Optimal Transport for Conditional Domain Matching and Label Shift

Alain Rakotomamonjy, Rémi Flamary, Gilles Gasso, M. El Alaya, Maxime Berar, Nicolas Courty

Match and Reweight Strategy for Generalized Target Shift

Alain Rakotomamonjy
LITIS, Univ. de Rouen
Criteo AI Lab, Paris
alain.rakoto@insa-rouen.fr

Rémi Flamary
Univ Côte d'Azur,CNRS , OCA Lagrange
remi.flamary@unice.fr

Gilles Gasso
LITIS, INSA de Rouen
gilles.gasso@insa-rouen.fr

Mokhtar El Alaya
LMAC, Univ. Technologique de Compiègne
elmokhtar.alaya@utc.fr

Maxime Berar
LITIS, Univ. de Rouen
maxime.berar@univ-rouen.fr

Nicolas Courty
Univ. Bretagne Sud, CNRS, Irisa
nicolas.courty@irisa.fr

Abstract

We address the problem of unsupervised domain adaptation under the setting of generalized target shift (both 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 probabilities are aligned accross 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 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*.

1 Introduction

Unsupervised Domain Adaptation (UDA) is a machine learning subfield that aims at addressing issues due to the discrepancy of train/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 ($p_S(x) \neq p_T(x)$ and $p_S(y|x) = p_T(y|x)$) 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 $p_S(y) \neq p_T(y)$ while $p_S(x|y) = p_T(x|y)$. 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; Shrikumar et al., 2020; Li et al., 2019; Redko et al., 2019).

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) ($p_S(z|y) \neq p_T(z|y)$). For this more general DA assumption, denoted as generalized target shift, fewer works have been proposed. Zhang et al. (2013) have been the first one that proposed a methodology for handling

Preprint. Under review.

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. Very 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 have zero distance, 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 Related works

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 introduce 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 learns invariant representations by using an assymetrically-relaxed distance. This relaxation allows to mitigate the effect of label distribution shift. In the case of Wasserstein distance, their approach consists in reweighting the marginal of the source distribution and in its dual form, their distance boils to

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

where the importance weight $w(\cdot)$ is actually a constant $\frac{1}{1+\beta}$. At the contrary, in our reweighting scheme $w(z)$ in an estimate of target label proportion. Hence, we believe that our approach would adapt better to imbalance without the need to tune β. The work of Combes et al. (2020) estimates $w(\cdot)$ by applying a technique tailored and grounded for problems without class-conditional shifts. Our differs in the way the weights $w(z)$ are estimated, an important distinction being that we consider a theoretically supported and consistent estimation of the target label proportion. We show in the experimental section that their estimator in some cases leads to poorer generalization. And more importantly, compared to their work, we provide theoretical guarantees and assumptions under which, the solution of our learning problem leads to matched class-conditionals which is key for generalization.

3 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 mapping from $\mathcal{X}$ to $\mathcal{Z}$. Similarly, $\mathcal{H}$ represents the hypothesis space, which is a set of function 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 $y_i^s \in \{1 \dots C\}$ 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 $U \in \{S, T\}$.

Domain adaptation framework Since the seminal work of Pan et al. (2010); Long et al. (2015); Ganin & 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:

$$\min_{h,g} \frac{1}{n} \sum_{i=1}^{n_s} L(y_i^s, h(g(x_i^s))) + \lambda D(p_S^g, p_T^g) + \Omega(h, g) \tag{2}$$

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).

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 $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. 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 Kantorovitch relaxation of the OT problem seeks for a transportation coupling $\mathbf{P}$ that minimizes the problem

$$\min_{\mathbf{P} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle \tag{3}$$

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 $\Pi(\mathbf{a}, \mathbf{b}) = \{\mathbf{P} \in \mathbb{R}_+^{n \times m} | \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b}\}$ 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).

4 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 $f : \mathcal{X} \mapsto [0, 1]$ be the domain-invariant labeling function, which is a classical assumption in DA (Wu et al., 2019; Shen et al., 2018). For a domain U, with $U = \{S, T\}$, the induced marginal probability of samples in $\mathcal{Z}$ is formally defined as $p^g_U(A) = p_U(g^{-1}(A))$ for any subset $A \subset \mathcal{Z}$ and $g^{-1}(A)$ being potentially a set ($p^g_U(A)$ is thus the push-forward of $p_U(x)$ by $g(\cdot)$). Similarly, we define the conditional distribution $g_U(\cdot|z)$ such that $p_U(x) = \int g_U(x|z)p^g_u(z)dz$ 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 $\varepsilon_U(h \circ g, f) \triangleq \mathbb{E}_{x\sim p_U}[|h(g(x)) - f(x)|] = \mathbb{E}_{z\sim p^z_U}[|h(z) - f^g_U(z)|] \triangleq \varepsilon^z_U(h, f^g_U)$ with f^g_U being a domain-dependent labeling function defined as $f^g_U(z) = \int f(x)g_U(x|z)dx$.

