_{te}}$, i.e., $R(\mathrm{h}) = \mathbb{E}_{\mathrm{p_{te}}}\{\ell(\mathrm{h}, (x, y))\}$.

The learning objective is to use the training samples to find a classification rule h that has small ℓ-risk $R(\mathrm{h})$. In this paper, we consider 0-1-loss and log-loss:

$$\ell_{01}(\mathrm{h}, (x, y)) = 1 - \mathrm{h}(y|x) \tag{1}$$

$$\ell_{\log}(\mathrm{h}, (x, y)) = -\log \mathrm{h}(y|x). \tag{2}$$

Covariate shift. Under covariate shift, the training samples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ follow a distribution $\mathrm{p_{tr}}(x, y)$ such that the marginal distributions of in-

stances differ, $\mathrm{p}_{\mathrm{te}}(x) \neq \mathrm{p}_{\mathrm{tr}}(x)$, but label conditionals coincide, $\mathrm{p}_{\mathrm{tr}}(y|x) = \mathrm{p}_{\mathrm{te}}(y|x)$. In addition, covariate shift methods assume that the ratio between $\mathrm{p}_{\mathrm{tr}}(x)$ and $\mathrm{p}_{\mathrm{te}}(x)$ is known or that t unsupervised samples $x_{n+1}, x_{n+2}, \ldots, x_{n+t}$ from $\mathrm{p}_{\mathrm{te}}(x)$ are known at training. In previous literature, it is also usually assumed that the training support contains that at testing or vice versa. In this paper, we consider general scenarios of covariate shift in which such supports are not required to contain each other.

2.1. Main existing approaches

Reweighted methods. Most of the techniques for covariate shift adaptation are based on the reweighted approach (Sugiyama & Kawanabe, 2012; Shimodaira, 2000; Zadrozny, 2004; Cortes et al., 2008; Dudík et al., 2005; Lin et al., 2002). These methods exploit that, for any function f, we have that

$$\mathbb{E}_{\mathrm{p}_{\mathrm{te}}} f(x,y) = \mathbb{E}_{\mathrm{p}_{\mathrm{tr}}} \beta(x) f(x,y), \text{ for } \beta(x) = \frac{\mathrm{p}_{\mathrm{te}}(x)}{\mathrm{p}_{\mathrm{tr}}(x)} \quad (3)$$

if $\mathrm{p}_{\mathrm{te}}(x) > 0 \Rightarrow \mathrm{p}_{\mathrm{tr}}(x) > 0$. Reweighted methods weight loss functions at training by means of the weight function $\beta(x)$ in (3), as detailed in Appendix A. Using these weights, such methods can account for the fact that some training instances are unlikely at testing, and assign low relevance to such instances at training (see Fig.1).

Reweighted methods assume the support of p_{tr} contains that of p_{te} (i.e., $\mathrm{p}_{\mathrm{te}}(x) > 0 \Rightarrow \mathrm{p}_{\mathrm{tr}}(x) > 0$) so that (3) is valid. Even if this condition is satisfied, such methods may achieve poor performances if the ratio $\beta(x)$ in (3) takes large values at certain training samples. In these cases, the learning process is dominated by few training samples (see e.g., (Cortes & Mohri, 2014; Gretton et al., 2008)). The flattening approach alleviates such problems using weights for training samples smoothed utilizing a hyperparameter γ as $(\mathrm{p}_{\mathrm{te}}(x)/\mathrm{p}_{\mathrm{tr}}(x))^{\gamma}$ in (Shimodaira, 2000) and as $\mathrm{p}_{\mathrm{te}}(x)/\left(\gamma \mathrm{p}_{\mathrm{te}}(x) + (1-\gamma)\mathrm{p}_{\mathrm{tr}}(x)\right)$ in (Yamada et al., 2011).

Robust methods. Robust methods under covariate shift (Liu & Ziebart, 2014; 2017) exploit that, for any f:

$$\mathbb{E}_{\mathrm{p}_{\mathrm{te}}} \alpha(x) f(x,y) = \mathbb{E}_{\mathrm{p}_{\mathrm{tr}}} f(x,y), \text{ for } \alpha(x) = \frac{\mathrm{p}_{\mathrm{tr}}(x)}{\mathrm{p}_{\mathrm{te}}(x)} \quad (4)$$

if $\mathrm{p}_{\mathrm{tr}}(x) > 0 \Rightarrow \mathrm{p}_{\mathrm{te}}(x) > 0$. Robust methods weight feature functions at testing[1] by means of the weight function $\alpha(x)$ in (4), as detailed in Appendix A. Using these weights, such methods can account for the fact that some testing instances are unlikely at training, and consider rules that assign low-confidence predictions to such instances (see Fig. 1).

Robust methods assume the support of p_{te} contains that of p_{tr} (i.e., $\mathrm{p}_{\mathrm{tr}}(x) > 0 \Rightarrow \mathrm{p}_{\mathrm{te}}(x) > 0$) so that (4) is valid. Even if this condition is satisfied, such methods may achieve poor performances if the ratio $\alpha(x)$ in (3) takes large values at certain testing samples. In these cases, the classification rule would only provide confident predictions at few testing samples.

The connection and symmetric relation between reweighted and robust methods are also shown in Theorem 3 in (Liu & Ziebart, 2014). They can both be regarded as special cases of the adversarial risk minimization framework (Fathony et al., 2016) when the feature expectation matching constraints are set to match different empirical estimates of the features. To enable covariate shift adaptation in general cases (e.g., training and testing supports not contained in each other), this paper proposes a learning framework that avoids the limitations of existing methods by utilizing a double-weighting approach.

2.2. Minimax Risk Classifiers

Similarly to other approaches based on robust risk minimization (RRM) (Farnia & Tse, 2016; Fathony et al., 2016), MRC methods (Mazuelas et al., 2022; 2023) do not require that the training samples follow the same distribution as the testing samples. MRCs minimize the worst-case expected loss with respect to distributions in uncertainty sets that can contain the true underlying distribution with high probability. The uncertainty sets are given by constraints on the expectation of a function $\Phi : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}^m$ referred to as feature mapping. Such a function can be defined using one-hot encodings of the elements of $\mathcal{Y}$ as $\Phi(x,y) = \boldsymbol{e}_y \otimes \mathbf{x}$, where $\boldsymbol{e}_y$ is the y-th element of the canonical basis of $\mathbb{R}^{|\mathcal{Y}|}$ and $\otimes$ denotes the Kronecker product.

Given the uncertainty set $\mathcal{U}$, we say that a classification rule $\mathrm{h}^{\mathcal{U}}$ is a ℓ-MRC for $\mathcal{U}$ if

$$\mathrm{h}^{\mathcal{U}} \in \arg \min_{\mathrm{h} \in \mathrm{T}(\mathcal{X},\mathcal{Y})} \max_{\mathrm{p} \in \mathcal{U}} \ell(\mathrm{h}, \mathrm{p}) \quad (5)$$

and, we denote by $R(\mathcal{U})$ the minimax risk against $\mathcal{U}$, i.e.,

$$R(\mathcal{U}) = \min_{\mathrm{h} \in \mathrm{T}(\mathcal{X},\mathcal{Y})} \max_{\mathrm{p} \in \mathcal{U}} \ell(\mathrm{h}, \mathrm{p}) \quad (6)$$

where $\ell(\mathrm{h}, \mathrm{p})$ denotes the expected loss of classification rule h w.r.t. distribution p.

3. Framework for Adaptation to General Covariate Shift

This section first describes the proposed double-weighting approach and the corresponding MRC learning methodo-

[1] The robust bias-aware prediction weight the feature functions in both training and testing as the weight appears in the predictive parametric form. Here we are emphasizing the weights at testing to show a symmetric view between these two methods.

logy. We then describe its relationship with existing techniques, present finite-sample generalization bounds for the proposed methods, and discuss the trade-off involved in the choice of weight functions.

3.1. Double-weighting

The proposed framework considers both training and testing weights, $\beta(x)$ and $\alpha(x)$ (see Fig.1). We exploit the fact that, for any function f, we have that

$$\mathbb{E}_{\mathrm{p_{te}}}\alpha(x)f(x,y) = \mathbb{E}_{\mathrm{p_{tr}}}\beta(x)f(x,y) \tag{7}$$

can be attained by multiple choices of weights $\alpha(x)$ and $\beta(x)$. For instance, it is satisfied taking

$$\alpha(x) = \min\left(C\frac{\mathrm{p_{tr}}(x)}{\mathrm{p_{te}}(x)}, 1\right), \beta(x) = \min\left(\frac{\mathrm{p_{te}}(x)}{\mathrm{p_{tr}}(x)}, C\right) \tag{8}$$

for any $C > 0$, since $\alpha(x)\mathrm{p_{te}}(x) = \beta(x)\mathrm{p_{tr}}(x)$, $\forall x \in \mathcal{X}$. Notice that the equality in (7) is satisfied taking weights as in (8) even if the supports of $\mathrm{p_{tr}}$ and $\mathrm{p_{te}}$ do not contain each other.

Such a double-weighting approach can avoid the limitations of reweighed and robust methods. For $x \in \mathcal{X}$ with large ratio $\mathrm{p_{te}}(x)/\mathrm{p_{tr}}(x)$, using a small $\alpha(x)$ can enable to have $\alpha(x)\mathrm{p_{te}}(x) = \beta(x)\mathrm{p_{tr}}(x)$ with moderate values of $\beta(x)$. Reciprocally, for $x \in \mathcal{X}$ with large ratio $\mathrm{p_{tr}}(x)/\mathrm{p_{te}}(x)$, using a small $\beta(x)$ can enable to have $\alpha(x)\mathrm{p_{te}}(x) = \mathrm{p_{tr}}(x)\beta(x)$ with moderate values of $\alpha(x)$. For instance, using weights as in (8) we have that $\beta(x) \leq C$ and $\alpha(x) \leq 1$ for any $x \in \mathcal{X}$. Considering both weights $\beta(x)$ and $\alpha(x)$, we can *both* assign low relevance to training instances that are unlikely at testing, *and also* assign low-confidence predictions to testing instances that are unlikely at training.

3.2. MRC learning framework using double-weighting

The proposed framework adapts to general covariate shift by constructing the uncertainty set $\mathcal{U}$ in (5) using both weights $\alpha(x)$ and $\beta(x)$. In particular, we use feature mappings weighted by $\alpha(x)$ as $\Phi_\alpha(x,y) = \alpha(x)\Phi(x,y)$ and constrain the difference between the expectation and empirical mean estimates of feature mappings as follows

$$\mathcal{U} = \{\mathrm{p} \in \Delta(\mathcal{X}\times\mathcal{Y}) : |\mathbb{E}_{\mathrm{p}}\Phi_\alpha(x,y) - \boldsymbol{\tau}| \preceq \boldsymbol{\lambda} \\ \text{and } \mathrm{p}(x) = \mathrm{p_{te}}(x), \forall x \in \mathcal{X}\} \tag{9}$$

where $\boldsymbol{\tau}$ denotes the mean vector of expectation estimates, and $\boldsymbol{\lambda}$ is a vector that determines the confidence with which $\mathrm{p_{te}}(x,y) \in \mathcal{U}$. The expectation of the feature mapping $\Phi_\alpha(x,y)$ is estimated using averages of training samples weighted by $\beta(x)$ as

$$\boldsymbol{\tau} = \frac{1}{n}\sum_{i=1}^{n}\Phi_\beta(x_i,y_i), \text{ for } \Phi_\beta(x,y) = \beta(x)\Phi(x,y). \tag{10}$$

Notice that the mean vector $\boldsymbol{\tau}$ is an unbiased estimator of $\mathbb{E}_{\mathrm{p_{te}}}\Phi_\alpha(x,y)$ for any choice of weights satisfying $\alpha(x)\mathrm{p_{te}}(x) = \beta(x)\mathrm{p_{tr}}(x)$, in particular for those given by (8). In addition, the accuracy of $\boldsymbol{\tau}$ can be improved using weights $\alpha(x)$ that avoid large weights $\beta(x)$, as discussed in Section 3.4 below.

