Machine Learning (2022) 111:1651–1670
https://doi.org/10.1007/s10994-021-06088-2

Optimal transport for conditional domain matching and label shift

A. Rakotomamonjy[1] · R. Flamary[2] · G. Gasso[3] · M. El Alaya[4] · M. Berar[5] · N. Courty[6]

Received: 24 February 2021 / Revised: 30 July 2021 / Accepted: 27 September 2021 /
Published online: 2 November 2021

Abstract
We address the problem of unsupervised domain adaptation under the setting of generalized target shift (joint class-conditional and label shifts). For this framework, we theoretically show that, for good generalization, it is necessary to learn a latent representation in which both marginals and class-conditional distributions are aligned across domains. For this sake, we propose a learning problem that minimizes importance weighted loss in the source domain and a Wasserstein distance between weighted marginals. For a proper weighting, we provide an estimator of target label proportion by blending mixture estimation and optimal matching by optimal transport. This estimation comes with theoretical guarantees of correctness under mild assumptions. Our experimental results show that our method performs better on average than competitors across a range domain adaptation problems including *digits*,*VisDA* and *Office*. Code for this paper is available at https://github.com/arakotom/mars_domain_adaptation.

1 Introduction

Unsupervised Domain Adaptation (UDA) is a machine learning subfield that aims at addressing issues due to the discrepancy of train/test, also denoted as source/test, data distributions. There exists a large amount of literature addressing the UDA problem under different assumptions. One of the most studied setting is based on the covariate shift assumption (marginal distributions on source and target $p_S(x) \neq p_T(x)$ and conditional

Editors: Annalisa Appice, Sergio Escalera, Jose A. Gamez, Heike Trautmann.

✉ A. Rakotomamonjy
alain.rakoto@insa-rouen.fr

1 Criteo AI Lab, Paris, France

2 CMAP, Ecole Polytechnique, Palaiseau, France

3 LITIS, INSA de Rouen Normandie, Saint-Étienne-du-Rouvray, France

4 LMAC, Université Technologique de Compiègne, Compiègne, France

5 LITIS, Université de Rouen Normandie, Saint Etienne du Rouvray, France

6 IRISA, Université de Bretagne Sud, Lorient, France

distributions $p_S(y|x) = p_T(y|x)$) for which methods perform importance weighting (Sugiyama et al., 2007) or aim at aligning the marginal distributions in some learned feature space (Pan et al., 2010; Long et al., 2015; Ganin & Lempitsky, 2015). Target shift, also denoted as label shift (Schölkopf et al., 2012) assumes that for the class prior probability, $p_S(y) \neq p_T(y)$ while, for the class-conditional distributions, we have $p_S(x|y) = p_T(x|y)$. For this problem, most works seek at estimating either the ratio $p_T(y)/p_S(y)$ or the label proportions (Lipton et al., 2018; Azizzadenesheli et al., 2019; Li et al., 2019; Redko et al., 2019; Shrikumar et al., 2020;).

However as most models now learn the latent representation space, in practical situations we have both a label shift ($p_S(y) \neq p_T(y)$) and class-conditional probability shift ($p_S(z|y) \neq p_T(z|y)$, z being a vector in the latent space). For this more general DA assumption, denoted as generalized target shift, fewer works have been proposed. Zhang et al. (2013) have been among the first authors that proposed a methodology for handling both shifts. They used a kernel embedding of distributions for estimating importance weights and for transforming samples so as to match class-conditional distributions. Gong et al. (2016) follow similar idea by assuming that there exists a linear mapping that maps source class-conditionals to the target ones. For addressing the same problem (Wu et al., 2019) introduced a so-called asymmetrically-relaxed distance on distributions that allows to mitigate the effect of label shift when aligning marginal distributions. Interestingly, they also show that, when marginals in the latent space are aligned, error in the target domain is lower-bounded by the mismatch of label distributions between the two domains. Recently, Combes et al. (2020) have presented a theoretical analysis of this problem showing that target generalization can be achieved by matching label proportions and class-conditionals in both domains. The key component of their algorithm relies on a importance weight estimation of the label distributions. Unfortunately, although relevant in practice, their label distribution estimator got theoretical guarantee only when class conditionals match across domains and empirically breaks as soon as class conditionals mismatch becomes large enough.

Our work addresses UDA with generalized target shift and we make the following contributions. From a theoretical side, we introduce a bound which clarifies the role of the label shift and class-conditional shift in the target generalization error bound. Our theoretical analysis emphasizes the importance of learning with same label distributions in source and target domains while seeking at minimizing class-conditional shifts in a latent space. Based on this theory, we derive a learning problem and an algorithm which aims at minimizing Wasserstein distance between weighted marginals while ensuring low empirical error in a weighted source domain. Since a weighting scheme requires the knowledge of the label distribution in the target domain, we solve this estimation problem by blending a consistent mixture proportion estimator and an optimal matching assignment problem. While conceptually simple, our strategy is supported by theoretical guarantees of correctness. Then, given the estimated label proportion in the target domain, we theoretically show that finding a latent space in which the Wasserstein distance between the weighted source marginal distribution and the target one is zero, guarantees that class-conditionals are also matched. We illustrate in our experimental analyses how our algorithm (named MARS from Match And Reweight Strategy) copes with label and class-conditional shifts and show that it performs better than other generalized target shift competitors on several UDA problems.

2 Notation and background

Let $\mathcal{X}$ and $\mathcal{Y}$ be the input and output space. We denote by $\mathcal{Z}$ the latent space and $\mathcal{G}$ the class of representation mappings from $\mathcal{X}$ to $\mathcal{Z}$. Similarly, $\mathcal{H}$ represents the hypothesis space, which is a set of functions from $\mathcal{Z}$ to $\mathcal{Y}$. A labeling function f is a function from $\mathcal{X}$ to $\mathcal{Y}$. Elements of $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are respectively noted as x, y and z. For our UDA problem, we assume a learning problem with source and target domains and respectively note as $p_S(x, y)$ and $p_T(x, y)$ their joint distributions of features and labels. We have at our disposal a labeled source dataset $\{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ with $y_i^s \in \{1 \dots C\}$ (or $\{0, 1\}$ for binary classification) and only unlabeled examples from the target domain $\{x_i^t\}_{i=1}^{n_t}$ with all $x_i \in \mathcal{X}$, sampled *i.i.d* from their respective distributions. We refer to the marginal distributions of the source and target domains in the latent space as $p_S^g(z)$ and $p_T^g(z)$. Class-conditional probabilities in the latent space and label proportion for class j will be respectively noted as $p_U^j \triangleq p_U(z|y = j)$ and $p_U^{y=j} \triangleq p_U(y = j)$ with $U \in \{S, T\}$. Finally, we defer proofs of the theoretical results to the appendix.

2.1 Domain adaptation framework

Since the seminal work of Pan et al. (2010), Long et al. (2015), Ganin and Lempitsky (2015), a common formulation of the covariate shift domain adaptation problem is to learn a mapping of the source and target samples into a latent representation space where the distance between their marginal distributions is minimized and to learn a hypothesis that correctly predicts labels of samples in the source domain. This typically translates into the following optimization problem:

$$\min_{h,g} \frac{1}{n} \sum_{i=1}^{n_s} L(y_i^s, h(g(x_i^s))) + \lambda D(p_S^g, p_T^g) + \Omega(h, g) \tag{1}$$

where $h(\cdot)$ is the hypothesis, $g(\cdot)$ a representation mapping and $L(\cdot, \cdot) : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$ is a continuous loss function differentiable on its second parameter and Ω a regularization term. Here, $D(\cdot, \cdot)$ is a distance metric between distributions that measures discrepancy between source and target marginal distributions as mapped in a latent space induced by g. Most used distance measures are MMD (Tzeng et al., 2014), Wasserstein distance (Shen et al., 2018) or Jensen–Shannon distance (Ganin et al., 2016).