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^g_S and p^g_T, with $p^g_U = \sum_{k=1}^{C} p^{y=k}_U p^k_U$, $U = \{S, T\}$. For all p^y_T, p^y_S and for any continuous class-conditional density distribution p^k_S and p^k_T 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 $w(z) = \frac{p^{y=k}_T}{p^{y=k}_S}$, 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 $p^{y=k}_S = p^{y=k}_T$ and $p^k_S = p^k_T$. Interestingly, Wu et al. (2019)'s results involve a similar term $\sup_z \frac{p^g_T(z)}{p^g_S(z)}$ 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*

$$\begin{aligned}\varepsilon_T(h \circ g, f) \leq \varepsilon_S(h \circ g, f) &+ 2K \cdot WD_1(p^g_S, p^g_T) \\ &+ \left[1 + \sup_{k,z} w(z)S_k(z))\right] \varepsilon_S(h^\star \circ g, f) \\ &+ \varepsilon^z_T(f^g_S, f^g_T)\end{aligned}$$

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 $\varepsilon^z_T(f^g_S, f^g_T) = \mathbb{E}_{z\sim p^z_T}[|f^g_T(z) - f^g_S(z)|]$ and WD_1 as defined in Equation 1 with $w(\cdot) = 1$.

Let us analyze the terms that bound the target generalization error. The first term $\varepsilon_S(h \circ g, f) \triangleq \varepsilon^z_S(h, f^g_S)$ 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 $\varepsilon_T(f^g_S, f^g_T)$ 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 $\varepsilon^z_T(f^g_S, f^g_T) = 0$.

5 Match and Reweight Strategy

5.1 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

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 $\theta_h, \theta_g, \theta_v$
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 $\mathcal{L}_{wd} = \sum_i w^\dagger(z_i^s)v(z_i^s) - \frac{1}{B}\sum_i v(z_i^t)$
    $\theta_v \leftarrow \theta_v + \alpha_v \nabla_{\theta_v}[\mathcal{L}_{wd} - \mathcal{L}_{grad}]$
  end for
  compute the weighted classification loss $\mathcal{L}_w = \sum_i w^\dagger(z_i^s)L(y_i^s, h(g(x_i^s)))$
  $\theta_h \leftarrow \theta_h + \alpha_h \nabla_{\theta_h}\mathcal{L}_w$
  $\theta_g \leftarrow \theta_g + \alpha_g \nabla_{\theta_g}[\mathcal{L}_w + \mathcal{L}_{wd}]$
until a convergence condition is met

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 $p_{\tilde{S}} = \sum_{k=1}^C p_T^{y=k} p_S^i$, 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 on $p_{\tilde{S}}$ and p_T, we propose to learn the functions h and g by solving the problem

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

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, the discrepancy between marginals is the Wasserstein distance

$$WD_1(\tilde{p}_s^g, p_t^g) = \sup_{\|v\|_L \leq 1} \mathbf{E}_{z\sim p_{\tilde{S}}^g} w^\dagger(z)v(z) - \mathbf{E}_{z\sim p_T^g} v(z).$$

The first term of equation (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 Equation (4),we employ a classical adversarial learning strategy (detailed in algorithm 1) and we use gradient penalty for estimating the Wasserstein distance (Gulrajani et al., 2017).

5.2 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 $p_{\tilde{S}}^g$ and p_T^g. Several approaches exist for estimating

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 (3) 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)

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, it 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.

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 4 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 to another set of finite number C of probability distributions, 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.

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 domain. 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*

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

then solving the optimal transport problem between ν *and* μ *as defined in equation* (3) *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 $\sum_j^C \mathcal{D}(p_S^j, p_T^j) \leq \sum_j^{\mathcal{D}} (p_S^j, p_T^{\sigma(j)}), \forall \sigma$. We can note that while our assumptions can be considered as strong, we illustrate in Figure 7 in the supplementary material, that those hypotheses hold for the VisDA problem.

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.

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.