Convex optimization. We next show how MRCs corresponding with uncertainty sets (9) can be learned by solving the convex optimization problem

$$\min_{\boldsymbol{\mu}} -\boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{\mu} + \mathbb{E}_{\mathrm{p_{te}}(x)}\varphi_\ell(\boldsymbol{\mu},x,\alpha(x)) + \boldsymbol{\lambda}^{\mathrm{T}}|\boldsymbol{\mu}| \tag{11}$$

where φ_ℓ is a function defined under different loss functions. For 0-1-loss, we have

$$\varphi_{01}(\boldsymbol{\mu},x,\alpha(x)) = 1 + \max_{\mathcal{C}\subseteq\mathcal{Y}}\frac{\sum_{y\in\mathcal{C}}\Phi_\alpha(x,y)^{\mathrm{T}}\boldsymbol{\mu} - 1}{|\mathcal{C}|} \tag{12}$$

and, for log-loss, we have

$$\varphi_{\log}(\boldsymbol{\mu},x,\alpha(x)) = \log\sum_{y\in\mathcal{Y}}\exp\left\{\Phi_\alpha(x,y)^{\mathrm{T}}\boldsymbol{\mu}\right\}. \tag{13}$$

Theorem 3.1. *Let $\boldsymbol{\tau},\boldsymbol{\lambda} \in \mathbb{R}^m$ be such that the uncertainty set $\mathcal{U}$ in (9) is not the empty set. If $\boldsymbol{\mu}^*$ is a solution of (11) for 0-1-loss, the classification rule*

$$\mathrm{h}^{\mathcal{U}}(y|x) = \left(\alpha(x)\Phi(x,y)^T\boldsymbol{\mu}^* - \varphi_{01}(\boldsymbol{\mu}^*,x,\alpha(x)) + 1\right)_+ \tag{14}$$

is a 0-1-MRC for $\mathcal{U}$. If $\boldsymbol{\mu}^*$ is a solution of (11) for log-loss, the classification rule

$$\mathrm{h}^{\mathcal{U}}(y|x) = \exp\left\{\alpha(x)\Phi(x,y)^T\boldsymbol{\mu}^* - \varphi_{\log}(\boldsymbol{\mu}^*,x,\alpha(x))\right\} \tag{15}$$

is a log-MRC for $\mathcal{U}$. In addition, the minimax risk $R(\mathcal{U})$ is given by

$$R(\mathcal{U}) = -\boldsymbol{\tau}^T\boldsymbol{\mu}^* + \mathbb{E}_{\mathrm{p_{te}}(x)}\varphi_\ell(\boldsymbol{\mu}^*,x,\alpha(x)) + \boldsymbol{\lambda}^T|\boldsymbol{\mu}^*|. \tag{16}$$

Proof. See Appendix B. □

Remarks. The optimization in (11) can be addressed in practice using conventional optimization methods such as stochastic gradient descent. If unlabeled instances from the test distribution are available at training, they can directly be used to obtain samples corresponding to the (sub)gradient of $\mathbb{E}_{\mathrm{p_{te}}(x)}\varphi_\ell(\boldsymbol{\mu},x,\alpha(x))$ since the function φ_ℓ does not depend on labels. If the marginals $\mathrm{p_{tr}}(x)$, $\mathrm{p_{te}}(x)$ are known, training samples can be used to obtain samples of the above gradient using (3). This theorem is novel as we apply weights α and β for covariate shift adaptation, even though the general form is analogous to the results in (Mazuelas et al., 2022; 2023), which studies MRC with train and test data sampled i.i.d. from the same distribution.

Regularization. The convex optimization problem (11) carries out an L1-type regularization, where the regularization parameter is given by vector $\boldsymbol{\lambda}$. The regularization term in (11) allows to penalize each component of parameter $\boldsymbol{\mu}$ differently, such that feature components with poorly estimated expectations (i.e., components i with large $\lambda^{(i)}$) are strongly penalized.

Classification rule. The deterministic classifier associated with $\mathrm{h}^{\mathcal{U}}$ classifies each instance with the label maximizing $\mathrm{h}^{\mathcal{U}}(y|x)$. For both losses, this deterministic classifier is given by

$$\begin{aligned}\arg\max_{y\in\mathcal{Y}} \mathrm{h}^{\mathcal{U}}(y|x) &= \arg\max_{y\in\mathcal{Y}} \alpha(x)\Phi(x,y)^{\mathrm{T}}\boldsymbol{\mu}^* \\ &= \arg\max_{y\in\mathcal{Y}} \Phi(x,y)^{\mathrm{T}}\boldsymbol{\mu}^*.\end{aligned} \tag{17}$$

Such deterministic classifiers, denoted by $\mathrm{h}^{\mathcal{U}}_{\mathrm{d}}$, allow us to classify testing samples even if we do not know the weights $\alpha(x)$ associated with them.

Predictive confidence. The values of $\alpha(x)$ adjust the confidence with which each sample is classified. For instance, for very small values of $\alpha(x)$, the classifier $\mathrm{h}^{\mathcal{U}}$ uniformly assigns labels in the set $\mathcal{Y}$ for both losses, i.e., $\mathrm{h}^{\mathcal{U}}(y|x) = 1/|\mathcal{Y}|$ for all $y \in \mathcal{Y}$.

Relation with existing approaches. The general framework proposed above encompasses existing approaches, as detailed in Appendix A for binary classification with log-loss. The usage of weights $\alpha(x) = 1, \beta(x) = \mathrm{p}_{\mathrm{te}}(x)/\mathrm{p}_{\mathrm{tr}}(x)$ leads to reweighted methods (Sugiyama & Kawanabe, 2012), approximating $\mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)}\varphi_\ell(\boldsymbol{\mu}, x, \alpha(x))$ in (11) using training instances. The usage of weights $\alpha(x) = \mathrm{p}_{\mathrm{tr}}(x)/\mathrm{p}_{\mathrm{te}}(x), \beta(x) = 1$ leads to robust methods (Liu & Ziebart, 2014), approximating the gradient of $\mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)}\varphi_\ell(\boldsymbol{\mu}, x, \alpha(x))$ in (11) using training instances.

3.3. Generalization bounds

The following shows the generalization bounds of the proposed methods in Section 3.2. Such bounds are given in terms of smallest minimax risk, R^∞, that corresponds with the uncertainty set given by the exact expectations, and is defined by

$$R^\infty = \min_{\boldsymbol{\mu}} -\mathbb{E}_{\mathrm{p}_{\mathrm{te}}}\Phi_\alpha(x,y)^{\mathrm{T}}\boldsymbol{\mu} + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)}\varphi_\ell(\boldsymbol{\mu}, x, \alpha(x)). \tag{18}$$

The MRC corresponding to that smallest minimax risk R^∞ could only be obtained by an exact estimation of the expectation of the feature mapping Φ_α that in turn would require an infinite amount of training samples. The theorem below shows risk bounds for the proposed MRCs in terms of minimax risks $R(\mathcal{U})$ and smallest minimax risks R^∞.

Theorem 3.2. *Let $\mathcal{U}$ be a non-empty uncertainty set given by (9) and $\mathrm{h}^{\mathcal{U}}$ be an ℓ-MRC for $\mathcal{U}$. If $\boldsymbol{\mu}^*$ and $\boldsymbol{\mu}_\infty$ are solutions to (11) and (18), respectively, then, we have that*

$$R\left(\mathrm{h}^{\mathcal{U}}\right) \leq R(\mathcal{U}) + \left(|\boldsymbol{\tau} - \mathbb{E}_{\mathrm{p}_{te}}\Phi_\alpha(x,y)| - \boldsymbol{\lambda}\right)^T |\boldsymbol{\mu}^*| \tag{19}$$

$$\begin{aligned}R\left(\mathrm{h}^{\mathcal{U}}\right) \leq & R^\infty + \boldsymbol{\lambda}^T(|\boldsymbol{\mu}_\infty| - |\boldsymbol{\mu}^*|) \\ & + |\boldsymbol{\tau} - \mathbb{E}_{\mathrm{p}_{te}}\Phi_\alpha(x,y)|^T |\boldsymbol{\mu}_\infty - \boldsymbol{\mu}^*|.\end{aligned} \tag{20}$$

Proof. See Appendix B. □

Note that the minimax risk $R(\mathcal{U})$ obtained at learning offers an upper bound for the ℓ-risk if $\boldsymbol{\lambda} \succeq |\boldsymbol{\tau} - \mathbb{E}_{\mathrm{p}_{\mathrm{te}}}\Phi_\alpha(x,y)|$ and an approximate upper bound for general $\boldsymbol{\lambda}$. In addition, the difference between the risk $R\left(\mathrm{h}^{\mathcal{U}}\right)$ and the smallest minimax risk R^∞ decreases with the estimation error $|\boldsymbol{\tau} - \mathbb{E}_{\mathrm{p}_{\mathrm{te}}}\Phi_\alpha(x,y)|$.

We next show how the proposed methods can lead to a significant increase in effective size compared with reweighted methods.

Corollary 3.3. *Let $\mathcal{U}$ be a non-empty uncertainty set given by (9) with $\boldsymbol{\lambda} = 0$, and $\mathrm{h}^{\mathcal{U}}$ be an ℓ-MRC for $\mathcal{U}$. If weights $\alpha(x)$ and $\beta(x)$ are given by (8) with $C = B/\sqrt{D}$ for $D \geq 1$ and*

$$B = \sup_{x\in\mathcal{X}} \mathrm{p}_{te}(x)/\mathrm{p}_{tr}(x). \tag{21}$$

Then, with probability at least $1-\delta$ we have that

$$R(\mathrm{h}^{\mathcal{U}}) \leq R^\infty + M\|\boldsymbol{\mu}_\infty - \boldsymbol{\mu}^*\|_\infty \sqrt{2\frac{B^2}{Dn}\log\frac{2}{\delta}} \tag{22}$$

where M is a constant satisfying $\|\Phi(x,y)\|_\infty \leq M$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$.

Proof. A direct consequence of Theorem 3.2 and Hoeffding's inequality. □

As described in (Cortes et al., 2010; Yu & Szepesvári, 2012), reweighted methods have an estimation error of the order $\sqrt{2\frac{B^2}{n}\log\frac{2}{\delta}}$ so that the methods proposed can achieve an effective sample size increased by a factor of D using the double-weighting given by (8) with $C = B/\sqrt{D}$. The next section more broadly discusses such an increase in effective sample size and the corresponding trade-off for predictions' confidence.

3.4. Choice of weight functions

Existing reweighted and robust methods, as well as the proposed general framework, utilize weights $\alpha(x)$ and $\beta(x)$ in the estimation of expectations:

$$\frac{1}{n}\sum_{i=1}^{n}\beta(x_i)f(x_i,y_i) \approx \mathbb{E}_{\mathrm{p}_{\mathrm{te}}}\alpha(x)f(x,y). \tag{23}$$

The error of such estimates is determined by the weights $\beta(x)$. If $\alpha(x)$ and $\beta(x)$ satisfy $\alpha(x)\mathrm{p}_{\mathrm{te}}(x) = \beta(x)\mathrm{p}_{\mathrm{tr}}(x)$, using Hoeffding's inequality we have that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\beta(x_i)f(x_i,y_i) - \mathbb{E}_{\mathrm{p_{te}}}\alpha(x)f(x,y)\right| \leq ||f||_\infty \sqrt{2\frac{||\beta||_\infty^2}{n}\log\frac{2}{\delta}} \quad (24)$$

with probability at least $1-\delta$.

In particular, for reweighted methods the bound (24) becomes $||f||_\infty\sqrt{2\frac{B^2}{n}\log\frac{2}{\delta}}$ with B given by (21) as shown in (Cortes et al., 2010; Yu & Szepesvári, 2012).

The error in the expectations estimates in (23) decreases when we choose the weights $\alpha(x)$ adequately. In particular, using small values of $\alpha(x)$ we can achieve $\alpha(x)\mathrm{p_{te}}(x) = \beta(x)\mathrm{p_{tr}}(x)$ with moderate values of $\beta(x)$. Such improvement comes at the expense of using classification rules with significant confidence only in the subregion of $\mathcal{X}$ in which $\alpha(x)$ is significantly larger than 0.

The above trade-off between error in expectations estimates and confidence of classification rules can be addressed using pairs of weights of the form (8) and varying the value of C. For values $C \geq B$, $\alpha(x) = 1$ and $\beta(x) = \mathrm{p_{te}}(x)/\mathrm{p_{tr}}(x)$ that corresponds to the reweighted approach. For values $C < B$, the expectations' estimates improve as we decrease C since $\|\beta\|_\infty = C$. However, the corresponding classification rules would only predict with significant confidence in the subregion of $\mathcal{X}$ where $\alpha(x)$ is significantly larger than 0. Such subregion shrinks when C decreases because it is composed by the $x \in \mathcal{X}$ where $\mathrm{p_{te}}(x)$ is not significantly larger than $C\mathrm{p_{tr}}(x)$. In the following, we present methods that obtain weights α and β addressing the above trade-off, and generalize conventional KMM methods.