2.2 Optimal transport (OT)

We provide here some background on optimal transport as it will be a key concept for assigning label proportion. More details can be found in Peyré et al. (2019). Optimal transport measures the distance between two distributions over a space $\mathcal{X}$ given a transportation cost $c : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$. It seeks for an optimal coupling between the two measures that minimizes a transportation cost. In a discrete case, we denote the two measures as $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{i=1}^m b_i \delta_{x_i'}$. The Kantorovich relaxation of the OT problem seeks for a transportation coupling $\mathbf{P}$ that minimizes the problem

$$\min_{\mathbf{P} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle \tag{2}$$

where $\mathbf{C} \in \mathbb{R}^{n \times m}$ is the matrix of all pairwise costs, $\mathbf{C}_{i,j} = c(x_i, x'_j)$ and $\Pi(\mathbf{a}, \mathbf{b}) = \{\mathbf{P} \in \mathbb{R}_+^{n \times m} | \mathbf{P1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}\}$ is the transport polytope between the two distributions. The above problem is known as the discrete optimal transport problem and in the specific case where $n = m$ and the weights $\mathbf{a}$ and $\mathbf{b}$ are positive and uniform then the solution of the above problem is a scaled permutation matrix (Peyré et al., 2019). One of the key features of OT that we are going to exploit for solving the domain adaptation problem is its ability to find correspondences between samples in an unsupervised way by exploiting the underlying space geometry. These features have been for instance exploited for unsupervised word translation (Alvarez-Melis et al., 2019; Alaux et al., 2019).

3 Theoretical insights

In this work, we are interested in a situation where both class-conditional and label shifts occur between source and target distributions *i.e* there exists some j so that $p_S(z|y=j) \neq p_T(z|y=j)$ and $p_S^{y=j} \neq p_T^{y=j}$. Because we have these two sources of mismatch, the resulting domain adaptation problem is difficult and aligning marginals is not sufficient (Wu et al., 2019).

For better understanding the key aspects of the problem, we provide an upper bound on the target generalization error which exhibits the role of class-conditional and label distribution mismatches. For a sake of simplicity, we will consider binary classification problem. Let $\mathcal{X}$ be the input space and assume that the function $f : \mathcal{X} \mapsto \{0, 1\}$ is the domain-invariant labeling function, which is a classical assumption in DA (Shen et al., 2018; Wu et al., 2019). For a domain U, with $U = \{S, T\}$, the induced marginal probability of samples in $\mathcal{Z}$ is formally defined as $p_U^g(A) = p_U(g^{-1}(A))$ for any subset $A \subset \mathcal{Z}$ and $g^{-1}(A)$ being potentially a set ($p_U^g(A)$ is thus the push-forward of $p_U(x)$ by $g(\cdot)$). Similarly, we define the conditional distribution $g_U(\cdot|z)$ such that $p_U(x) = \int g_U(x|z)p_u^g(z)dz$ holds for all $x \in \mathcal{X}$. For a representation mapping g, an hypothesis h and the labeling function f, the expected risk is defined as $\varepsilon_U(h \circ g, f) \triangleq \mathbb{E}_{x \sim p_U}[|h(g(x)) - f(x)|] = \mathbb{E}_{z \sim p_U^z}[|h(z) - f_U^g(z)|] \triangleq \varepsilon_U^z(h, f_U^g)$ with f_U^g being a domain-dependent labeling function defined as $f_U^g(z) = \int f(x)g_U(x|z)dx$.

Now, we are in position to derive a bound on the target error but first, we introduce a key intermediate result.

Lemma 1 *Assume two marginal distributions p_S^g and p_T^g, with $p_U^g = \sum_{k=1}^C p_U^{y=k} p_U^k$, $U = \{S, T\}$. For all p_T^y, p_S^y and for any continuous class-conditional density distribution p_S^k and p_T^k such that for all z and k, we have $p_S(z|y=k) > 0$ and $p_S(y=k) > 0$, the inequality $\sup_{k,z}[w(z)S_k(z)] \geq 1$ holds with $S_k(z) = \frac{p_T(z|y=k)}{p_S(z|y=k)}$ and $w(z) = \frac{p_T^{y=k}}{p_S^{y=k}}$, if z is of class k.*

Intuitively, this lemma says that the maximum ratio between class-conditionals weighted by label proportion ratio is lower-bounded by 1, and that potentially, this bound can be achieved when both $p_S^{y=k} = p_T^{y=k}$ and $p_S^k = p_T^k$. Interestingly, Wu et al. (2019)'s results involve a similar term $\sup_z \frac{p_T^g(z)}{p_S^g(z)}$ for defining their assymetrically-relaxed distribution distance. But we use a finer modeling that allows us to explicitly disentangle the role of the class-conditionals and label distribution ratio. In our case, owing to this inequality, we can bound one of the key term that upper bounds the generalization error in the target domain.

Theorem 1 *Under the assumption of Lemma 1, and assuming that any function $h \in \mathcal{H}$ is K-Lipschitz and g is a continuous function then for every function h and g, we have*

$$\begin{aligned}\varepsilon_T(h\circ g,f) \leq \varepsilon_S(h\circ g,f) + 2K \cdot WD_1\left(p_S^g, p_T^g\right) \\ + \left[1 + \sup_{k,z} w(z)S_k(z))\right]\varepsilon_S\left(h^\star\circ g,f\right) \\ + \varepsilon_T^z\left(f_S^g,f_T^g\right)\end{aligned}$$

where $S_k(z)$ and $w(z)$ are as defined in Lemma 1, $h^\star = \arg\min_{h\in\mathcal{H}} \varepsilon_S(h\circ g;f)$ and $\varepsilon_T^z(f_S^g,f_T^g) = \mathbb{E}_{z\sim p_T^z}[|f_T^g(z) - f_S^g(z)|]$ and WD_1 defined through its dual form as

$$WD_1\left(p_S^g, p_T^g\right) = \sup_{\|v\|_L\leq 1} \mathbf{E}_{z\sim p_S^g} w(z)v(z) - \mathbf{E}_{z\sim p_T^g} v(z) \tag{3}$$

with $w(\cdot) = 1$.

Let us analyze the terms that bound the target generalization error. The first term $\varepsilon_S(h\circ g,f) \triangleq \varepsilon_S^z(h,f_S^g)$ can be understood as the error induced by the hypothesis h and the mapping g. This term is controllable through an empirical risk minimization approach as we have some supervised training data available from the source domain. The second term is the Wasserstein distance between the marginals of the source and target distribution in the latent space. Again, this can be minimized based on empirical examples and the Lipschitz constant K can be controlled either by regularizing the model $g(\cdot)$ or by properly setting the architecture of the neural network model used for $g(\cdot)$. The last term $\varepsilon_T(f_S^g,f_T^g)$ is not directly controllable (Wu et al., 2019) but it becomes zero if the latent space labelling function is domain-invariant which is a reasonable assumption especially when latent joint distributions of the source and target domains are equal. The remaining term that we have to analyze is $\sup_{k,z}[w(z)S_k(z)]$ which according to Lemma 1 is lower bounded by 1. This lower bound is attained when the label distributions are equal and class-conditional distributions are all equal and in this case, the joint distributions in the source and target domains are equal and thus $\varepsilon_T^z(f_S^g,f_T^g) = 0$.

4 Match and reweight strategy

4.1 The learning problem

The bound in Theorem 1 suggests that a good model should: (i) look for a latent representation mapping g and a hypothesis h that generalizes well on the source domain, ii) have minimal Wasserstein distance between marginal distributions of the latent representations while having class-conditional probabilities that match, and iii) learn from source data with equal label proportions as the target so as to have $w(z) = 1$ for all z. For yielding our learning problem, we will translate these properties into an optimization problem.

At first, let us note that one simple and efficient way to handle mismatch in label distribution is to consider importance weigthing in the source domain. Hence, instead of learning from the marginal source distribution $p_S = \sum_{k=1}^{C} p_S^{y=k} p_S^k$, we learn from a reweighted version denoted as $p_{\tilde{S}} = \sum_{k=1}^{C} p_T^{y=k} p_S^k$, as proposed by Sugiyama et al. (2007), Combes et al. (2020), so that no label shift occurs between $p_{\tilde{S}}$ and p_T . This approach needs an estimation of $p_T^{y=k}$ that we will detail in the next subsection, but interestingly, in this case, for Theorem 1, we will have $w(z) = \frac{p_T^{y=k}}{p_{\tilde{S}}^{y=k}} = \frac{p_T^{y=k}}{p_T^{y=k}} = 1$. Then, based on the bound in Theorem 1 applied to $p_{\tilde{S}}$ and p_T, we propose to learn the functions h and g by solving the problem

$$\min_{g,h} \frac{1}{n} \sum_{i=1}^{n_s} w^\dagger\left(x_i^s\right) L\left(y_i^s, h\left(g\left(x_i^s\right)\right)\right) + \lambda WD_1\left(p_{\tilde{S}}^g, p_T^g\right) + \Omega(h,g) \tag{4}$$

where the importance weight $w^\dagger(x_i^s) = \frac{p_T^{y=y_i}}{p_S^{y=y_i}}$ allows to simulate sampling from $p_{\tilde{S}}^g$ given p_S^g, and the discrepancy between marginals is the Wasserstein distance

$$WD_1\left(\tilde{p}_s^g, p_t^g\right) = \sup_{\|v\|_L \le 1} \mathbf{E}_{z\sim p_S^g} w^\dagger(z)v(z) - \mathbf{E}_{z\sim p_T^g} v(z). \tag{5}$$

The first term of Eq. (4) corresponds to the empirical loss related to the error $\varepsilon_{\tilde{S}}$ in Theorem 1 while the distribution divergence aims at minimizing distance between marginal probabilities, the second term in that theorem. In the next subsections, we will make clear why the Wasserstein distance is used as the divergence and provide conditions and guarantees for having $WD_1(\tilde{p}_S^g, p_T^g) = 0 \implies WD(p_S^k, p_T^k) = 0$, i.e. perfect class-conditionals matching, and thus $S_k(z) = 1$ for all k, z. Recall that in this case, the lower bound on $\max_{k,z}[w(z)S_k(z)]$ will be attained.