5.3 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* $\gamma = \frac{1}{C} diag(\mathbb{1})$ *then* $\gamma' = diag(\mathbf{a})$ *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

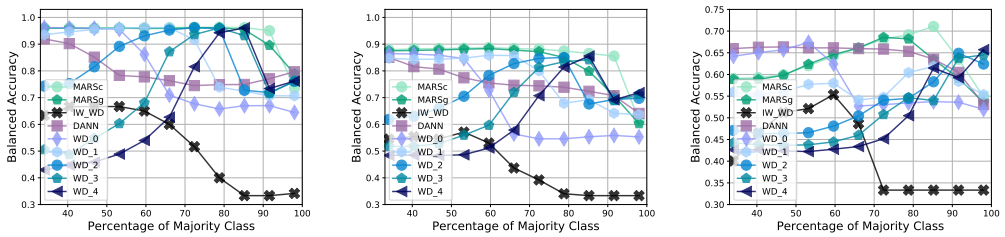

Figure 1: Performance of the compared algorithms for three different covariance matrices of the Gaussians composing the toy dataset with respect to the imbalance. The x-axis is given with respect to the percentage of majority class which is the class 1. (left) Low-error setting. (middle) mid-error setting.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

Table 1: Table of averaged **balanced accuracy** for the compared models and different domain adaptation problems and label proportion imbalance settings. 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 presented in the supplementary (including experiments on *Office*). #Win includes the statistical ties.

Setting	Source	DANN	$\text{WD}_{\beta=0}$	$\text{WD}_{\beta=1}$	$\text{WD}_{\beta=2}$	$\text{WD}_{\beta=3}$	$\text{WD}_{\beta=4}$	IW-WD	MARSg	MARSc
MNIST-USPS 10 modes										
Balanced	76.89±3.7	79.74±3.5	93.71±0.7	74.27±4.3	51.33±4.0	76.61±3.3	71.90±5.7	95.28±0.4	**95.61±0.7**	**95.64±1.0**
Mid	80.41±3.1	78.65±3.0	94.30±0.7	75.36±3.4	55.55±4.3	78.98±3.1	72.32±4.2	**95.60±0.5**	89.70±2.3	90.39±2.6
High	78.13±4.9	81.79±4.0	**93.86±1.1**	87.44±1.7	83.83±5.2	85.65±2.5	83.65±3.0	**94.08±1.0**	88.30±1.5	89.65±2.3
USPS-MNIST 10 modes										
Balanced	77.04±2.6	80.49±2.2	73.35±2.8	66.70±2.9	49.86±2.8	55.83±2.9	52.12±3.5	80.52±2.2	**84.59±1.7**	**85.50±2.1**
Mid	**79.54±2.8**	**78.88±1**.8	75.85±1.6	63.33±2.3	53.22±2.8	47.20±2.4	48.29±2.9	**78.36±3.5**	**79.73±3.6**	**78.49±2.5**
High	**78.48±2.4**	**77.79±2.0**	**76.14±2.7**	63.00±3.3	57.56±4.8	51.19±4.4	49.31±3.3	71.53±4.7	75.62±1.8	**77.14±2.4**
MNIST-MNISTM 10 modes										
Setting 1	58.34±1.3	**61.22±1.1**	57.44±1.7	50.20±4.4	47.01±2.0	57.85±1.1	55.95±1.3	**63.10±3.1**	58.08±2.3	56.58±4.6
Setting 2	59.94±1.1	61.09±1.0	58.08±1.4	53.39±3.5	48.61±2.4	59.74±0.7	58.14±0.8	**65.03±3.5**	57.69±2.3	55.64±2.1
Setting 3	58.14±1.2	**60.39±1.4**	57.68±1.2	47.72±4.9	42.15±7.3	57.09±1.0	53.52±1.1	52.46±14.8	53.68±7.2	53.72±3.3
VisdDA 3 modes										
setting 1	79.28±4.3	78.83±9.1	91.83±0.7	73.78±2.0	61.65±2.2	65.62±2.7	58.58±2.6	**94.11±0.6**	92.47±1.2	92.13±1.8
setting 4	80.15±5.3	75.46±9.3	72.75±1.2	86.86±7.5	86.82±1.2	80.16±6.9	75.71±2.0	85.88±5.7	87.69±3.0	**91.29±4.8**
setting 2	81.47±3.5	68.46±14.7	68.81±1.3	84.45±1.2	**93.15±0.4**	73.65±14.2	60.67±0.9	78.73±10.8	84.04±4.3	**91.80±3.4**
setting 3	78.35±3.2	58.93±15.9	64.13±1.9	79.17±0.8	77.12±10.3	89.93±0.5	**94.38±0.3**	77.96±9.3	75.68±4.1	73.81±13.2
setting 5	83.52±3.5	80.83±14.5	63.82±0.6	73.70±7.3	50.91±1.1	76.52±6.7	59.28±1.0	**90.40±3.6**	**89.01±0.9**	**89.03±3.5**
setting 6	80.84±4.2	54.76±19.8	45.27±2.4	63.70±5.1	67.05±6.1	42.86±10.8	62.21±1.4	**94.36±1.0**	**93.70±0.4**	**93.86±1.0**
setting 7	79.22±3.7	42.94±2.5	57.51±1.5	55.39±2.0	50.22±4.3	43.66±8.3	62.47±0.8	**88.52±4.9**	78.56±3.2	**82.33±7.5**
VisDA 12 modes										
setting 1	41.90±1.5	52.79±2.1	45.81±4.3	44.23±3.0	35.45±4.6	40.96±3.0	37.59±3.4	50.35±2.3	53.31±0.9	**55.05±1.6**
setting 2	41.75±1.5	50.82±1.6	45.72±8.9	40.49±4.8	36.21±5.0	36.12±4.6	31.86±5.7	48.59±1.8	53.09±1.6	**55.33±1.6**
setting 3	40.64±4.3	49.17±1.3	47.12±1.6	42.10±3.0	36.32±4.4	37.26±3.5	34.96±5.4	46.59±1.3	50.78±1.6	**52.08±1.2**
#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**