4. Double-weighting Kernel Mean Matching

The KMM method obtains weights $\boldsymbol{\beta} \in \mathbb{R}^n$ for n training instances $x_1, x_2, \ldots, x_n$ using t testing instances $x_{n+1}, x_{n+2}, \ldots, x_{n+t}$ (Huang et al., 2006; Gretton et al., 2008). We propose the double-weighting KMM (DW-KMM) method that obtains weights $\boldsymbol{\beta} \in \mathbb{R}^n$ for the n training instances together with weights $\boldsymbol{\alpha} \in \mathbb{R}^t$ for the t testing instances by solving the optimization problem

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \left\|\frac{1}{t}\sum_{i=1}^{t}\alpha^{(i)}K(x_{n+i}) - \frac{1}{n}\sum_{i=1}^{n}\beta^{(i)}K(x_i)\right\|_{\mathcal{H}}^2 \\ \text{s.t.} \quad & 0 \leq \beta^{(i)} \leq B/\sqrt{D}, \text{ for } i = 1,\ldots,n \\ & 0 \leq \alpha^{(i)} \leq 1, \text{ for } i = 1,\ldots,t \\ & \left|\frac{1}{n}\sum_{i=1}^{n}\beta^{(i)} - \frac{1}{t}\sum_{i=1}^{t}\alpha^{(i)}\right| \leq \epsilon \\ & ||\boldsymbol{\alpha} - \mathbf{1}|| \leq \left(1 - \frac{1}{\sqrt{D}}\right)\sqrt{t} \end{aligned} \quad (25)$$

where $K : \mathcal{X} \longrightarrow \mathcal{H}$ is a feature map corresponding with a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with kernel $k(x,\bar{x}) = \langle K(x), K(\bar{x})\rangle_{\mathcal{H}}$.

As described above, the hyperparameter $D \geq 1$ in (25) balances the trade-off between error in expectation estimates and confidence of the classification. For $D = 1$, the optimization problem becomes that of KMM for reweighted methods (Huang et al., 2006; Gretton et al., 2008).

Performance guarantees. The proposed approach is an empirical version of the following (population) problem given by exact expectations

$$\begin{aligned}\min_{\alpha(x),\beta(x)} \quad & \left\|\mathbb{E}_{\mathrm{p_{te}}(x)}\alpha(x)K(x) - \mathbb{E}_{\mathrm{p_{tr}}(x)}\beta(x)K(x)\right\|_{\mathcal{H}}^2 \\ \text{s.t} \quad & 0 \leq \beta(x) \leq B/\sqrt{D},\ 0 \leq \alpha(x) \leq 1,\ \forall x \in \mathcal{X} \\ & \mathbb{E}_{\mathrm{p_{te}}(x)}\alpha(x) = \mathbb{E}_{\mathrm{p_{tr}}(x)}\beta(x) \\ & \mathbb{E}_{\mathrm{p_{te}}(x)}\left\{(\alpha(x)-1)^2\right\} \leq \left(1 - 1/\sqrt{D}\right)^2. \end{aligned} \quad (26)$$

The minimum value of (26) is zero since (8) with $C = B/\sqrt{D}$ is a feasible solution. Then, solutions of (26), $\hat{\beta}(x)$, $\hat{\alpha}(x)$, provide consistent estimators of expectations because

$$\mathbb{E}_{\mathrm{p_{te}}(x,y)}\hat{\alpha}(x)\Phi(x,y) = \mathbb{E}_{\mathrm{p_{tr}}(x,y)}\hat{\beta}(x)\Phi(x,y) \quad (27)$$

is satisfied if the kernel k is characteristic or if $\mathbb{E}_{\mathrm{p_{te}}(y|x)}\Phi(x,y)$ belongs to $\mathcal{H}$, analogously as shown in (Yu & Szepesvári, 2012).

With finite samples, the following theorem shows bounds for the difference between the empirical means in feature space for solutions of (26).

Theorem 4.1. *If $\hat{\beta}(x)$ and $\hat{\alpha}(x)$ are solutions of (26), with probability at least $1-\delta$ we have that*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}(x_i)K(x_i) - \frac{1}{t}\sum_{i=1}^{t}\hat{\alpha}(x_{n+i})K(x_{n+i})\right\|_{\mathcal{H}} \leq \left(1 + \sqrt{2\log\frac{2}{\delta}}\right)\kappa\sqrt{\left(\frac{B^2}{Dn} + \frac{1}{t}\right)} \quad (28)$$

where the constant κ satisfies $|k(x,x)| \leq \kappa^2$ for all $x \in \mathcal{X}$.

Proof. See Appendix C. □

Relation with conventional KMM. The solutions $\hat{\beta}(x)$ for conventional KMM in reweighted methods satisfy with probability at least $1-\delta$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}(x_i)K(x_i) - \frac{1}{t}\sum_{i=1}^{t}K(x_{n+i})\right\|_{\mathcal{H}} \leq \left(1 + \sqrt{2\log\frac{2}{\delta}}\right)\kappa\sqrt{\left(\frac{B^2}{n} + \frac{1}{t}\right)} \quad (29)$$

as shown in Lemma 4 of (Huang et al., 2006) and equation (10) in (Yu & Szepesvári, 2012). Therefore, the proposed DW-KMM allows to significantly improve the effective sample by exploiting the usage of weights α. Analogously to the results shown in Section 3.4, the effective sample size of the methods proposed is D times larger than that of existing KMM for reweighted methods.

5. Practical Algorithm

In this section, we present a practical algorithm for the proposed Double-Weighting for General Covariate Shift (DW-GCS), detailed in Algorithm 1. We first compute weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by solving (25), then, we learn the classifier's parameters by solving (11) using mean vector $\boldsymbol{\tau}$ defined in (10) and confidence vector $\boldsymbol{\lambda}$.

Algorithm 1 The proposed algorithm: DW-GCS

Input: Training samples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
Testing instances $x_{n+1}, x_{n+2}, \ldots, x_{n+t}$, D
Output: Weights $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$
Classifier parameters $\boldsymbol{\mu}^*$, Minimax risk $R(\mathcal{U})$

1: $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}} \leftarrow$ solution of (25)
2: $\boldsymbol{\tau} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}^{(i)} \Phi(x_i, y_i)$
3: $\boldsymbol{\lambda} \leftarrow$ solution of (31)
4: $\boldsymbol{\mu}^* \leftarrow$ solution of (30) using (12) for 0-1-loss, and (13) for log-loss
5: $R(\mathcal{U}) \leftarrow -\boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{\mu}^* + \frac{1}{t}\sum_{i=1}^{t} \varphi_\ell(\boldsymbol{\mu}^*, x_{n+i}, \hat{\alpha}^{(i)}) + \boldsymbol{\lambda}^{\mathrm{T}}|\boldsymbol{\mu}^*|$

Computing weights and learning MRCs. Weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are computed solving the convex optimization (25), which is a quadratic problem as detailed in Appendix D.

The optimization in (11) can be addressed by approximating the expectation by means of the t instances in testing $x_{n+1}, x_{n+2}, \ldots, x_{n+t}$ as

$$\min_{\boldsymbol{\mu}} -\boldsymbol{\tau}^T\boldsymbol{\mu} + \frac{1}{t}\sum_{i=1}^{t} \varphi_\ell(\boldsymbol{\mu}, x_{n+i}, \alpha^{(i)}) + \boldsymbol{\lambda}^T|\boldsymbol{\mu}| \qquad (30)$$

that is an unconstrained convex optimization problem and can be efficiently solved by conventional methods.

Hyperparameters. In principle, both hyperparameters $\boldsymbol{\lambda}$ and D can be obtained by cross-validation. However, standard cross-validation is not valid under covariate shift (Sugiyama et al., 2007). We hence avoid cross-validation and determine both parameters as follows.

As detailed in Section 3.4, the hyperparameter D serves to address the trade-off between error in expectation estimates and confidence of classification rules. For instance, values of D close to 1 can be effective in situations with a large number of samples while higher values of D can be effective with a reduced number of samples. This is shown by the theoretical results in the paper, since the estimation error in the proposed methods is of the order $\mathcal{O}(1/\sqrt{Dn})$, as described by the performance bounds in Corollary 3.3 and Theorem 4.1.

In practice, we propose to select the hyperparameter D taking advantage of the minimax risk provided at the learning stage by the methods presented. Specifically, we select the value of D to achieve the lowest minimax risk over a certain range $D \geq 1$. Note that, as described in Theorem 3.2, the minimax risk $R(\mathcal{U})$ obtained at learning offers an upper bound for the risk if $\boldsymbol{\lambda} \succeq |\boldsymbol{\tau} - \mathbb{E}_{\mathrm{p_{te}}}\Phi_\alpha(x, y)|$, and an approximate upper bound for general $\boldsymbol{\lambda}$. Therefore, the proposed selection method in the paper uses the value of D that results in the lowest upper bound over a range of values for D. Appendix E further illustrates the adequacy of such approach in practice.

The second hyperparameter $\boldsymbol{\lambda}$ is determined solving

$$\begin{aligned} \min_{\mathrm{p},\boldsymbol{\lambda}} \ & \mathbf{1}^{\mathrm{T}}\boldsymbol{\lambda} \\ \text{s.t.} \ & \boldsymbol{\tau} - \boldsymbol{\lambda} \preceq \sum_{i=1}^{t}\sum_{y\in\mathcal{Y}} \mathrm{p}(y|x_{n+i})\Phi_\alpha(x_{n+i}, y) \preceq \boldsymbol{\tau} + \boldsymbol{\lambda} \\ & \boldsymbol{\lambda}, \mathbf{p} \succeq \mathbf{0} \\ & \sum_{y\in\mathcal{Y}} \mathrm{p}(y|x_{n+i}) = 1/t \text{ for } i = 1, \ldots, t \end{aligned} \qquad (31)$$

that ensures the uncertainty set used is non-empty.

Complexity and implementation without testing instances. The computational complexity of the methods proposed is similar to existing methods for covariate shift adaptation. Specifically, the step for DW-KMM that obtains weights has a similar complexity as that for conventional KMM. The main difference is that (25) has t additional variables and $t + 1$ additional constraints corresponding to the weights $\boldsymbol{\alpha}$. The step that obtains the classifier parameters solving convex optimization problem (11) has the same complexity as that for conventional methods. Finally, the step that determines hyperparameters not only avoids the usage of cross-validation but can also reduce complexity. In particular, cross-validation with P partitions would require solving (11) P times for each candidate value for hyperparameters, while the methods proposed only require solving (31) and (11) once, for each candidate value of D.

Algorithm 1 details the implementation of DW-GCS in cases where testing instances are available at training. The methods proposed can be implemented with small modifications in cases where only training instances are available and the marginals (or their ratios) are known. In these cases, weights $\alpha(x)$ and $\beta(x)$ can be determined using (8) with $C = B/\sqrt{D}$ instead of solving (25), and optimization (11) can be addressed using the training instances instead of testing instances making use of equality (3).

6. Experiments

This section shows experimental results for the proposed approach in comparison with existing methods on synthetic and real datasets. Reweighted and robust approaches are implemented as in (Sugiyama et al., 2007; Liu & Ziebart, 2014) and described in Appendix A, the flattening method is implemented as in (Shimodaira, 2000), the RuLSIF is implemented as in (Yamada et al., 2011), the KMM method is implemented as in (Huang et al., 2006), and the methods proposed are implemented as described in Alg. 1. The source code for the methods presented is publicly available in the library MRCpy (Bondugula et al., 2023) and the experimental setup in https://github.com/MachineLearningBCAM/MRCs-for-Covariate-Shift-Adaptation.

For existing methods, the regularization parameter has been fine-tuned as shown in Appendix E. For the proposed methods, hyperparameters are obtained as described in Section 5. The results in this section are complemented by those in Appendix E that provide further implementation details and experimental results. In particular, the appendix shows that selecting the hyperparameter D with lowest minimax risk results in performances near those obtained with the best value for D by grid search.