Algorithmically, for solving the problem in Eq. (4), we employ a classical adversarial learning strategy. It is based on a standard back-propagation strategy using stochastic gradient descent (detailed in Algorithm 1). We estimate the label proportion using Algorithm 2 and then use this proportion for computing the importance weights $w(\cdot)$. The first part of the algorithm consists then in computing the weighted Wassertein distance using gradient penalty (Gulrajani et al., 2017). Once this distance is computed, we back-propagate the error through the parameters of the feature extractor g and the classifier f. In practice, we use weight decay as regularizer Ω over the representation mapping and classifier functions g and h.

Algorithm 1 Training the full MARS model

Require: $\{(x_i^s, y_i^s)\}, \{x_i^t\}$, number of classes C, batch size B, number of critic iterations n

Initialize representation mapping g, the classifier h and the domain critic $v(\cdot)$, with parameters θ_h, θ_g, θ_v

repeat

  estimate $\mathbf{p}_T$ from $\{x_i^t\}$ using Algorithm 2 {done every 10 iterations}

  sample minibatches $\{(x_B^s, y_B^s)\}, \{x_B^t\}$ from $\{(x_i^s, y_i^s)\}$ and $\{x_i^t\}$

  compute $\{w_i^\dagger\}_{i=1}^C$ given $\{(x_B^s, y_B^s)\}$ and $\mathbf{p}_T$

  for $t = 1, \cdots, n$ **do**

    $z^s \leftarrow g(x_B^s)$, $z^t \leftarrow g(x_B^t)$

    compute gradient penalty $\mathcal{L}_{\text{grad}}$

    compute empirical Wasserstein dual loss $\mathcal{L}_{wd} = \sum_i w^\dagger(z_i^s) v(z_i^s) - \frac{1}{B}\sum_i v(z_i^t)$

    $\theta_v \leftarrow \theta_v + \alpha_v \nabla_{\theta_v}[\mathcal{L}_{wd} - \mathcal{L}_{\text{grad}}]$

  end for

  compute the weighted classification loss $\mathcal{L}_w = \sum_i w^\dagger(z_i^s) L(y_i^s, h(g(x_i^s)))$

  $\theta_h \leftarrow \theta_h - \alpha_h \nabla_{\theta_h}\mathcal{L}_w$

  $\theta_g \leftarrow \theta_g - \alpha_g \nabla_{\theta_g}[\mathcal{L}_w + \lambda \mathcal{L}_{wd}]$

until a convergence condition is met

4.2 Estimating target label proportion using optimal assignment

The above learning problem needs an estimation of $P_T(y)$ for weighting the classification loss and for computing the Wasserstein distance between p_S^g and p_T^g. Several approaches exist for estimating p_T^y when class-conditional distributions in source and target matches (Redko et al., 2019; Combes et al., 2020). However, this is not the case in our general setting. Hence, in order to make the problem tractable, we will introduce some assumptions on the structure and geometry of the class-conditional distributions in the target domain that allow us to provide guarantee on the correct estimation of p_T^y.

For achieving this goal, we first consider the target marginal distribution as a mixture of models and estimate the proportions of the mixture. Next we aim at finding a permutation $\sigma(\cdot)$ that guarantees, under mild assumptions, correspondence between the class-conditional probabilities of same class in the source and target domain. Then, this permutation allows us to correctly assign a class to each mixture proportion leading to a proper estimation of each class label proportion in the target domain.

In practice, for the first step, we assume that the target distribution is a mixture model with C components $\{p_T^j\}$ and we want to estimate the mixture proportion of each component. For this purpose, we have considered two alternative strategies coming from the literature : i) applying agglomerative clustering on the target samples tells us about the membership class of each sample and thus, the resulting clustering provides the proportion of each component in the mixture. ii) learning a Gaussian mixture model over the data in the target domain. This gives us both the estimate components $\{p_T^j\}$ and the proportion of the mixture $\mathbf{p}_u$. Under some conditions on its initialization and assuming the model is well-calibrated, Zhao et al. (2020) have shown that the sample estimator asymptotically converges towards the true mixture model.

Algorithm 2 Label proportion estimation in the target domain

Require: $\{(x_i^s, y_i^s)\}, \{x_i^t\}$, number of classes C

Ensure: $\mathbf{p}_T$: Estimated label proportion

1: $\{p_T^j\}, \mathbf{p}_u \leftarrow$ Estimate a mixture with C modes and related proportions from $\{x_i^t\}$.

2: $\mathbf{D} \leftarrow$ Given $\mathcal{D}$, compute the matrix pairwise distance $\{p_S^i\}$ and $\{p_T^j\}$ modes.

3: $\mathbf{P}^\star \leftarrow$ Solve OT problem (2) with $\mathbf{D}$ and uniform marginals as in Proposition 1.

4: $\mathbf{p}_T \leftarrow C \cdot \mathbf{P}^\star \mathbf{p}_u$ Permute the mixture proportion on source ($C \cdot \mathbf{P}^\star$ is a permutation matrix)

Matching class-conditionals with OT Since, we do not know to which class each component of the mixture in target domain is related to, we assume that the conditional distribution in the source and target domain of the same class can be matched owing to optimal assignment. The resulting permutation would then help us assign each label proportion estimated as above to the correct class-conditional. Figure 1 in the appendix illustrates this matching problem.

Let us suppose that we have an estimation of all C class-conditional probabilities on source and target domain (based on empirical distributions). We want to solve an optimal assignment problem with respect to the class-conditional probabilities $\{p_S^i\}_{i=1}^C$ and $\{p_T^j\}_{j=1}^C$ and we clarify under which conditions on distance between class-conditional probabilities, the assignment problem solution achieves a correct matching of classes (*i.e* p_S^i is correctly assigned to p_T^i for all i). Formally, denote as $\mathbb{P}$ the set of probability distributions over $\mathbb{R}^d$ and assume a metric over $\mathbb{P}$. We want to optimally assign a finite number C of probability distributions of $\mathbb{P}$ to another set of finite number C of probability distributions belonging to $\mathbb{P}$, in a minimizing distance sense. Based on a distance $\mathcal{D}$ between couple of class-conditional probability distributions, the assignment problems looks for the permutation that solves $\min_\sigma \frac{1}{C} \sum_j \mathcal{D}(p_S^j, p_T^{\sigma(j)})$. Note that the best permutation $\sigma^\star$ solution to this problem can be retrieved by solving a Kantorovitch relaxed version of the optimal transport (Peyré et al., 2019) with marginals $\mathbf{a} = \mathbf{b} = \frac{1}{C}\mathbb{1}$. Hence, this OT-based formulation of the matching problem can be interpreted as an optimal transport one between discrete measures of probability distributions of the form $\frac{1}{C}\sum_{j=1}^C \delta_{p_U^j}$. In order to be able to correctly match class-conditional probabilities in source and target domain by optimal assignement, we ask ourselves:

Under which conditions the retrieved permutation matrix would correctly match the class-conditionals?

In other word, we are looking for conditions of identifiability of classes in the target domain based on their geometry with respect to the classes in source domain. Our proposition below presents an abstract sufficient condition for identifiability based on the notion of cyclical monotonicity and then we exhibit some practical situations in which this property holds.