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 $p_{\tilde{S}}^g$).

6 Numerical Experiments

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 apply adversarial domain learning using approximation of the $\mathcal{H}$ divergence and the Wasserstein distance computed in the dual as distances for measuring discrepancy between p_S and p_T, denoted as DANN and $\text{WD}_{\beta=0}$. 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

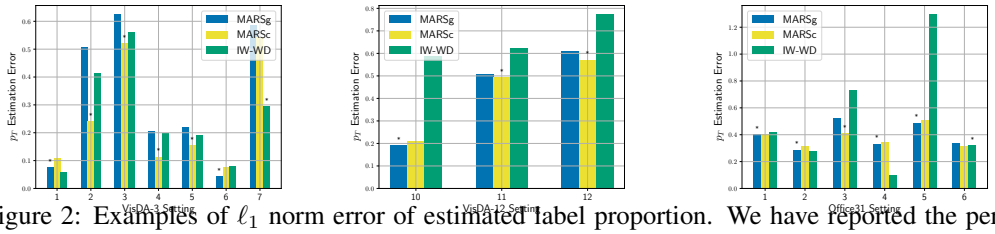

Figure 2: Examples of ℓ_1 norm error of estimated label proportion. We have reported the performance of our two methods (MARSg and MARSc) as well as the performance of IW-WD. The three panels are related to the (left) VisDA-3, (middle) VisDA-12, (right) Office 31 and the different experimental imbalance settings (see Table 3). 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.

asymmetrically-relaxed Wasserstein distance so as to make it similar to our approach and also reported results for different values of the relaxation β. This model is named WD$_\beta$ 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 $w(x_i)$ 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 Equation 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.

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 1 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$_\beta$ 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 estimated 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 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.

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. 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 2 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.

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.

References

Alaux, J., Grave, E., Cuturi, M., and Joulin, A. Unsupervised hyper-alignment for multilingual word embeddings. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL `https://openreview.net/forum?id=HJe62s09tX`.

Alvarez-Melis, D., Jegelka, S., and Jaakkola, T. S. Towards optimal transport with global invariances. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1870–1879. PMLR, 16–18 Apr 2019.

Ambrosio, L. and Gigli, N. A users guide to optimal transport. In *Modelling and optimisation of flows on networks*, pp. 1–155. Springer, 2013.

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://openreview.net/forum?id=rJl0r3R9KX`.

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

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124. IEEE, 2010.

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

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1180–1189, Lille, France, 07–09 Jul 2015. PMLR. URL `http://proceedings.mlr.press/v37/ganin15.html`.

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

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

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

Johansson, F. D., Sontag, D. A., and Ranganath, R. Support and invertibility in domain-invariant representations. In Chaudhuri, K. and Sugiyama, M. (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 527–536. PMLR, 2019. URL `http://proceedings.mlr.press/v89/johansson19a.html`.

Li, Y., Murias, M., Major, S., Dawson, G., and Carlson, D. On target shift in adversarial domain adaptation. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 616–625. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/li19b.html.