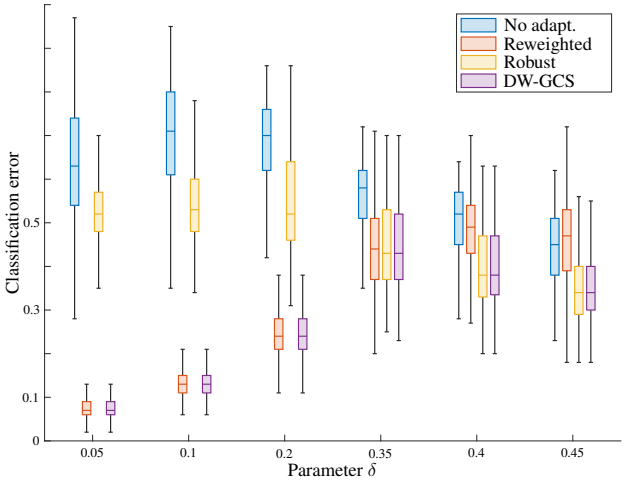

Figure 2. Classification error for different types of covariate shift. In the case $\delta = 0.05$, the training support contains that at testing, while in the case $\delta = 0.45$ we have the opposite.

Experiments with synthetic data. In the first set of results we show how the proposed approach can achieve covariate shift adaptation in situations where existing methods are challenged. For such results, the training and testing samples are drawn from distributions

$$\begin{aligned} \mathrm{p}_{\mathrm{tr}}(x) &= (0.5-\delta)N(\mathbf{m}_1, \mathbf{\Sigma}_1) + (0.5+\delta)N(\mathbf{m}_2, \mathbf{\Sigma}_2) \\ \mathrm{p}_{\mathrm{te}}(x) &= (1-\delta)N(\mathbf{m}_1, \mathbf{\Sigma}_1) + \delta N(\mathbf{m}_2, \mathbf{\Sigma}_2) \end{aligned} \tag{32}$$

with $\mathbf{m}_1 = [-3/2, 0]^{\mathrm{T}}$, $\mathbf{m}_2 = [3/2, 0]^{\mathrm{T}}$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = (1/4)\mathbf{I}$, and labels are $y = 1$ if $x^{(1)}x^{(2)} \geq 0$ and $y = 2$ otherwise. We use values $\delta \in \{0.05, 0.1, 0.2, 0.35, 0.4, 0.45\}$ to simulate different relations between the marginals of training and testing instances. We utilize the non-linear feature mapping given by instances components and their products and implement existing and proposed methods using the exact marginals. In addition, for each type of covariate shift (value of δ) we carry out 1,000 random repetitions with 100 training and testing samples.

Results. Figure 2 shows box-plots corresponding to the classification error of exiting and proposed approaches in comparison to that obtained without covariate shift adaptation ($\alpha(x) = \beta(x) = 1$). The results in the figure show how reweighted (resp. robust) methods obtain poor performances in situations where the support of training (resp. testing) instances does not contain that of testing (resp. training) instances. On the other hand, the methods proposed can leverage the presented double-weighting approach and adapt to more general covariate shifts.

Experiments with real datasets. In the second set of results, we assess the performance of the proposed methods in comparison with existing techniques using real datasets. In particular, reweighted and robust methods are implemented with marginal distributions estimated using log-linear models as shown in (Bickel et al., 2007; 2009).

We generate covariate shift in the datasets following (Huang et al., 2006) and (Gretton et al., 2008). In particular, we select training and testing samples with different probabilities based on the medians of the first 3 features, and based on the median of the first principal component of features. In (Huang et al., 2006) and (Gretton et al., 2008), covariate shift is generated with a biased sampling for testing instances that are drawn with probability δ_{te} if the first principal component or feature is larger than a certain value. In those works, the training samples are uniformly sampled, so that the generated covariate shifts correspond to situations where the support of training samples contains that of testing samples. In the numerical results of the table below, we generate more general covariate shifts by using a biased sampling both for training and testing instances (using probabilities $\delta_{\mathrm{tr}} = 0.7$ and $\delta_{\mathrm{te}} = 0.3$). These covariate shifts correspond to situations where the support of training and testing samples have certain overlap but they do not need to be contained in each other. Additionally, we include experimental results using the "News20groups" dataset that is intrinsically affected by a covariate shift since the training and testing partitions correspond to different times (Zhang et al., 2013). We consider the same 5 binary problems used in (Zhang et al., 2013), utilize the 1,000 features with highest Pearson's correlation, and randomly sample 1,000 training and testing samples in each repetition.

Results. Table 1 shows the averaged classification error

Table 1. Classification errors in 21 scenarios show that the proposed methods can more adequately adapt to general covariate shift. Values in bold show best classification error in each scenario.

Datasets	Reweighted	Flattening	RuLSIF	Robust	KMM	DW-GCS 0-1	DW-GCS log
Blood							
Feature 1	$.55 \pm .08$	$.48 \pm .11$	$\mathbf{.29 \pm .04}$	$.34 \pm .06$	$.32 \pm .03$	$.30 \pm .03$	$.31 \pm .03$
Feature 2	$.39 \pm .03$	$.38 \pm .03$	$.39 \pm .03$	$.40 \pm .03$	$.39 \pm .04$	$\mathbf{.38 \pm .05}$	$\mathbf{.38 \pm .05}$
Feature 3	$.43 \pm .05$	$.41 \pm .05$	$.36 \pm .04$	$.39 \pm .04$	$.36 \pm .04$	$\mathbf{.34 \pm .03}$	$.35 \pm .03$
PCA	$.48 \pm .05$	$.48 \pm .05$	$.29 \pm .05$	$.44 \pm .05$	$.30 \pm .05$	$\mathbf{.28 \pm .04}$	$\mathbf{.28 \pm .04}$
BreastCancer							
Feature 1	$.05 \pm .02$	$.05 \pm .03$	$.05 \pm .02$	$.06 \pm .03$	$.06 \pm .02$	$\mathbf{.04 \pm .02}$	$\mathbf{.04 \pm .02}$
Feature 2	$.06 \pm .02$	$.05 \pm .02$	$.06 \pm .03$	$.07 \pm .03$	$.06 \pm .03$	$\mathbf{.04 \pm .02}$	$\mathbf{.04 \pm .02}$
Feature 3	$.05 \pm .02$	$.05 \pm .02$	$.05 \pm .02$	$.06 \pm .03$	$.05 \pm .02$	$\mathbf{.04 \pm .02}$	$\mathbf{.04 \pm .02}$
PCA	$.03 \pm .01$	$.03 \pm .01$	$.03 \pm .01$	$.03 \pm .01$	$.03 \pm .01$	$\mathbf{.02 \pm .01}$	$\mathbf{.02 \pm .01}$
Haberman							
Feature 1	$.48 \pm .07$	$.47 \pm .08$	$.31 \pm .06$	$.41 \pm .09$	$.34 \pm .10$	$\mathbf{.28 \pm .07}$	$.29 \pm .06$
Feature 2	$.46 \pm .08$	$.44 \pm .08$	$.31 \pm .06$	$.39 \pm .08$	$.36 \pm .10$	$\mathbf{.29 \pm .08}$	$.30 \pm .07$
Feature 3	$\mathbf{.33 \pm .05}$	$\mathbf{.33 \pm .05}$	$\mathbf{.33 \pm .05}$	$.36 \pm .06$	$.42 \pm .08$	$.35 \pm .07$	$.36 \pm .06$
PCA	$.43 \pm .12$	$.42 \pm .12$	$\mathbf{.29 \pm .05}$	$.42 \pm .11$	$.35 \pm .08$	$.30 \pm .08$	$.31 \pm .07$
Ringnorm							
Feature 1	$.27 \pm .02$	$.26 \pm .02$	$\mathbf{.25 \pm .02}$	$.26 \pm .02$	$\mathbf{.25 \pm .02}$	$\mathbf{.25 \pm .02}$	$\mathbf{.25 \pm .02}$
Feature 2	$.28 \pm .02$	$.27 \pm .02$	$\mathbf{.25 \pm .02}$	$.27 \pm .02$	$.26 \pm .02$	$\mathbf{.25 \pm .02}$	$\mathbf{.25 \pm .02}$
Feature 3	$.28 \pm .02$	$.27 \pm .02$	$\mathbf{.25 \pm .02}$	$.27 \pm .02$	$.26 \pm .03$	$\mathbf{.25 \pm .02}$	$\mathbf{.25 \pm .02}$
PCA	$.32 \pm .03$	$.29 \pm .03$	$\mathbf{.25 \pm .02}$	$.26 \pm .02$	$.28 \pm .02$	$.27 \pm .02$	$.26 \pm .02$
20 Newsgroups							
comp vs sci	$.41 \pm .02$	$.41 \pm .02$	$.41 \pm .02$	$.42 \pm .03$	$.40 \pm .02$	$\mathbf{.22 \pm .02}$	$\mathbf{.22 \pm .02}$
comp vs talk	$.37 \pm .03$	$.37 \pm .03$	$.37 \pm .03$	$.40 \pm .05$	$.34 \pm .03$	$\mathbf{.11 \pm .02}$	$\mathbf{.11 \pm .02}$
rec vs sci	$.43 \pm .02$	$.42 \pm .02$	$.42 \pm .02$	$.42 \pm .03$	$.41 \pm .02$	$\mathbf{.17 \pm .02}$	$\mathbf{.17 \pm .02}$
rec vs talk	$.40 \pm .03$	$.40 \pm .03$	$.40 \pm .03$	$.41 \pm .03$	$.38 \pm .03$	$\mathbf{.15 \pm .02}$	$\mathbf{.15 \pm .02}$
sci vs talk	$.41 \pm .03$	$.41 \pm .02$	$.41 \pm .02$	$.41 \pm .04$	$.39 \pm .02$	$\mathbf{.20 \pm .02}$	$\mathbf{.20 \pm .02}$

corresponding to different datasets and covariate shift situations, together with their standard deviations over 100 random partitions as detailed in Appendix E. The first column of the table describes the different covariate shift datasets generated as described above.

Overall, the experimental results show that the proposed method provides improved adaptation to general covariate shifts, even in situations where the supports of training and testing samples are not contained in each other. These results agree with the discussion in Sections 2.1 and 3.1 as well as the theoretical results in Corollary 3.3 and Theorem 4.1 that show how the proposed methodology can be effective in situations where existing methods based on single weights are challenged. The improvement obtained by the methods presented can be clearly observed by comparing the results obtained by the KMM method, since that technique is the most closely related to the proposed method. In particular, the results show that significant performance improvements can be obtained using a double weighting of both training and testing samples solving (25) instead of using the existing KMM method (that solves (25) fixing the weights α to be one).

7. Conclusion

Existing approaches for covariate shift adaptation use the ratios between marginal distributions to either weight training or testing samples. However, the performance of such approaches can be poor when the marginals' supports are not contained in each other or when marginals' ratios take large values. This paper proposes a minimax risk classification (MRC) approach for covariate shift adaptation that avoids such limitations by weighting both training and testing samples. We present effective techniques that obtain both sets of weights generalizing the conventional kernel mean matching method that only obtains weights for training samples. In addition, we present generalization bounds for the proposed methods that show a significant increase in effective sample size. The unifying approach and the learning methods proposed can enable techniques capable to adapt to more general scenarios affected by covariate shift.

Acknowledgments

Funding in direct support of this work has been provided by projects PID2019-105058GA-I00, CNS2022-135203, and CEX2021-001142-S funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR, programmes ELKARTEK and BERC-2022-2025 funded by the Basque Government, the project "Early Prognosis of COVID-19 Infections via Machine Learning" funded by the AXA Research Fund, and by the JHU-Amazon AI2AI Faculty Award.

References

Altun, Y. and Smola, A. Unifying divergence minimization and statistical inference via convex duality. In *Proceedings of the 19th Annual Conference on Computational Learning Theory*, pp. 139 – 153, 2006.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 81 – 88, 2007.

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

Bondugula, K., Alvarez, V., Segovia-Martín, J. I., Pérez, A., and Mazuelas, S. MRCpy: A library for minimax risk classifiers. *arXiv preprint arXiv:2108.01952*, 2023.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Chen, X., Monfort, M., Liu, A., and Ziebart, B. D. Robust covariate shift regression. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 1270 – 1279, 2016.

Cortes, C. and Mohri, M. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103 – 126, 2014.

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. Sample selection bias correction theory. In *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, pp. 38 – 53, 2008.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, pp. 442 – 450, 2010.

Dua, D. and Graff, C. UCI Machine Learning Repository, 2017. URL http://archive.ics.uci.edu/ml.