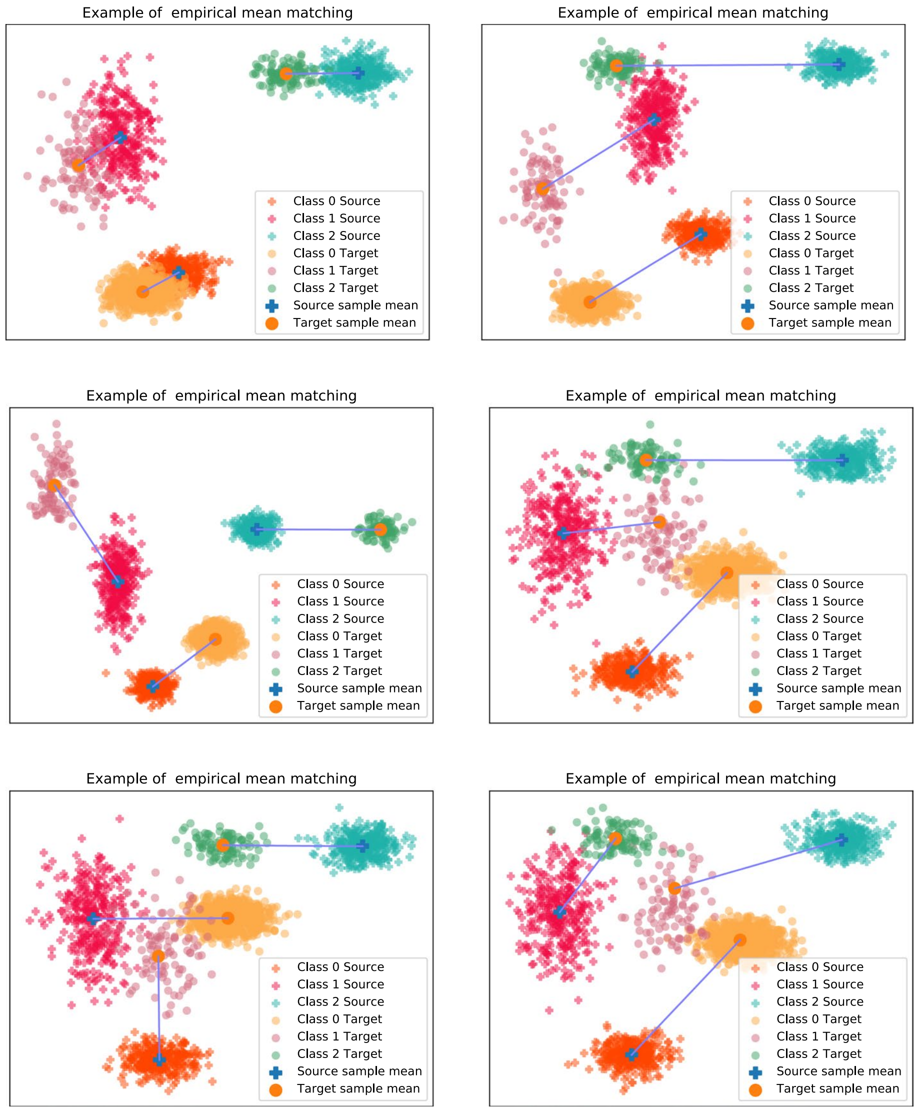


Fig. 1 Example of geometrical arrangments of the source and target class-conditional distributions that allows correct and incorrect matching of classes by optimal transport of empirical means (assuming correct estimation of these means). Blue lines denote the matching. (top-left) In this setting, the displacements of each class-conditionals is so that for each class i $\|\mathbf{m}_S^i - \mathbf{m}_T^i\|_2 \leq \|\mathbf{m}_S^i - \mathbf{m}_T^j\|_2$, for all j. We are thus in the first example that we gave as satisfying Proposition 1. (top-right) Class-conditionals have been displaced such that the "nearness" hypothesis is not respected anymore. However, target class-conditional distributions are obtained by a linear Monge map of their source counterparts. This ensures that optimal transport allows their matchings (based on their means). (middle) We have illustrated two other examples of distribution arrangments that allow class matching. (bottom) Two examples that break our assumption. In both cases, one target class-conditional is "near" another source class, without the global displacements of all target class-conditionals being uniform in direction (Color figure online)

Proposition 1 *Denote as* $\nu = \frac{1}{C}\sum_{j=1}^{C}\delta_{p_S^j}$ *and* $\mu = \frac{1}{C}\sum_{j=1}^{C}\delta_{p_T^j}$*, representing respectively the balanced weighted sum of class-conditionals probabilities in source and target domains. Given* $\mathcal{D}$ *a distance over probability distributions, assume that for any permutation* σ *of* C *elements, the following assumption, known as the* $\mathcal{D}$*-cyclical monotonicity relation, holds*

$$\sum_j \mathcal{D}\Big(p_S^j, p_T^j\Big) \leq \sum_j \mathcal{D}\Big(p_S^j, p_T^{\sigma(j)}\Big)$$

then solving the optimal transport problem between ν *and* μ *as defined in Eq. (2) using* $\mathcal{D}$ *as the ground cost matches correctly class-conditional probabilities.*

While the cyclical monotonicity assumption above can be hard to grasp, there exists several situations where it applies. One condition that is simple and intuitive is when class-conditionals of same source and target classes are "near" each other in the latent space. More formally, if we assume that $\forall j\ \mathcal{D}(p_S^j, p_T^j) \leq \mathcal{D}(p_S^j, p_T^k) \quad \forall k$, then summing over all possible j, and choosing k so that all the couples of (j, k) form a permutation, we recover the cyclical monotonicity condition $\sum_j^C \mathcal{D}(p_S^j, p_T^j) \leq \sum_j^{\mathcal{D}}(p_S^j, p_T^{\sigma(j)}), \forall\sigma$. Another more general condition on the identifiability of the target class-conditional can be retrieved by exploiting the fact that, for discrete optimal transport with uniform marginals, the support of optimal transport plan satisfies the cyclical monotonicity condition (Santambrogio, 2015). This is for instance the case, when p_S^j and p_T^j are Gaussian distributions of same covariance matrices and the mean m_T^j of each p_T^j is obtained as a linear symmetric positive definite mapping of the mean m_S^j of p_S^j and the distance $\mathcal{D}(p_S^j, p_T^j)$ is $\|m_S^j - m_T^j\|_2^2$ (Courty et al., 2016). This situation would correspond to a linear shift of the class-conditionals of the source domain to get the target ones. Figure 1 illustrates how our class-conditional matching algorithm performs on a simple toy problem. While our assumptions can be considered as strong, we illustrate in Fig. 4, that the above hypotheses hold for the VisDA problem, and lead afterwards to a correct matching of the class-conditionals.

It is interesting to compare our assumptions on identifiability to other hypotheses proposed in the literature for solving (generalized) target shift problems. When handling only target shift, one common hypothesis (Redko et al., 2019) is that class-conditional probabilities are equal. This in our case boils down to have a 0 distance between $\mathcal{D}(P_S^j, P_T^j)$ guaranteeing matching under our more general assumptions. When both shifts occur on labels and class-conditionals, Wu et al. (2019) assume that there exists continuity of support between the $p(z|y)$ in source and target domains. Again, this assumption may be related to the above minimum distance hypothesis if class-conditionals in source domain are far enough. Interestingly, one of the hypothesis of Zhang et al. (2013) for handling generalized target shift is that there exists a linear transformation between the class-conditional probabilities in source and target domains. This is a particular case of our Proposition 1 and subsequent discussion, where the mapping between class-conditionals is supposed to be linear. Our conditions for correct matching and thus for identifying classes in the target domain are more general than those proposed in the current literature.

4.3 When matching marginals lead to matched class-conditionals?

In our learning problem, since one term we aim at minimizing is $WD_1(p_{\tilde{S}}^g, p_T^g)$, with $p_{\tilde{S}}^g = \sum_j p_T^{y=j} p_S^j$ and $p_T^g = \sum_j p_T^{y=j} p_T^j$, we want to understand under which assumptions $WD_1(p_{\tilde{S}}^g, p_T^g) = 0$ implies that $p_S(z|y=j) = p_T(z|y=j)$ for all j, which is key for a good generalization as stated in Theorem 1. Interestingly, the assumptions needed for guaranteeing this implication are the same as those in Proposition 1.

Proposition 2 *Denote as* γ *the optimal coupling plan for distributions* ν *and* μ *defined as balanced weighted sum of class-conditionals that is* $\nu = \frac{1}{C}\sum_{j=1}^{C}\delta_{p_S^j}$ *and* $\mu = \frac{1}{C}\sum_{j=1}^{C}\delta_{p_T^j}$ *under assumptions given in Proposition 1. Assume that the classes are ordered so that we have* $\gamma = \frac{1}{C}\mathrm{diag}(\mathbb{1})$. *Then* $\gamma' = \mathrm{diag}(\mathbf{a})$ *is also optimal for the transportation problem with marginals* $\nu' = \sum_{j=1}^{C} a_j\delta_{p_S^j}$ *and* $\mu' = \sum_{j=1}^{C} a_j\delta_{p_T^j}$, *with* $a_j > 0, \forall j$. *In addition, if the Wasserstein distance between* ν' *and* μ' *is* 0, *it implies that the distance between class-conditionals are all* 0.

Applying this proposition with $a_j = p_T^{y=j}$ brings us the guarantee that under some geometrical assumptions on the class-conditionals in the latent space, having $WD_1(\tilde{p}_S^g, p_T^g) = 0$ implies matching of the class-conditionals, resulting in a minimization of $\max_{k,z} w(z)S_k(z)$ (remind that $w(z) = 1$ as mixture components p_S^j and p_T^j of $p_{\tilde{S}}^g$ and p_T^g are both weighted by $p_T^{y=j}$ for all j, since we learn using $p_{\tilde{S}}^g$).

5 Discussions

From a theoretical point of view, several works have pointed out the limitations of learning domain invariant representations. Johansson et al. (2019), Zhao et al. (2019) and Wu et al. (2019) have introduced some generalization bounds on the target error that show the key role of label distribution and conditional distribution shifts when learning invariant representations. Importantly, Zhao et al. (2019) and Wu et al. (2019) have shown that in a label shift situation, minimizing source error while achieving invariant representation will tend to increase the target error. In our work, we introduce an upper bound that clarifies the importance of learning invariant representations that also align class-conditional representations in source and target domains.

Algorithmically, most related works are the one by Wu et al. (2019) and Combes et al. (2020) that also address generalized target shift. The first approach does not seek at estimating label proportion but instead allows flexibility in the alignment by using an assymetrically-relaxed distance. In the case of Wasserstein distance, the approach of Wu et al. (2019) consists in reweighting the marginal of the source distribution and in its dual form, their distance boils to

$$WD_w(p_S, p_T) = \sup_{\|v\|_L \leq 1} \mathbf{E}_{x\sim p_S} w(x)v(x) - \mathbf{E}_{x\sim p_T} v(x)$$

where $w(\cdot)$ is actually a constant $\frac{1}{1+\beta}$. We can note that the adversarial loss we propose is a general case of this one. Indeed, in the above, the same amount of weighting applies to all the samples of the source distribution. At the contrary, our reweighting scheme depends on the class-conditional probability and their estimate target label proportion. Hence, we believe that our approach would adapt better to imbalance without the need to tune β (by validation for instance, which is hard in unsupervised domain adaptation). The work of Combes et al. (2020) and our differs only in the way the weights $w(x)$ are estimated. In our case, we consider a theoretically supported and consistent estimation of the target label proportion, while they directly estimate $w(\cdot)$ by applying a technique tailored and grounded for problems without class-conditional shifts. We will show in the experimental section that their estimator in some cases lead to poor generalization.

Still in the context of reweighting, Yan et al. (2017) proposed a weighted Maximum Mean discrepancy distance for handling target shift in UDA. However, their weights are estimated based on pseudo-labels obtained from the learned classifier and thus, it is difficult to understand whether they provide accurate estimation of label proportion even in simple setting. While their distance is MMD-transposed version of our weighted Wasserstein, our approach applies to representation learning and is more theoretically grounded as the label proportion estimation is based on sound algorithm with proven convergence guarantees (see below) and our optimal assignment assumption provides guarantees on situations under which class-conditional probability matching is correct.

The idea of matching moment of distributions have already been proven to be an effective for handling distribution mismatch. About ten years ago, Huang et al. (2007), Gretton et al. (2009), Yu and Szepesvári (2012) already leveraged such an idea for handling covariate shift by matching means of distributions in some reproducing kernel Hilbert space. Li et al. (2019) recycled the same idea for label proportion estimation and extended the idea to distribution matching. Interestingly, our approach differs on its usage. While most above works employ mean matching for density ratio estimation or for label proportion estimation, we use it as a mean for identifying displacement of class-conditional distributions through optimal assignment/transport. Hence, it allows us to assign estimated label proportion to the appropriate class.

For estimating the label proportion, we have proposed to learn a Gaussian mixture model of the target distribution. By doing so we are actually trying to solve a harder problem than necessary. However, once the target distribution estimation has been evaluated and class-conditional probabilities being assigned from the source class, one can use that Gaussian mixture model for labelling the target samples. Note however that Gaussian mixture learned by expectation-minimization can be hard to estimate especially in high-dimension (Zhao et al., 2020) and that the speed of convergence of the EM algorithm depends on smallest mixture weights (Naim & Gildea, 2012). Hence, in high-dimension and/or highly imbalanced situations, one may get a poor estimate of the target distribution. Nonetheless, one can consider other non-EM approach (Kannan et al., 2005; Arora et al., 2005). Hence, in practice, we can expect the approach GMM estimation and OT-based matching to be a strong baseline in low-dimension and well-clustered mixtures setting but to break in high-dimension one.

6 Numerical experiments

We present in this section some experimental analyses of the proposed algorithm on a toy dataset as well as on real-world visual domain adaptation problems. The code for reproducing part of the experiments is available at https://github.com/arakotom/mars_domain_adaptation.

6.1 Experimental setup

Our goal is to show that among algorithms tailored for handling generalized target shift, our method is the best performing one (on average). Hence, we compare with two very recent methods designed for generalized target shift and with two domain adaptation algorithms tailored for covariate shift for sanity check.

As a baseline, we consider a model, denoted as Source trained for f and g on the source examples and tested without adaptation on the target examples. Two other competitors use respectively an adversarial domain learning (Ganin et al., 2016) and the Wasserstein distance (Shen et al., 2018) computed in the dual as distances for measuring discrepancy between p_S and p_T, denoted as DANN and $\mathrm{WD}_{\beta=0}$. We consider the model proposed by Wu et al. (2019) and Combes et al. (2020) as competing algorithms able to cope with generalized target shift. For this former approach, we use the asymmetrically-relaxed Wasserstein distance so as to make it similar to our approach and also report results for different values of the relaxation β. This model is named WD_β with $\beta \geq 1$. The Combes et al. (2020)'s method, named IW-WD (for importance weighted Wasserstein distance) solves the same learning problem as ours and differs only on the way the ratio $w(x)$ is estimated. Our approaches are denoted as MARSc or MARSg respectively when estimating proportion by hierarchical clustering or by Gaussian mixtures. All methods differ only in the metric used for computing the distance between marginal distributions and most of them except DANN use a Wassertein distance. The difference essentially relies on the reweighting strategy of the source samples. For all models, learning rate and the hyperparameter λ in Eq. (4) have been chosen based on a reverse cross-validation strategy. The metric that we have used for comparison is the balanced accuracy (the average recall obtained on each class) which is better suited for imbalanced problems (Brodersen et al., 2010). All presented results have been obtained as averages over 20 runs.