Dudík, M., Schapire, R. E., and Phillips, S. J. Correcting sample selection bias in maximum entropy density estimation. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, pp. 323 – 330, 2005.

Farnia, F. and Tse, D. A minimax approach to supervised learning. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, pp. 4240 – 4248, 2016.

Fathony, R., Liu, A., Asif, K., and Ziebart, B. D. Adversarial multiclass classification: A risk minimization perspective. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, pp. 559 – 567, 2016.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Covariate shift by kernel mean matching. In *Dataset shift in machine learning*, pp. 131 – 160. MIT Press, 2008.

Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M., and Schölkopf, B. Correcting sample selection bias by unlabeled data. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, pp. 601 – 608, 2006.

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

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

Liu, A. and Ziebart, B. D. Robust classification under sample selection bias. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, pp. 37 – 45, 2014.

Liu, A. and Ziebart, B. D. Robust covariate shift prediction with general losses and feature views. *arXiv preprint arXiv:1712.10043*, 2017.

Liu, S., Yamada, M., Collier, N., and Sugiyama, M. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72 – 83, 2013.

Mazaheri, B., Jain, S., and Bruck, J. Robust correction of sampling bias using cumulative distribution functions. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, pp. 3546 – 3556, 2020.

Mazuelas, S., Shen, Y., and Perez, A. Generalized maximum entropy for supervised classification. *IEEE Transactions on Information Theory*, 68(4):2530–2550, 2022.

Mazuelas, S., Romero, M., and Grünwald, P. Minimax risk classifiers with 0-1 loss. *arXiv preprint arXiv:2201.06487*, 2023.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. MIT Press, 2008.

Reddi, S. J., Póczos, B., and Smola, A. Doubly robust covariate shift correction. In *Proceedings of the 29th*

AAAI Conference on Artificial Intelligence, pp. 2949 – 2955, 2015.

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.

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

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

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

Wen, J., Yu, C.-N., and Greiner, R. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 631 – 639, 2014.

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., and Sugiyama, M. Relative density-ratio estimation for robust distribution comparison. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, pp. 594 – 602, 2011.

Yu, Y.-L. and Szepesvári, C. Analysis of kernel mean matching under covariate shift. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1147 – 1154, 2012.

Zadrozny, B. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the 21st International Conference on Machine Learning*, pp. 114, 2004.

Zhang, K., Zheng, V. W., Wang, Q., Kwok, J. T., Yang, Q., and Marsic, I. Covariate shift in hilbert space: A solution via surrogate kernels. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 388 – 395, 2013.

A. Detailed derivations describing existing methods and relation with the proposed framework

The following describes reweighted and robust methods for binary classification with $\mathcal{Y} \in \{-1, 1\}$ and log-loss. In particular, we show how, using the specific weights in (3) and (4), such methods can be obtained from Theorem 3.1 in Section 3.2 corresponding to the proposed framework.

Reweighted methods consider classification rules of the form

$$\mathrm{h}(y|x) = \frac{1}{1 + \exp\left\{-y\mathbf{x}^T\boldsymbol{\mu}\right\}} \tag{33}$$

and learn the parameter $\boldsymbol{\mu}$ using the fact that equality (3) in Section 2.1 allows to estimate expected losses with respect to the test distribution using training samples since

$$\mathbb{E}_{\mathrm{p_{te}}(x,y)} \log\left(1 + \exp\left\{-y\mathbf{x}^T\boldsymbol{\mu}\right\}\right) = \mathbb{E}_{\mathrm{p_{tr}}(x,y)}\beta(x) \log\left(1 + \exp\left\{-y\mathbf{x}^T\boldsymbol{\mu}\right\}\right).$$

for $\beta(x) = \mathrm{p_{te}}(x)/\mathrm{p_{tr}}(x)$.

Robust methods consider classification rules of the form

$$\mathrm{h}(y|x) = \frac{1}{1 + \exp\left\{-\alpha(x)y\mathbf{x}^T\boldsymbol{\mu}\right\}} \tag{34}$$

with $\alpha(x) = \mathrm{p_{tr}}(x)/\mathrm{p_{te}}(x)$. Such methods learn the parameter $\boldsymbol{\mu}$ using the fact that equality (4) in Section 2.1 allows to estimate the expected gradient of losses with respect to the test distribution using training samples since

$$\begin{aligned}\mathbb{E}_{\mathrm{p_{te}}(x,y)} \nabla_{\boldsymbol{\mu}} \log\left(1 + \exp\left\{-\frac{\mathrm{p_{tr}}(x)}{\mathrm{p_{te}}(x)} y\mathbf{x}^T\boldsymbol{\mu}\right\}\right) \\ = \mathbb{E}_{\mathrm{p_{te}}(x,y)}\alpha(x)\left(\frac{-y\mathbf{x}^T}{1 + \exp\left\{\frac{\mathrm{p_{tr}}(x)}{\mathrm{p_{te}}(x)} y\mathbf{x}^T\boldsymbol{\mu}\right\}}\right) = \mathbb{E}_{\mathrm{p_{tr}}(x,y)}\frac{-y\mathbf{x}^T}{1 + \exp\left\{\frac{\mathrm{p_{tr}}(x)}{\mathrm{p_{te}}(x)} y\mathbf{x}^T\boldsymbol{\mu}\right\}}.\end{aligned}$$

for $\alpha(x) = \mathrm{p_{tr}}(x)/\mathrm{p_{te}}(x)$.

For their derivation from the Theorem 3.1 corresponding to the proposed framework; taking $\Phi(x, y) = y\mathbf{x}/2$, we have that optimization problem in (11) of Theorem 3.1 becomes

$$\min_{\boldsymbol{\mu}} -\frac{1}{n}\sum_{i=1}^{n}\beta(x_i)\frac{y_i\mathbf{x}_i^T}{2}\boldsymbol{\mu} + \mathbb{E}_{\mathrm{p_{te}}(x)}\left\{\log\left(\exp\left\{\alpha(x)\frac{\mathbf{x}^T}{2}\boldsymbol{\mu}\right\} + \exp\left\{-\alpha(x)\frac{\mathbf{x}^T}{2}\boldsymbol{\mu}\right\}\right)\right\} \tag{35}$$

in binary classification with log-loss.

If $\alpha(x) = 1$, $\beta(x) = \mathrm{p_{te}}(x)/\mathrm{p_{tr}}(x)$, the classifier in (15) of Theorem 3.1 coincides with that of reweighted methods in (33). In addition, using (3) in Section 2.1 and approximating the expectation with training samples, the optimization in (35) becomes

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{p_{te}}(x_i)}{\mathrm{p_{tr}}(x_i)}\log\left(1 + \exp\left\{-y_i\mathbf{x}_i^T\boldsymbol{\mu}\right\}\right) \tag{36}$$

that coincides with that of reweighted logistic regression (Sugiyama & Kawanabe, 2012).

If $\alpha(x) = \mathrm{p_{tr}}(x)/\mathrm{p_{te}}(x)$, $\beta(x) = 1$, the classifier in (15) of Theorem 3.1 coincides with that of robust methods in (34). In addition, using (4) in Section 2.1, the gradient of objective function in (35) becomes

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i^T y_i}{2} + \mathbb{E}_{\mathrm{p_{tr}}(x)}\frac{\mathbf{x}^T}{2}\frac{1 - \exp\left\{-\frac{\mathrm{p_{tr}}(x)}{\mathrm{p_{te}}(x)}\mathbf{x}^T\boldsymbol{\mu}\right\}}{1 + \exp\left\{-\frac{\mathrm{p_{tr}}(x)}{\mathrm{p_{te}}(x)}\mathbf{x}^T\boldsymbol{\mu}\right\}} \tag{37}$$

that coincides with that shown in equation (7) in (Liu & Ziebart, 2014) for robust methods.

B. Proofs for Section 3

The proofs of Theorem 3.1 and 3.2 below are done for the case of finite $\mathcal{X}$. The proofs for infinite $\mathcal{X}$ can be carried out analogously using Fenchel duality instead of Lagrange duality, similarly to as is done in (Altun & Smola, 2006; Mazuelas et al., 2023).

Proof of Theorem 3.1. Firstly, for each $\mathrm{h} \in \mathrm{T}(\mathcal{X}, \mathcal{Y})$, we have that

$$\begin{aligned} \max_{\mathrm{p}\in\mathcal{U}} \ell(\mathrm{h}, \mathrm{p}) = \quad & \max_{\mathbf{p}} \quad \mathbf{l}^{\mathrm{T}}\mathbf{p} - I_+(\mathbf{p}) \\ & \text{s.t.} \quad \textstyle\sum_{y\in\mathcal{Y}} \mathrm{p}(x, y) = \mathrm{p}_{\mathrm{te}}(x),\ \forall x \in \mathcal{X} \\ & \quad\ \boldsymbol{\tau} - \boldsymbol{\lambda} \preceq \boldsymbol{\Phi}_\alpha^{\mathrm{T}}\mathbf{p} \preceq \boldsymbol{\tau} + \boldsymbol{\lambda} \end{aligned} \tag{38}$$

where $\mathbf{l}$, $\mathbf{p}$, and $\boldsymbol{\Phi}_\alpha$ denote the vectors and matrix with rows $\ell(\mathrm{h}, (x, y))$, $\mathrm{p}(x, y)$, and $\Phi_\alpha(x, y)^{\mathrm{T}}$, respectively, for $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and

$$I_+(\mathbf{p}) = \begin{cases} 0 & \text{if } \mathbf{p} \succeq \mathbf{0} \\ \infty & \text{otherwise.} \end{cases}$$

Optimization problem (38) has Lagrange dual

$$\begin{aligned} \min_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \nu(x)} \quad & -\big(\boldsymbol{\tau} - \boldsymbol{\lambda}\big)^{\mathrm{T}}\boldsymbol{\mu}_1 + \big(\boldsymbol{\tau} + \boldsymbol{\lambda}\big)^{\mathrm{T}}\boldsymbol{\mu}_2 + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)}\nu(x) + f^*(\boldsymbol{\Phi}_\alpha(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \boldsymbol{\nu}) \\ \text{s.t.} \quad & \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \succeq \mathbf{0} \end{aligned}$$

where $\boldsymbol{\nu}$ is the vector in $\mathbb{R}^{|\mathcal{X}||\mathcal{Y}|}$ with component corresponding with (x, y) for $x \in \mathcal{X}$, $y \in \mathcal{Y}$ given by $\nu(x)$, and f^* is the conjugate function of $f(\mathbf{p}) = -\mathbf{l}^{\mathrm{T}}\mathbf{p} + I_+(\mathbf{p})$ given by

$$f^*(\mathbf{w}) = \sup_{\mathbf{p}\succeq\mathbf{0}} \mathbf{w}^{\mathrm{T}}\mathbf{p} + \mathbf{l}^{\mathrm{T}}\mathbf{p} = \begin{cases} 0 & \text{if } \mathbf{w} \preceq -\mathbf{l} \\ \infty & \text{otherwise} \end{cases}.$$

Therefore, the Lagrange dual above becomes

$$\begin{aligned} \min_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \nu(x)} \quad & -\big(\boldsymbol{\tau} - \boldsymbol{\lambda}\big)^{\mathrm{T}}\boldsymbol{\mu}_1 + \big(\boldsymbol{\tau} + \boldsymbol{\lambda}\big)^{\mathrm{T}}\boldsymbol{\mu}_2 + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)}\nu(x) \\ \text{s.t.} \quad & \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \succeq \mathbf{0} \\ & \Phi_\alpha(x, y)^{\mathrm{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \nu(x) \leq -\ell(\mathrm{h}, (x, y)),\ \forall x \in \mathcal{X}, y \in \mathcal{Y}. \end{aligned}$$

It is easy to see that the solution of such optimization problem $\bar{\boldsymbol{\mu}}_1, \bar{\boldsymbol{\mu}}_2$ satisfies that $\bar{\mu}_1^{(i)}\bar{\mu}_2^{(i)} = 0$ for any i such that $\lambda^{(i)} > 0$. Then $\boldsymbol{\lambda}^{\mathrm{T}}(\bar{\boldsymbol{\mu}}_1 + \bar{\boldsymbol{\mu}}_2) = \boldsymbol{\lambda}^{\mathrm{T}}|\bar{\boldsymbol{\mu}}_1 - \bar{\boldsymbol{\mu}}_2|$ and taking $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ the Lagrange dual above is equivalent to

$$\begin{aligned} \min_{\boldsymbol{\mu}, \nu(x)} \quad & -\boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{\mu} + \boldsymbol{\lambda}^{\mathrm{T}}|\boldsymbol{\mu}| + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)}\nu(x) \\ & \Phi_\alpha(x, y)^{\mathrm{T}}\boldsymbol{\mu} - \nu(x) \leq -\ell(\mathrm{h}, (x, y)),\ \forall x \in \mathcal{X}, y \in \mathcal{Y} \end{aligned}$$

that has the same value as $\max_{\mathrm{p}\in\mathcal{U}} \ell(\mathrm{h}, \mathrm{p})$ since the constraints in (38) are affine and $\mathcal{U}$ is non-empty.