6.2 Toy dataset

The toy dataset is a 3-class problem in which class-conditional probabilities are Gaussian distributions. For the source distribution, we fix the mean and the covariance matrix of each of the three Gaussians and for the target, we simply shift the means (by a fixed translation). We have carried out two sets of experiments where we have fixed the shift and modified the label proportion imbalance and another one with fixed imbalance and increasing shift. For space reasons, we have deported to the supplementary the results of the latter. Figure 2 show how models perform for varying imbalance and fixed shift. The plots nicely show what we expect. DANN performs worse as the imbalance increases. WD_β works well for all balancing but its parameter β needs to increase with the imbalance level. Because of the shift in class-conditional probabilities, IW-WD is not able to properly estimate the importance weights and fails. Our approaches are adaptive to the imbalance and perform very well over a large range for both a low-noise and mid-noise setting (examples of how the Gaussians are mixed are provided in the supplementary material). For the hardest

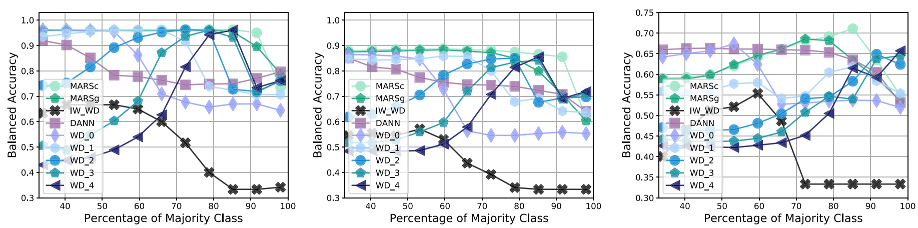


Fig. 2 Performance of the compared algorithms for three different covariance matrices of the Gaussians composing the toy dataset with respect to the imbalance. The x-axis is given with respect to the percentage of majority class which is the class 1. (left) Low-error setting. (middle) mid-error setting. (right) high-error setting. Example of the source and target samples for the different cases are provided in the supplementary material

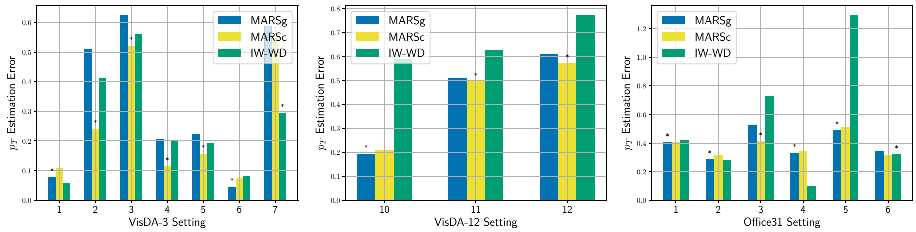


Fig. 3 Examples of ℓ_1 norm error of estimated label proportion. We have reported the performance of our two methods (MARSg and MARSc) as well as the performance of IW-WD. The three panels are related to the (left) VisDA-3, (middle) VisDA-12, (right) Office 31 and the different experimental imbalance settings (see Table 1). We have also reported, with a '*' on top, among the three approaches, the best performing one in term of balanced accuracy. We note that MARSc provides better estimation than IW-WD on 12 out of 16 experiments. Note also the correlation between better $\mathbf{p}_T$ estimation and accuracy

Fig. 4 *t-sne* embeddings of the target sample for the VisDA-3 problem and imbalance setting 2 ($\mathbf{p}_S = [0.4, 0.2, 0.4]$ and $\mathbf{p}_T = [0.2, 0.6, 0.2]$). The columns depict the embeddings obtained (left) after training on the source data without adaptation for about 10 iterations, which is sufficient for 0 training error. (right) after adaptation by minimizing the appropriate discrepancy loss between marginal distributions. From top to bottom, we have : (first-row) DANN, (second-row) $\mathrm{WD}_{\beta=1}$, (third-row), IW-WD (last row) MARSc. From the right column, we note how DANN and $\mathrm{WD}_{\beta=1}$ struggle in aligning the class conditionals, especially those of Class 1, which is the class that varies the most in term of label proportion. IW-WD manages to aligns the classes "0" and "2" but is not able to correctly match the class "1". Instead, our MARSc approach achieves high performance and correctly aligns the class conditionals, although some few examples seem to be mis-classified. Importantly, we can remark from the left column that for this example, before alignment, the embeddings seem to satisfy our Proposition 1 hypothesis. At the contrary, the assumption needed for correctly estimating $\mathbf{p}_T$ for IW-WD is not satisfied, justifying thus the good and poor performance of those models

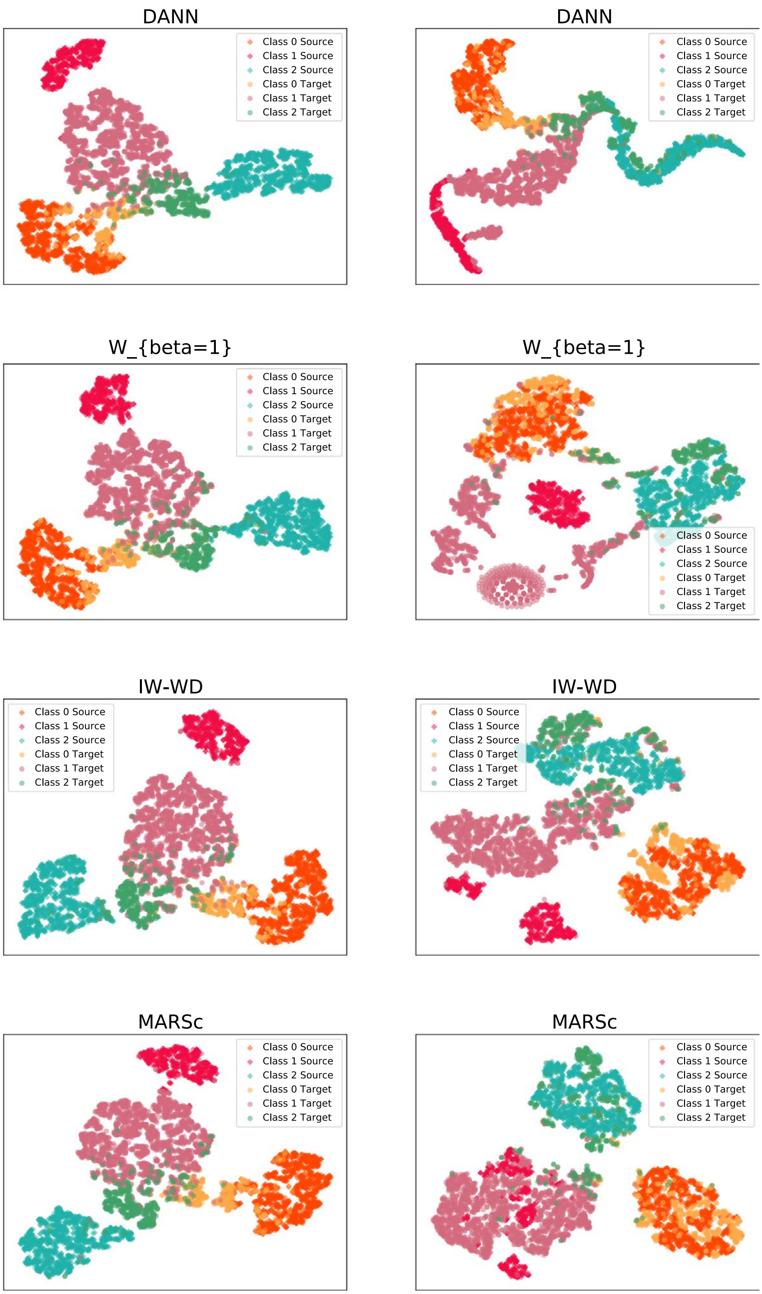
problem (most-right panel), all models have difficulties and achieve only a balanced accuracy of 0.67 over some range of imbalance. Note that for this low-dimension toy problem, as expected, the approach GMM and OT-based matching achieves the best performance as reported in the supplementary material.

6.3 Digits, VisDA and office

We present some UDA experiments on computer vision datasets (Peng et al., 2017; Venkateswara et al., 2017), with different imbalanced settings. Details of problem configurations as well as model architecture and training procedure can be found in the appendix.

Our first result provides an illustration in Fig. 4 of the latent representation we obtain for the VisDA problem after training on the source domain only and after convergence of the different DA algorithms. We first note that for this problem, the assumptions for correct matching seem to hold and this leads to very good visual matching of class-conditionals for MARS.

Table 1 reports the averaged balanced accuracy achieved by the different models for only a fairly chosen subset of problems. The full table is in the supplementary. Results presented here are not comparable to results available in the literature as they mostly consider covariate shift DA (hence with balanced proportions). For these subsets of problems, our approaches yield the best average ranking. They perform better than competitors except on the MNIST-MNISTM problems where the change in distribution might violate our assumptions. Figure 3 presents some quantitative results label proportion estimation in the



target domain between our method and IW-WD. We show that MARSc provides better estimation than this competitor 12 out of 16 experiments. As the key issue in generalized target shift problem is the ability to estimate accurately the importance weight or the target label proportion, we believe that the learnt latent representation fairly satisfies our OT hypothesis leading to good performance.