Therefore,

$$\begin{aligned} \min_{\mathrm{h}\in\mathrm{T}(\mathcal{X},\mathcal{Y})} \max_{\mathrm{p}\in\mathcal{U}} \ell(\mathrm{h}, \mathrm{p}) = \min_{\mathrm{h}, \boldsymbol{\mu}, \nu(x)} \quad & -\boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{\mu} + \boldsymbol{\lambda}^{\mathrm{T}}|\boldsymbol{\mu}| + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)}\nu(x) \\ & \Phi_\alpha(x, y)^{\mathrm{T}}\boldsymbol{\mu} - \nu(x) \leq -\ell(\mathrm{h}, (x, y)),\ \forall x \in \mathcal{X}, y \in \mathcal{Y}. \end{aligned}$$

For 0-1-loss we have that

$$\begin{aligned} & \Phi_\alpha(x, y)^{\mathrm{T}}\boldsymbol{\mu} - \nu(x) \leq -1 + \mathrm{h}(y|x),\ \forall x \in \mathcal{X}, y \in \mathcal{Y} \\ & \quad \Rightarrow \sum_{y\in\mathcal{C}} \big(\Phi_\alpha(x, y)^{\mathrm{T}}\boldsymbol{\mu} - \nu(x) + 1\big) \leq 1,\ \forall \mathcal{C} \subseteq \mathcal{Y}, x \in \mathcal{X} \\ & \quad \Rightarrow \nu(x) \geq 1 + \frac{\sum_{y\in\mathcal{C}} \Phi_\alpha(x, y)^{\mathrm{T}}\boldsymbol{\mu} - 1}{|\mathcal{C}|},\ \forall \mathcal{C} \subseteq \mathcal{Y}, x \in \mathcal{X} \\ & \quad \Rightarrow \nu(x) \geq \varphi_{01}(\boldsymbol{\mu}, x, \alpha(x)),\ \forall x \in \mathcal{X}. \end{aligned}$$

Therefore, for each $\boldsymbol{\mu}$, we have that any classification rule satisfying

$$\mathrm{h}(y|x) \geq \Phi_\alpha(x,y)^{\mathrm{T}}\boldsymbol{\mu} - \varphi_{01}(\boldsymbol{\mu}, x, \alpha(x)) + 1,\ \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

is solution of

$$\begin{aligned} \min_{\mathrm{h},\nu(x)} \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)}\nu(x) \qquad & = \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)}\varphi_{01}(\boldsymbol{\mu}, x, \alpha(x)) \\ \Phi_\alpha(x,y)^{\mathrm{T}}\boldsymbol{\mu} - \nu(x) + 1 \leq \mathrm{h}(y|x),\ \forall x \in \mathcal{X}, y \in \mathcal{Y} & \end{aligned}$$

and the result is obtained because for any $x \in \mathcal{X}$, we have that

$$\sum_{y\in\mathcal{Y}} \left(\Phi_\alpha(x,y)^{\mathrm{T}}\boldsymbol{\mu} - \varphi_{01}(\boldsymbol{\mu}, x, \alpha(x)) + 1\right)_+ = 1$$

because otherwise there would exist $\nu_x < \varphi_{01}(\boldsymbol{\mu}, x, \alpha(x))$ such that

$$1 = \sum_{y\in\mathcal{Y}} \left(\Phi_\alpha(x,y)^{\mathrm{T}}\boldsymbol{\mu} - \nu_x + 1\right)_+ = \max_{\mathcal{C}\subseteq\mathcal{Y}} \sum_{y\in\mathcal{C}} \left(\Phi_\alpha(x,y)^{\mathrm{T}}\boldsymbol{\mu} - \nu_x + 1\right)$$

which contradicts the definition of $\varphi_{01}(\boldsymbol{\mu}, x, \alpha(x))$.

The case of log-loss is analogous to the case for 0-1-loss above taking into account that

$$\begin{aligned} \Phi_\alpha(x,y)^{\mathrm{T}}\boldsymbol{\mu} - \nu(x) \leq \log(\mathrm{h}(y|x)),\ \forall x \in \mathcal{X}, y \in \mathcal{Y} & \\ \Rightarrow \sum_{y\in\mathcal{Y}} \exp\{\Phi_\alpha(x,y)^{\mathrm{T}}\boldsymbol{\mu} - \nu(x)\} \leq 1,\ \forall x \in \mathcal{X} & \\ \Rightarrow \nu(x) \geq \log\big(\sum_{y\in\mathcal{Y}} \exp\{\Phi_\alpha(x,y)^{\mathrm{T}}\boldsymbol{\mu}\}\big),\ \forall x \in \mathcal{X} & \\ \Rightarrow \nu(x) \geq \varphi_{\log}(\boldsymbol{\mu}, x, \alpha(x)),\ \forall x \in \mathcal{X}. & \end{aligned}$$

□

The lemma below is used in the proof of Theorem 3.2.

Lemma B.1. *Let $\mathcal{U}$ be the uncertainty set given by (9) for $\boldsymbol{\tau}, \boldsymbol{\lambda} \in \mathbb{R}^m$, and h be a classification rule. If*

$$\overline{R}_{01}(\mathcal{U}, \mathrm{h}) = \min_{\boldsymbol{\mu}} -\boldsymbol{\tau}^T\boldsymbol{\mu} + \mathbb{E}_{\mathrm{p}_{te}(x)} \max_{y\in\mathcal{Y}} \left\{1 + \alpha(x)\Phi(x,y)^T\boldsymbol{\mu} - \mathrm{h}(y|x)\right\} + \boldsymbol{\lambda}^T|\boldsymbol{\mu}| \tag{39}$$

$$\overline{R}_{\log}(\mathcal{U}, \mathrm{h}) = \min_{\boldsymbol{\mu}} -\boldsymbol{\tau}^T\boldsymbol{\mu} + \mathbb{E}_{\mathrm{p}_{te}(x)} \max_{y\in\mathcal{Y}} \left\{\alpha(x)\Phi(x,y)^T\boldsymbol{\mu} - \log\mathrm{h}(y|x)\right\} + \boldsymbol{\lambda}^T|\boldsymbol{\mu}| \tag{40}$$

then, for any $\mathrm{p} \in \mathcal{U}$

$$\ell_{01}(\mathrm{h}, \mathrm{p}) \leq \overline{R}_{01}(\mathcal{U}, \mathrm{h}) \tag{41}$$

$$\ell_{\log}(\mathrm{h}, \mathrm{p}) \leq \overline{R}_{\log}(\mathcal{U}, \mathrm{h}). \tag{42}$$

Proof of Lemma B.1. The proof is analogous to the proof of Theorem 5 of (Mazuelas et al., 2023). The case $\mathcal{U} = \emptyset$ is trivial. For the case where $\mathcal{U} \neq \emptyset$, we will first calculate the Lagrange dual of the optimization problem $\min_{\hat{\mathrm{p}}\in\mathcal{U}} \mathbb{E}_{\hat{\mathrm{p}}} q$ for a general function $q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Then we will consider the fact that for any $\mathrm{p} \in \mathcal{U}$ and $\mathrm{h} \in \mathrm{T}(\mathcal{X}, \mathcal{Y})$,

$$\min_{\hat{\mathrm{p}}\in\mathcal{U}} \ell(\mathrm{h}, \hat{\mathrm{p}}) \leq \ell(\mathrm{h}, \mathrm{p}) \leq \max_{\hat{\mathrm{p}}\in\mathcal{U}} \ell(\mathrm{h}, \hat{\mathrm{p}})$$

and

$$\begin{aligned} \max_{\hat{\mathrm{p}}\in\mathcal{U}} \ell_{01}(\mathrm{h}, \hat{\mathrm{p}}) &= -\min_{\hat{\mathrm{p}}\in\mathcal{U}} \mathbb{E}_{\hat{\mathrm{p}}} \left\{\mathrm{h}(y|x) - 1\right\} \\ \max_{\hat{\mathrm{p}}\in\mathcal{U}} \ell_{\log}(\mathrm{h}, \hat{\mathrm{p}}) &= -\min_{\hat{\mathrm{p}}\in\mathcal{U}} \mathbb{E}_{\hat{\mathrm{p}}} \log\mathrm{h}(y|x) \end{aligned}$$

for 0-1-loss and log-loss respectively.

First, we have that $\min_{\hat{\mathrm{p}}\in\mathcal{U}} \mathbb{E}_{\hat{\mathrm{p}}} q$ is equal to

$$\begin{aligned} \min_{\hat{\mathbf{p}}} \;& \boldsymbol{q}^T\hat{\mathbf{p}} + I_+(\hat{\mathbf{p}}) \\ \text{s.t.} \;& -\sum_{y\in\mathcal{Y}} \hat{\mathrm{p}}(x,y) = -\mathrm{p}_{\text{te}}(x) \text{ for all } x\in\mathcal{X} \\ & \boldsymbol{\tau}-\boldsymbol{\lambda} \preceq \boldsymbol{\Phi}_\alpha^T\hat{\mathbf{p}} \preceq \boldsymbol{\tau}+\boldsymbol{\lambda} \end{aligned} \tag{43}$$

where $\hat{\mathbf{p}}$, $\boldsymbol{q}$, $\boldsymbol{\Phi}_\alpha$ denote the vectors and matrix with rows $\hat{\mathrm{p}}(x,y)$, $q(x,y)$ and $\alpha(x)\Phi(x,y)^T$, respectively, for $x\in\mathcal{X}$, $y\in\mathcal{Y}$, and

$$I_+(\hat{\mathbf{p}}) = \begin{cases} 0 & \text{if } \hat{\mathbf{p}} \succeq \mathbf{0} \\ \infty & \text{otherwise.} \end{cases}$$

Optimization problem (43) has Lagrange dual

$$\begin{aligned} \max_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\nu(x)} \;& (\boldsymbol{\tau}-\boldsymbol{\lambda})^T\boldsymbol{\mu}_1 - (\boldsymbol{\tau}+\boldsymbol{\lambda})^T\boldsymbol{\mu}_2 + \mathbb{E}_{\mathrm{p}_{\text{te}}(x)}\nu(x) - f^*\left(\boldsymbol{\Phi}_\alpha(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\boldsymbol{\nu}\right) \\ \text{s.t.} \;& \boldsymbol{\mu}_1,\boldsymbol{\mu}_2 \succeq \mathbf{0} \end{aligned}$$

where $\boldsymbol{\nu}$ denotes the vector in $\mathbb{R}^{|\mathcal{X}||\mathcal{Y}|}$ with component corresponding with (x,y) for $x\in\mathcal{X}$, $y\in\mathcal{Y}$ given by $\nu(x)$, and f^* is the conjugate function of $f(\hat{\mathbf{p}}) = \mathbf{q}^T\hat{\mathbf{p}} + I_+(\hat{\mathbf{p}})$ that becomes

$$f^*(\boldsymbol{w}) = \begin{cases} 0 & \text{if } \mathbf{w} \preceq \mathbf{q} \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, the previous Lagrange dual becomes

$$\begin{aligned} \max_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\nu(x)} \;& (\boldsymbol{\tau}-\boldsymbol{\lambda})^T\boldsymbol{\mu}_1 - (\boldsymbol{\tau}+\boldsymbol{\lambda})^T\boldsymbol{\mu}_2 + \mathbb{E}_{\mathrm{p}_{\text{te}}(x)}\nu(x) \\ \text{s.t.} \;& \boldsymbol{\mu}_1,\boldsymbol{\mu}_2 \succeq \mathbf{0} \\ & \boldsymbol{\Phi}_\alpha(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\boldsymbol{\nu} \preceq \boldsymbol{q} \end{aligned}$$

which is equivalent to

$$\begin{aligned} \max_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2} \;& (\boldsymbol{\tau}-\boldsymbol{\lambda})^T\boldsymbol{\mu}_1 - (\boldsymbol{\tau}+\boldsymbol{\lambda})^T\boldsymbol{\mu}_2 + \mathbb{E}_{\mathrm{p}_{\text{te}}(x)} \min_{y\in\mathcal{Y}} \left\{q(x,y) - \alpha(x)\Phi(x,y)^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right\} \\ \text{s.t.} \;& \boldsymbol{\mu}_1,\boldsymbol{\mu}_2 \succeq \mathbf{0}. \end{aligned}$$

Taking $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ the Lagrange dual problem is equivalent to

$$\max_{\boldsymbol{\mu}} \boldsymbol{\tau}^T\boldsymbol{\mu} + \mathbb{E}_{\mathrm{p}_{\text{te}}(x)} \min_{y\in\mathcal{Y}} \left\{q(x,y) - \alpha(x)\Phi(x,y)^T\boldsymbol{\mu}\right\} - \boldsymbol{\lambda}^T|\boldsymbol{\mu}|$$

that has the same value as its primal $\min_{\hat{\mathrm{p}}\in\mathcal{U}} \mathbb{E}_{\hat{\mathrm{p}}} q$ since the constraints defining $\mathcal{U}$ are affine and $\mathcal{U} \neq \emptyset$. Then, we have that

$$\begin{aligned} \max_{\hat{\mathrm{p}}\in\mathcal{U}} \ell_{01}(\mathrm{h},\hat{\mathrm{p}}) &= -\min_{\hat{\mathrm{p}}\in\mathcal{U}} \mathbb{E}_{\hat{\mathrm{p}}}\left\{\mathrm{h}(y|x) - 1\right\} = \min_{\boldsymbol{\mu}} -\boldsymbol{\tau}^T\boldsymbol{\mu} + \mathbb{E}_{\mathrm{p}_{\text{te}}(x)} \max_{y\in\mathcal{Y}} \left\{1 + \alpha(x)\Phi(x,y)^T\boldsymbol{\mu} - \mathrm{h}(y|x)\right\} + \boldsymbol{\lambda}^T|\boldsymbol{\mu}| \\ \max_{\hat{\mathrm{p}}\in\mathcal{U}} \ell_{\log}(\mathrm{h},\hat{\mathrm{p}}) &= -\min_{\hat{\mathrm{p}}\in\mathcal{U}} \mathbb{E}_{\hat{\mathrm{p}}} \log \mathrm{h}(y|x) \quad = \min_{\boldsymbol{\mu}} -\boldsymbol{\tau}^T\boldsymbol{\mu} + \mathbb{E}_{\mathrm{p}_{\text{te}}(x)} \max_{y\in\mathcal{Y}} \left\{\alpha(x)\Phi(x,y)^T\boldsymbol{\mu} - \log \mathrm{h}(y|x)\right\} + \boldsymbol{\lambda}^T|\boldsymbol{\mu}| \end{aligned}$$

for 0-1-loss and log-loss respectively. □

Proof of Theorem 3.2. For inequality (19), let $\mathcal{U}_\infty$ be the uncertainty set given by the exact mean vector $\tau_\infty = \mathbb{E}_{\mathrm{p}_{\text{te}}}\Phi_\alpha(x,y)$, i.e.,

$$\begin{aligned} \mathcal{U}_\infty = \{\mathrm{p}\in\Delta(\mathcal{X}\times\mathcal{Y}) : & |\mathbb{E}_{\mathrm{p}}\Phi_\alpha(x,y) - \boldsymbol{\tau}_\infty| \preceq \boldsymbol{\lambda} \\ & \text{and } \mathrm{p}(x) = \mathrm{p}_{\text{te}}(x), \forall x\in\mathcal{X}\}. \end{aligned} \tag{44}$$

It is clear that we have $\mathrm{p}_{\mathrm{te}}(x,y) \in \mathcal{U}_\infty$, then for 0-1-loss, using Lemma B.1 and the definition of $\mathrm{h}(y|x)$ in (14), we have that

$$\begin{aligned}
R(\mathrm{h}^{\mathcal{U}}) \leq \overline{R}_{01}(\mathcal{U}_\infty, \mathrm{h}^{\mathcal{U}}) &= \min_{\boldsymbol{\mu}} -\boldsymbol{\tau}_\infty^T \boldsymbol{\mu} + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)} \max_{y\in\mathcal{Y}} \left\{1 + \alpha(x)\Phi(x,y)^T\boldsymbol{\mu} - \mathrm{h}(y|x)\right\} \\
&\leq -\boldsymbol{\tau}_\infty^T \boldsymbol{\mu}^* + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)} \max_{y\in\mathcal{Y}} \left\{1 + \alpha(x)\Phi(x,y)^T\boldsymbol{\mu}^* - \mathrm{h}(y|x)\right\} \quad (45)\\
&\leq -\boldsymbol{\tau}_\infty^T \boldsymbol{\mu}^* + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)} \max_{y\in\mathcal{Y}} \varphi_{01}(\boldsymbol{\mu}^*, x, \alpha(x)) \quad (46)\\
&= -\boldsymbol{\tau}_\infty^T \boldsymbol{\mu}^* + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)} \varphi_{01}(\boldsymbol{\mu}^*, x, \alpha(x)) \quad (47)\\
&= R(\mathcal{U}) + (\boldsymbol{\tau} - \boldsymbol{\tau}_\infty)^T \boldsymbol{\mu}^* - \boldsymbol{\lambda}^T |\boldsymbol{\mu}^*|. \quad (48)
\end{aligned}$$

where, for inequality (45)-(46), we have used the fact that $\mathrm{h}(y|x) \geq \alpha(x)\Phi(x,y)^T\boldsymbol{\mu}^* + 1 - \varphi_{01}(\boldsymbol{\mu}^*, x, \alpha(x))$ and for inequality (47)-(48) we have added and subtracted $\boldsymbol{\tau}^T\boldsymbol{\mu}^*$ and $\boldsymbol{\lambda}^T\boldsymbol{\mu}^*$, and used the definition of $R(\mathcal{U})$ in (16).

For log-loss, using Lemma B.1 and the definition of $\mathrm{h}(y|x)$ in (15), we have that

$$\begin{aligned}
R(\mathrm{h}^{\mathcal{U}}) \leq \overline{R}_{\log}(\mathcal{U}_\infty, \mathrm{h}^{\mathcal{U}}) &= \min_{\boldsymbol{\mu}} -\boldsymbol{\tau}_\infty^T \boldsymbol{\mu} + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)} \max_{y\in\mathcal{Y}} \left\{\alpha(x)\Phi(x,y)^T\boldsymbol{\mu} - \log \mathrm{h}(y|x)\right\} \\
&\leq -\boldsymbol{\tau}_\infty^T \boldsymbol{\mu}^* + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)} \max_{y\in\mathcal{Y}} \left\{\alpha(x)\Phi(x,y)^T\boldsymbol{\mu}^* - \log \mathrm{h}(y|x)\right\} \quad (49)\\
&= -\boldsymbol{\tau}_\infty^T \boldsymbol{\mu}^* + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)} \max_{y\in\mathcal{Y}} \varphi_{\log}(\boldsymbol{\mu}^*, x, \alpha(x)) \quad (50)\\
&= -\boldsymbol{\tau}_\infty^T \boldsymbol{\mu}^* + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)} \varphi_{\log}(\boldsymbol{\mu}^*, x, \alpha(x)) \quad (51)\\
&= R(\mathcal{U}) + (\boldsymbol{\tau} - \boldsymbol{\tau}_\infty)^T \boldsymbol{\mu}^* - \boldsymbol{\lambda}^T |\boldsymbol{\mu}^*| \quad (52)
\end{aligned}$$

where, for inequality (49)-(50) we have used the fact that $\log \mathrm{h}(y|x) = \alpha(x)\Phi(x,y)^T\boldsymbol{\mu}^* - \varphi_{\log}(\boldsymbol{\mu}^*, x, \alpha(x))$ and for inequality (51)-(52) we have added and subtracted $\boldsymbol{\tau}^T\boldsymbol{\mu}^*$ and $\boldsymbol{\lambda}^T\boldsymbol{\mu}^*$, and used the definition of $R(\mathcal{U})$ in (16).

For inequality (20), note that using the definition of $\boldsymbol{\mu}^*$ and (47) (resp. (51)) for 0-1-loss (resp. log-loss), we have that

$$\begin{aligned}
R(\mathrm{h}^{\mathcal{U}}) &\leq -\boldsymbol{\tau}_\infty^T \boldsymbol{\mu}^* + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)} \varphi_\ell(\boldsymbol{\mu}^*, x, \alpha(x)) \\
&\leq -\boldsymbol{\tau}^T \boldsymbol{\mu}_\infty + \mathbb{E}_{\mathrm{p}_{\mathrm{te}}(x)} \varphi_\ell(\boldsymbol{\mu}_\infty, x, \alpha(x)) + \boldsymbol{\lambda}^T |\boldsymbol{\mu}_\infty| + (\boldsymbol{\tau} - \boldsymbol{\tau}_\infty)^T \boldsymbol{\mu}^* - \boldsymbol{\lambda}^T |\boldsymbol{\mu}^*| \\
&= R^\infty + \boldsymbol{\lambda}^T \left(|\boldsymbol{\mu}_\infty| - |\boldsymbol{\mu}^*|\right) + (\boldsymbol{\tau}_\infty - \boldsymbol{\tau})^T \boldsymbol{\mu}_\infty + (\boldsymbol{\tau} - \boldsymbol{\tau}_\infty)^T \boldsymbol{\mu}^* \\
&\leq R^\infty + \boldsymbol{\lambda}^T \left(|\boldsymbol{\mu}_\infty| - |\boldsymbol{\mu}^*|\right) + |\boldsymbol{\tau} - \boldsymbol{\tau}_\infty|^T |\boldsymbol{\mu}_\infty - \boldsymbol{\mu}^*|.
\end{aligned}$$

□

C. Proofs for Section 4

Proof of Theorem 4.1. The proof is analogous to Example 6.3 in (Boucheron et al., 2013) that shows a Hoeffding-type inequality in Hilbert space.

We consider $n+t$ independent random variables taking values in the Hilbert space $\mathcal{H}$ as follows

$$f_i = \begin{cases} \frac{1}{n}\hat{\beta}(x_i)K(x_i) & \text{if } i = 1, 2, \ldots, n \\ -\frac{1}{t}\hat{\alpha}(x_i)K(x_i) & \text{if } i = n+1, n+2, \ldots, n+t. \end{cases} \quad (53)$$

and we want to bound $||\sum_{i=1}^{n+t} f_i||_{\mathcal{H}}$. We have that,

$$||f_i||_{\mathcal{H}} \leq \begin{cases} \frac{1}{n}\frac{B}{\sqrt{D}}\kappa & \text{if } i = 1, 2, \ldots, n \\ \frac{1}{t}\kappa & \text{if } i = n+1, n+2, \ldots, n+t. \end{cases} \quad (54)$$

Taking $v = \kappa^2 \left(\frac{B^2}{Dn} + \frac{1}{t} \right)$ and using the bounded differences inequality (Theorem 6.2 in (Boucheron et al., 2013)), we have that, for all $l \geq \sqrt{v}$

$$\begin{aligned} \mathbb{P}\left\{ \left\| \sum_{i=1}^{n+t} f_i \right\|_{\mathcal{H}} > l \right\} &= \mathbb{P}\left\{ \left\| \sum_{i=1}^{n+t} f_i \right\|_{\mathcal{H}} - \mathbb{E}\left\| \sum_{i=1}^{n+t} f_i \right\|_{\mathcal{H}} > l - \mathbb{E}\left\| \sum_{i=1}^{n+t} f_i \right\|_{\mathcal{H}} \right\} \\ &\leq \exp\left\{ - \frac{\left(l - \mathbb{E}\left\| \sum_{i=1}^{n+t} f_i \right\|_{\mathcal{H}} \right)^2}{2v} \right\}. \end{aligned} \tag{55}$$

Finally, using Hölder's inequality and by independence, we have that

$$\mathbb{E}\left\| \sum_{i=1}^{n+t} f_i \right\|_{\mathcal{H}} \leq \sqrt{\mathbb{E}\left\| \sum_{i=1}^{n+t} f_i \right\|_{\mathcal{H}}^2} = \sqrt{\sum_{i=1}^{n+t} \mathbb{E}\,||f_i||_{\mathcal{H}}^2} \leq \sqrt{v}.$$