Table 1 Table of averaged **balanced accuracy** for the compared models and different domain adaptation problems and label proportion imbalance settings

Setting	Source	DANN	$WD_{\beta=0}$	$WD_{\beta=1}$	$WD_{\beta=2}$	$WD_{\beta=3}$	$WD_{\beta=4}$	IW-WD	MARSg	MARSc
MNIST-USPS 10 modes										
Balanced	76.9±3.7	79.7±3.5	93.7±0.7	74.3±4.3	51.3±4.0	76.6±3.3	71.9±5.7	95.3±0.4	**95.6±0.7**	**95.6±1.0**
Mid	80.4±3.1	78.7±3.0	94.3±0.7	75.4±3.4	55.6±4.3	79.0±3.1	72.3±4.2	**95.6±0.5**	89.7±2.3	90.4±2.6
High	78.1±4.9	81.8±4.0	**93.9±1.1**	87.4±1.7	83.8±5.2	85.7±2.5	83.6±3.0	**94.1±1.0**	88.3±1.5	89.7±2.3
USPS-MNIST 10 modes										
Balanced	77.0±2.6	80.5±2.2	73.4±2.8	66.7±2.9	49.9±2.8	55.8±2.9	52.1±3.5	80.5±2.2	**84.6±1.7**	**85.5±2.1**
Mid	**79.5±2.8**	**78.9±1.8**	75.8±1.6	63.3±2.3	53.2±2.8	47.2±2.4	48.3±2.9	**78.4±3.5**	**79.7±3.6**	**78.5±2.5**
High	**78.5±2.4**	**77.8±2.0**	**76.1±2.7**	63.0±3.3	57.6±4.8	51.2±4.4	49.3±3.3	71.5±4.7	75.6±1.8	**77.1±2.4**
MNIST-MNISTM 10 modes										
Setting 1	58.3±1.3	**61.2±1.1**	57.4±1.7	50.2±4.4	47.0±2.0	57.9±1.1	60.0±1.3	**63.1±3.1**	58.1±2.3	56.6±4.6
Setting 2	60.0±1.1	61.1±1.0	58.1±1.4	53.4±3.5	48.6±2.4	59.7±0.7	58.1±0.8	65.0±3.5	57.7±2.3	55.7±2.1
Setting 3	58.1±1.2	**60.4±1.4**	57.7±1.2	47.7±4.9	42.2±7.3	57.1±1.0	53.5±1.1	52.5±14.8	53.7±7.2	53.7±3.3
VisdDA 3 modes										
setting 1	79.3±4.3	78.9±9.1	91.8±0.7	73.8±2.0	61.7±2.2	65.6±2.7	58.6±2.6	**94.1±0.6**	92.5±1.2	92.1±1.8
setting 4	80.2±5.3	75.5±9.3	72.8±1.2	86.9±7.5	86.8±1.2	80.2±6.9	75.7±2.0	85.9±5.7	87.7±3.0	**91.3±4.8**
setting 2	81.5±3.5	68.5±14.7	68.8±1.3	84.5±1.2	**93.2±0.4**	73.7±14.2	60.7±0.9	78.7±10.8	84.0±4.3	**91.8±3.4**
setting 3	78.4±3.2	59.0±15.9	64.1±1.9	79.2±0.8	77.1±10.3	90.0±0.5	**94.4±0.3**	78.0±9.3	75.7±4.1	73.9±13.2
setting 5	83.5±3.5	80.9±14.5	63.9±0.6	73.7±7.3	50.9±1.1	76.5±6.7	59.3±1.0	**90.4±3.6**	**89.0±0.9**	**89.0±3.5**
setting 6	80.9±4.2	54.8±19.8	45.3±2.4	63.7±5.1	67.1±6.1	42.9±11	62.2±1.4	**94.4±1.0**	**93.7±0.4**	**93.9±1.0**
setting 7	79.2±3.7	42.9±2.5	57.5±1.5	55.4±2.0	50.2±4.3	43.7±8.3	62.5±0.8	**88.5±4.9**	78.6±3.2	**82.3±7.5**
VisdDA 12 modes										
setting 1	41.9±1.5	52.8±2.1	45.8±4.3	44.2±3.0	35.5±4.6	41.0±3.0	37.6±3.4	50.4±2.3	53.3±0.9	**55.1±1.6**
setting 2	41.8±1.5	50.8±1.6	45.7±8.9	40.5±4.8	36.2±5.0	36.1±4.6	31.9±5.7	48.6±1.8	53.1±1.6	**55.3±1.6**
setting 3	40.6±4.3	49.2±1.3	47.1±1.6	42.1±3.0	36.3±4.4	37.3±3.5	35.0±5.4	46.6±1.3	50.8±1.6	**52.1±1.2**
Office 31										
A - D	73.7±1.4	74.3±1.8	**77.2±0.7**	65.1±2.0	62.7±2.6	71.5±1.2	63.9±1.1	75.7±1.6	76.1±0.9	**78.2±1.3**
D - W	83.7±1.1	81.9±1.5	82.6±0.6	83.5±0.8	82.8±0.7	80.1±0.5	**87.1±0.9**	78.9±1.5	**86.3±0.6**	86.2±0.8

Table 1 (continued)

Setting	Source	DANN	$WD_{\beta=0}$	$WD_{\beta=1}$	$WD_{\beta=2}$	$WD_{\beta=3}$	$WD_{\beta=4}$	IW-WD	MARSg	MARSc
W - A	54.1±0.9	52.2±1.0	48.9±0.4	56.8±0.4	53.0±0.5	58.8±0.4	54.9±0.5	52.2±0.7	**60.7±0.8**	55.2±0.8
W - D	92.8±0.9	87.8±1.4	95.1±0.3	93.1±0.5	87.6±0.9	94.7±0.6	91.2±0.6	**97.0±0.9**	95.1±0.8	93.8±0.6
D - A	52.5±0.9	48.1±1.2	49.8±0.4	48.8±0.5	50.1±0.4	50.3±0.7	50.8±0.5	41.4±1.8	**54.7±0.9**	**55.0±0.9**
A - W	67.5±1.5	70.2±1.0	67.1±0.6	60.6±2.1	52.9±1.4	64.0±1.3	59.7±0.8	68.8±1.6	**73.1±1.5**	**71.9±1.2**
Office Home										
Art - Clip	37.7±0.7	36.8±0.6	33.4±1.2	31.4±1.6	27.1±1.6	31.6±5.2	29.3±6.6	37.7±0.6	37.6±0.5	**38.65±0.5**
Art - Product	49.7±0.9	50.0±0.9	39.4±3.6	38.8±2.3	35.1±2.3	35.1±3.4	32.9±3.6	49.0±0.3	**55.3±0.7**	52.2±0.4
Art - Real	58.2±1.0	53.7±0.5	51.1±2.3	50.4±1.8	46.4±2.4	51.5±4.5	45.3±11.0	57.7±0.7	**63.88±0.5**	58.8±0.7
Clip - Art	35.3±1.4	35.7±1.5	28.9±2.9	23.1±2.0	18.4±1.5	22.0±3.1	20.4±2.3	28.7±1.2	**41.2±0.6**	**40.7±0.8**
Clip - Product	**51.9±1.3**	**52.1±0.8**	39.2±7.9	39.3±2.6	34.7±1.9	39.6±2.8	39.5±2.9	34.5±2.1	51.7±0.5	**52.1±0.5**
Clip - Real	50.7±1.2	51.4±1.0	43.2±2.2	40.1±2.1	32.7±1.4	39.2±2.4	35.8±2.8	35.7±1.1	54.0±0.3	**56.6±0.5**
Product - Art	**39.6±1.6**	39.5±1.5	39.2±1.0	36.1±1.0	38.8±1.1	39.5±0.6	38.2±0.6	34.0±1.4	37.8±1.1	**39.3±1.3**
Product - Clip	32.7±0.9	**37.2±1.0**	33.8±0.5	28.4±0.7	28.4±0.6	29.7±0.5	31.8±0.8	24.9±1.0	30.9±0.8	29.3±0.9
Product - Real	**62.1±1.3**	**62.5±1.2**	**62.6±0.7**	58.1±0.5	57.6±0.6	59.3±0.6	57.1±0.8	59.2±0.9	60.5±0.6	**62.2±0.7**
Real - Product	68.3±1.0	**70.4±0.8**	**70.2±0.5**	61.7±0.8	63.4±0.9	61.5±1.0	65.5±0.6	64.5±1.5	64.8±3.6	66.5±1.1
Real - Art	**40.3±0.9**	**41.3±1.0**	39.2±0.7	33.5±1.3	31.6±1.5	36.9±0.9	36.1±0.9	36.9±1.9	39.9±1.4	39.2±1.6
Real - Clip	**42.7±1.1**	40.9±1.0	40.4±0.5	35.6±0.8	34.9±0.9	40.4±0.5	35.6±0.8	35.6±2.0	38.7±2.1	38.8±2.5
#Wins (/34)	7	9	5	0	1	0	2	9	12	21
Aver. Rank	4.16	4.73	5.32	6.97	8.38	6.59	7.57	4.95	3.38	2.95

Reported in bold are the best performances as well as other methods which achieve performance that are statistically similar according to a Wilcoxon signrank test with $p = 0.01$. Last lines present the summary of 34 experiments. #Win includes the statistical ties

7 Conclusion

The paper proposed a strategy for handling generalized target shift in domain adaptation. It builds upon the simple idea that if the target label proportion where known, then reweighting class-conditional probabilities in the source domain is sufficient for designing a distribution discrepancy that takes into account those shifts. In practice, our algorithm estimates the label proportion using Gaussian Mixture models or agglomerative clustering and then matches source and target class-conditional components for allocating the label proportion estimations. Resulting label proportion is then plugged into an weighted Wasserstein distance. When used for adversarial domain adaptation, we show that our approach outperforms competitors and is able to adapt to imbalance in target domains.

Several points are worth to be extended in future works. Our main assumption, for achieving estimations of class-conditionals, is the cyclical monotonicity of the class-conditional distributions in the latent space. However, unfortunately, we do not have any method for checking whether this assumption holds after training the representation on the source domain, especially as it supposed the knowledge of the class in the target domain. Hence, it would be interesting to enforce this assumption to hold, for instance by defining a regularization term based on the notion of cyclical monotonicity.

Furthermore, at the present time, we have considered simple mean-based approach for matching distributions, it is worth investigating whether higher-order moments are useful for improving the matching. Our algorithm relies mostly on our ability to estimate label proportion, we would be interested on in-depth theoretical analysis label proportion estimation and their convergence and convergence rate guarantees.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10994-021-06088-2.

Acknowledgements This work benefited from the support of the project OATMIL ANR-17-CE23-0012 of the French, LEAUDS ANR-18-CE23, and was performed using computing resources of CRIANN (Normandy, France). This work is partially funded through the projects 3IA Côte d'Azur Investments ANR-19-P3IA-0002 of the French National Research Agency (ANR). This research was produced within the framework of Energy4Climate Interdisciplinary Center (E4C) of IP Paris and Ecole des Ponts ParisTech. This research was supported by 3rd Programme d'Investissements d'Avenir ANR-18-EUR-0006-02. This action benefited from the support of the Chair "Challenging Technology for Responsible Energy" led by l'X – Ecole polytechnique and the Fondation de l'Ecole polytechnique, sponsored by TOTAL.

References

Alaux, J., Grave, E., Cuturi, M., & Joulin, A. (2019). Unsupervised hyper-alignment for multilingual word embeddings. In *7th International conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*.

Alvarez-Melis, D., Jegelka, S., & Jaakkola, T. S. (2019). Towards optimal transport with global invariances. In K. Chaudhuri, M. Sugiyama (Eds.), *Proceedings of machine learning research, vol. 89*, pp. 1870–1879.

Ambrosio, L. & Gigli, N. (2013). A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pp. 1–155. Springer.

Arora, S., Kannan, R., et al. (2005). Learning mixtures of separated nonspherical Gaussians. *The Annals of Applied Probability, 15*(1A), 69–92.

Azizzadenesheli, K., Liu, A., Yang, F., & Anandkumar, A. (2019). Regularized learning for domain adaptation under label shifts. In *International conference on learning representations (ICLR)*.

Birkhoff, G. (1946). Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucumán Rev. Ser. A*.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th International conference on pattern recognition*, pp. 3121–3124. IEEE.

Combes, R. T. D., Zhao, H., Wang, Y.-X., & Gordon, G. (2020). Domain adaptation with conditional distribution matching and generalized label shift. *arXiv preprint*arXiv:2003.04475.

Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(9), 1853–1865.

Ganin, Y. & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In F. Bach, D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning, vol. 37 of proceedings of machine learning research*, pp. 1180–1189, Lille, France.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research, 17*(1), 2096–2030.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., & Schölkopf, B. (2016). Domain adaptation with conditional transferable components. In *International conference on machine learning*, pp. 2839–2848.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., & Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning, 3*(4), 5.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of Wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., & Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pp. 601–608.

Johansson, F. D., Sontag, D. A., & Ranganath, R. (2019). Support and invertibility in domain-invariant representations. In K. Chaudhuri, M. Sugiyama (Eds.) *The 22nd international conference on artificial intelligence and statistics, AISTATS 2019, 16–18 April 2019, Naha, Okinawa, Japan, vol. 89 of proceedings of machine learning research*, pp. 527–536.

Kannan, R., Salmasian, H., & Vempala, S. (2005). The spectral method for general mixture models. In *International conference on computational learning theory*, pp. 444–457. Springer.

Li, Y., Murias, M., Major, S., Dawson, G., & Carlson, D. (2019). On target shift in adversarial domain adaptation. In K. Chaudhuri, M. Sugiyama (Eds.), *Proceedings of machine learning research, vol. 89 of proceedings of machine learning research*, pp. 616–625.

Lipton, Z. C., Wang, Y.-X., & Smola, A. (2018). Detecting and correcting for label shift with black box predictors. *arXiv preprint*arXiv:1802.03916.

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105.

Naim, I. & Gildea, D. (2012). Convergence of the EM algorithm for Gaussian mixtures with unbalanced mixing coefficients. In *Proceedings of the 29th international conference on machine learning, ICML 2012, Edinburgh, Scotland, UK, June 26–July 1, 2012*.

Pan, S. J., Tsang, I. W., Kwok, J. T., & Yang, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks, 22*(2), 199–210.

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., & Saenko, K. (2017). Visda: The visual domain adaptation challenge. *arXiv preprint*arXiv:1710.06924

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport. *Foundations and Trends®*

Redko, I., Courty, N., Flamary, R., & Tuia, D. (2019). Optimal transport for multi-source domain adaptation under target shift. In K. Chaudhuri, M. Sugiyama (Eds.), *Proceedings of machine learning research, vol. 89*, pp. 849–858.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*. Birkäuser, NY, 55(58–63), 94.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., & Mooij, J.(2012). On causal and anticausal learning. In *Proceedings of the 29th international conference on international conference on machine learning, ICML'12*, pp. 459–466, Madison, WI, USA. Omnipress.

Shen, J., Qu, Y., Zhang, W., & Yu, Y. (2018). Wasserstein distance guided representation learning for domain adaptation. In *Thirty-second AAAI conference on artificial intelligence*.

Shrikumar, A., Alexandari, A. M., & Kundaje, A. (2020). Adapting to label shift with bias-corrected calibration.

Sugiyama, M., Krauledat, M., & Muller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research, 8*(May), 985–1005.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., & Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. *arXiv preprint*arXiv:1412.3474

Venkateswara, H., Eusebio, J., Chakraborty, S., & Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. In *(IEEE) Conference on computer vision and pattern recognition (CVPR)*.

Wu, Y., Winston, E., Kaushik, D., Lipton, Z. (2019). Domain adaptation with asymmetrically-relaxed distribution alignment. In K. Chaudhuri, R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning*, vol. 97, pp. 6872–6881, Long Beach, California, USA.

Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., & Zuo, W. (2017). Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2272–2281.

Yu, Y. & Szepesvári, C. (2012). Analysis of kernel mean matching under covariate shift. In *Proceedings of the 29th international conference on machine learning, ICML 2012*, Edinburgh, Scotland, UK, June 26–July 1, 2012.

Zhang, K., Schölkopf, B., Muandet, K., & Wang, Z. (2013). Domain adaptation under target and conditional shift. In *International conference on machine learning*, pp. 819–827.

Zhao, H., Combes, R. T. D., Zhang, K., & Gordon, G. (2019). *On learning invariant representations for domain adaptation. vol. 97 of Proceedings of machine learning research*, pp. 7523–7532, Long Beach, California, USA.

Zhao, R., Li, Y., Sun, Y., et al. (2020). Statistical convergence of the em algorithm on Gaussian mixture models. *Electronic Journal of Statistics, 14*(1), 632–660.