Therefore,

$$\exp\left\{ - \frac{(l - \sqrt{v})^2}{2v} \right\} = \exp\left\{ - \frac{\left(l - \sqrt{\kappa^2 \left(\frac{B^2}{Dn} + \frac{1}{t} \right)} \right)^2}{2\kappa^2 \left(\frac{B^2}{Dn} + \frac{1}{t} \right)} \right\}$$

so that,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}(x_i) K(x_i) - \frac{1}{t} \sum_{i=n+1}^{n+t} \hat{\alpha}(x_{n+i}) K(x_{n+i}) \right\|_{\mathcal{H}} \leq \left(1 + \sqrt{2 \log \frac{1}{\delta}} \right) \kappa \sqrt{\left(\frac{B^2}{Dn} + \frac{1}{t} \right)}$$

with probability at least $1 - \delta$. □

D. Quadratic version of DW-KMM

The convex optimization in (25) is a quadratic problem since the squared norm in $\mathcal{H}$ can be written as

$$\begin{aligned} &\left\| \frac{1}{t} \sum_{i=1}^{t} \alpha^{(i)} K(x_{n+i}) - \frac{1}{n} \sum_{i=1}^{n} \beta^{(i)} K(x_i) \right\|_{\mathcal{H}}^2 \\ &= \frac{1}{t^2} \sum_{i,j=1}^{t} \alpha^{(i)} \alpha^{(j)} k(x_{n+i}, x_{n+j}) + \frac{1}{n^2} \sum_{i,j=1}^{n} \beta^{(i)} \beta^{(j)} k(x_i, x_j) - \frac{2}{nt} \sum_{i=1}^{t} \sum_{j=1}^{n} \alpha^{(i)} \beta^{(j)} k(x_{n+i}, x_j) \\ &= \frac{\boldsymbol{\alpha}^T}{t} \begin{bmatrix} k(x_{n+1}, x_{n+1}) & \cdots & k(x_{n+1}, x_{n+t}) \\ \vdots & \ddots & \vdots \\ k(x_{n+t}, x_{n+1}) & \cdots & k(x_{n+t}, x_{n+t}) \end{bmatrix} \frac{\boldsymbol{\alpha}}{t} + \frac{\boldsymbol{\beta}^T}{n} \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} \frac{\boldsymbol{\beta}}{n} \\ &\quad - 2 \frac{\boldsymbol{\beta}^T}{n} \begin{bmatrix} k(x_1, x_{n+1}) & \cdots & k(x_1, x_{n+t}) \\ \vdots & \ddots & \vdots \\ k(x_n, x_{n+1}) & \cdots & k(x_n, x_{n+t}) \end{bmatrix} \frac{\boldsymbol{\alpha}}{t} \\ &= \left[\boldsymbol{\beta}^T / n, -\boldsymbol{\alpha}^T / t \right] \mathbf{K} \begin{bmatrix} \boldsymbol{\beta}/n \\ -\boldsymbol{\alpha}/t \end{bmatrix} \end{aligned}$$

where $\mathbf{K}$ is the kernel matrix given by $\mathrm{K}^{(i,j)} = k(x_i, x_j)$.

Therefore, the optimization problem (25) is equivalent to the quadratic optimization problem

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \ & \left[\boldsymbol{\beta}^T/n, -\boldsymbol{\alpha}^T/t\right]\mathbf{K}\begin{bmatrix}\boldsymbol{\beta}/n \\ -\boldsymbol{\alpha}/t\end{bmatrix} \\ \text{s.t.} \ & \mathbf{0} \preceq \boldsymbol{\beta} \preceq (B/\sqrt{D})\mathbf{1}, \quad \mathbf{0} \preceq \boldsymbol{\alpha} \preceq \mathbf{1} \\ & \left|\boldsymbol{\beta}^T\mathbf{1}/n - \boldsymbol{\alpha}^T\mathbf{1}/t\right| \leq \epsilon \\ & ||\boldsymbol{\alpha} - \mathbf{1}|| \leq \left(1 - \frac{1}{\sqrt{D}}\right)\sqrt{t}. \end{aligned} \tag{56}$$

E. Implementation details and additional experimental results

This appendix details the datasets and settings used for the experiments in Section 6 and shows additional experiments.

For the experiments in Section 6, we have considered four binary classification datasets, available in the UCI repository (Dua & Graff, 2017), and previously used in multiple papers on covariate shift adaptation (Gretton et al., 2008; Huang et al., 2006; Kanamori et al., 2009; Mazaheri et al., 2020; Wen et al., 2014). In addition, we use the dataset "News20groups" that is intrinsically affected by covariate shift (Zhang et al., 2013).

Table 2 details the characteristics of the datasets used in the experiments. The table also shows the parameter σ used in the computation of the kernel matrix $\mathbf{K}$ for the RuLSIF, KMM and DW-KMM methods, which is determined using the common heuristic based on nearest neighbors with $K = 50$, as is done in (Wen et al., 2014). For the results obtained using the flattening method in (Shimodaira, 2000) and the RuLSIF method in (Yamada et al., 2011) we considered the hyperparameter $\gamma = 0.5$, which is the default value used in those papers.

Table 2. Datasets used in the experiments.

Dataset	Covariates	Samples		Ratio of majority class	σ
Blood	3	748		76.20%	0.7491
BreastCancer	9	683		65.01%	1.6064
Haberman	3	306		75.53%	1.3024
Ringnorm	20	7400		50.49%	3.8299
comp vs sci	1000	5309	3534	55.31%	23.5628
comp vs talk	1000	4888	3256	60.06%	23.4890
rec vs sci	1000	4762	3169	50.17%	24.5642
rec vs talk	1000	4341	2891	55.02%	25.1129
sci vs talk	1000	4325	2880	54.85%	24.8320

In the additional experiments we study the effectiveness of the proposed selection method for hyperparameter D. Specifically, Tables 3 and 4 show the average classification error varying the value of D for the datasets and covariate shifts shown in Table 1. The first column of these tables shows the classification error obtained when selecting D with the proposed method that minimizes the minimax risk, while the other columns show the classification error obtained using specific values of D. The values of the hyperparameter D have been chosen based on the last inequality in the optimization problem (25). Specifically, we take the values for D such that $1 - 1/\sqrt{D} \in \{0, 0.1, \ldots, 0.9\}$. As can be seen from the tables, the proposed selection method results in performances near those obtained with the best values of D.

Table 3. Classification error in 21 scenarios using DW-GCS methods with 0-1-loss varying the value of the hyperparameter D.

Dataset	proposed selection	$D = 1$	$D = 1.2$	$D = 1.6$	$D = 2$	$D = 2.8$	$D = 4$	$D = 6.3$	$D = 11$	$D = 25$	$D = 100$
Blood											
Feature 1	0.30	0.32	0.31	0.31	0.31	0.30	0.30	0.30	**0.29**	**0.29**	0.30
Feature 2	0.38	0.40	0.40	0.40	0.40	0.39	0.39	0.38	0.38	**0.37**	0.38
Feature 3	0.34	0.39	0.37	0.36	0.36	0.35	0.34	0.34	0.34	**0.33**	0.34
PCA	0.28	0.32	0.30	0.29	0.28	**0.27**	**0.27**	0.28	0.28	0.28	0.28
BreastCancer											
Feature 1	**0.04**	0.05	0.05	**0.04**	**0.04**	**0.04**	**0.04**	0.05	0.05	**0.04**	**0.04**
Feature 2	**0.04**	0.06	0.06	0.05	0.05	0.05	0.06	0.06	0.06	**0.04**	**0.04**
Feature 3	**0.04**	0.06	0.05	0.05	0.05	0.06	0.05	0.06	0.05	**0.04**	**0.04**
PCA	**0.02**	0.03	0.03	**0.02**	**0.02**	0.03	0.03	0.03	**0.02**	**0.02**	**0.02**
Haberman											
Feature 1	**0.28**	0.39	0.36	0.33	0.30	0.29	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**
Feature 2	**0.29**	0.39	0.37	0.35	0.33	0.32	0.30	0.30	0.30	0.30	**0.29**
Feature 3	0.35	0.46	0.45	0.43	0.40	0.37	0.35	0.35	**0.34**	**0.34**	**0.34**
PCA	0.30	0.40	0.38	0.35	0.32	0.30	**0.29**	**0.29**	**0.29**	0.30	**0.29**
Ringnorm											
Feature 1	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**
Feature 2	**0.25**	**0.25**	0.26	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**
Feature 3	**0.25**	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	**0.25**	**0.25**
PCA	0.27	0.30	0.29	0.29	0.28	0.28	0.27	0.27	**0.26**	0.27	0.27
20 Newsgroups											
comp vs sci	0.22	0.25	0.24	0.23	0.22	**0.21**	**0.21**	**0.21**	**0.21**	**0.21**	**0.21**
comp vs talk	0.11	0.17	0.16	0.14	0.12	0.11	**0.10**	**0.10**	**0.10**	0.11	0.11
rec vs sci	0.17	0.19	0.18	0.18	0.17	0.17	**0.16**	**0.16**	**0.16**	**0.16**	**0.16**
rec vs talk	0.15	0.18	0.17	0.16	0.15	0.15	**0.14**	**0.14**	**0.14**	**0.14**	**0.14**
sci vs talk	0.20	0.22	0.21	0.20	0.19	0.19	**0.18**	**0.18**	**0.18**	0.19	0.19

Table 4. Classification error in 21 scenarios using DW-GCS methods with log-loss varying the value of the hyperparameter D.

Dataset	proposed selection	$D = 1$	$D = 1.2$	$D = 1.6$	$D = 2$	$D = 2.8$	$D = 4$	$D = 6.3$	$D = 11$	$D = 25$	$D = 100$
Blood											
Feature 1	0.31	0.32	0.32	0.32	0.31	0.30	0.30	0.30	**0.29**	**0.29**	0.30
Feature 2	**0.38**	0.41	0.40	0.40	0.40	0.39	**0.38**	**0.38**	**0.38**	**0.38**	**0.38**
Feature 3	0.35	0.38	0.38	0.37	0.36	0.36	0.35	**0.34**	**0.34**	**0.34**	**0.34**
PCA	**0.28**	0.32	0.30	0.29	0.29	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**
BreastCancer											
Feature 1	**0.04**	0.05	0.05	0.05	0.05	**0.04**	**0.04**	0.05	0.05	**0.04**	**0.04**
Feature 2	**0.04**	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	**0.04**	**0.04**
Feature 3	**0.04**	0.06	0.06	0.06	0.06	0.06	0.05	0.06	0.05	**0.04**	**0.04**
PCA	**0.02**	0.03	0.03	0.03	0.03	0.03	0.03	0.03	**0.02**	**0.02**	**0.02**
Haberman											
Feature 1	0.29	0.38	0.36	0.34	0.31	0.30	0.29	0.29	0.29	0.29	**0.28**
Feature 2	**0.30**	0.39	0.37	0.35	0.33	0.32	0.31	0.31	0.31	0.31	**0.30**
Feature 3	0.36	0.46	0.44	0.43	0.40	0.37	0.36	0.35	0.35	0.35	**0.34**
PCA	0.31	0.39	0.38	0.36	0.33	0.31	**0.30**	**0.30**	**0.30**	0.31	**0.30**
Ringnorm											
Feature 1	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**
Feature 2	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	0.24	**0.25**
Feature 3	**0.25**	0.26	0.26	0.26	**0.25**	0.26	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**
PCA	**0.26**	0.30	0.29	0.29	0.28	0.27	0.27	**0.26**	**0.26**	**0.26**	0.27
20 Newsgroups											
comp vs sci	0.22	0.25	0.24	0.23	0.22	**0.21**	**0.21**	**0.21**	**0.21**	**0.21**	**0.21**
comp vs talk	0.11	0.17	0.16	0.14	0.12	0.11	**0.10**	**0.10**	**0.10**	0.11	0.11
rec vs sci	0.17	0.19	0.18	0.18	0.17	0.17	**0.16**	**0.16**	**0.16**	**0.16**	**0.16**
rec vs talk	0.15	0.18	0.17	0.16	0.15	0.15	**0.14**	**0.14**	**0.14**	**0.14**	**0.14**
sci vs talk	0.20	0.22	0.21	0.20	0.19	0.19	**0.18**	**0.18**	**0.18**	0.19	0.19