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

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

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

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

Peyré, G., Cuturi, M., et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Redko, I., Courty, N., Flamary, R., and Tuia, D. Optimal transport for multi-source domain adaptation under target shift. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 849–858. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/redko19a.html.

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

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML12, pp. 459466, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Shrikumar, A., Alexandari, A. M., and Kundaje, A. Adapting to label shift with bias-corrected calibration, 2020. URL https://openreview.net/forum?id=rkx-wA4YPS.

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

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

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *(IEEE) Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Wu, Y., Winston, E., Kaushik, D., and Lipton, Z. Domain adaptation with asymmetrically-relaxed distribution alignment. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6872–6881, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/wu19f.html.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pp. 819–827, 2013.

Zhao, H., Combes, R. T. D., Zhang, K., and Gordon, G. On learning invariant representations for domain adaptation. volume 97 of *Proceedings of Machine Learning Research*, pp. 7523–7532, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/zhao19a.html.

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

Supplementary material for Match and Reweight for Generalized Target Shift

This supplementary material presents some details of the theoretical and algorithmic aspects of the work as well as as some additional results. They are listed as below.

1. Theoretical details and proofs
2. The full algorithm for MARS is detailed and a pseudo-code is given in Algorithm 1
3. Dataset details and architecture details are given in Section 9.1 and 9.2
4. Figure 3 presents some samples of the 3-class toy data set for different configurations of covariance matrices making the problem easy, of mid-difficulty or difficult.
5. Examples of source and target class-conditionals that allow class matching through optimal transport 4 as discussed in Proposition 1.
6. Figure 5 exhibits the performances of the compared algorithms depending on the shift of the class-conditional distributions.
7. Figure 6 shows for the imbalanced toy problem, the results obtained by all competitors including a GMM.
8. Table 2 shows the performance of Source only and a simple GMM+OT on a Visda 3-class problem.
9. Table 3 depicts the different configurations of the dataset we used in our experiments
10. The full table presenting the experimental results for all competitors on different dataset settings is in Table 4.
11. Examples of label proportion error estimation is given in Figure **??**.
12. Examples of *t-sne* embeddings on the VisDA-3 problem, given in Figure 7 illustrating the features obtained by DANN, $\text{WD}_\beta = 1$, IW-WD and MARSc.

8 Theoretical and algorithmic details

8.1 Lemma 1 and its proof

Lemma 1. *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 following inequality holds.*

$$\sup_{k,z}[w(z)S_k(z)] \geq 1$$

with $S_k(z) = \frac{p_T^g(z|y=k)}{p_S^g(z|y=k)}$ and $w(z) = \frac{p_T^{y=k}}{p_S^{y=k}}$, if z is of class k.

Proof. Let first show that for any k the ratio $\sup_z \frac{p_T^k}{p_S^k} \geq 1$. Suppose that there does not exist a z such that $\frac{p_T^k}{p_S^k} \geq 1$. This means that : $\forall z\ p_T^k < p_S^k$. By integrating those positive and continuous functions on their domains lead to the contradiction that the integral of one of them is not equal to 1. Hence, there must exists a z such that $\frac{p_T^k}{p_S^k} \geq 1$. Hence, we indeed have ratio $\sup_z \frac{p_T^k}{p_S^k} \geq 1$.

Using a similar reasoning, we can show that $\sup_k \frac{p_T^{y=k}}{p_S^{y=k}} \geq 1$. For a sake of completeness, we provide it here. Assume that $\forall k,\ p_T^{y=k} < p_S^{y=k}$. We thus have $\sum_k p_T^{y=k} < \sum_k p_S^{y=k}$. Since noth sums should be equal to 1 leads to a contradiction.

By exploiting these two inequalities, we have :

$$\sup_{k,z}[w(z)S_k(z)] = \sup_k \left[w(z) \sup_z S_k(z) \right] \geq \sup_k w(z) \geq 1$$

□

8.2 Theorem 1 and its proof

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*

$$\varepsilon_T(h\circ g, f) \leq \varepsilon_S(h\circ g, f) + 2K \cdot WD_1(p^g_S, p^g_T) + \left[1 + \sup_{k,z} w(z)S_k(z))\right] \varepsilon_S(h^\star \circ g, f) + \varepsilon^z_T(f^g_S, f^g_T)$$

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 $\varepsilon^z_T(f^g_S, f^g_T) = \mathbb{E}_{z\sim p^z_T}[|f^g_T(z) - f^g_S(z)|]$

Proof. At first, let us remind the following result due to Shen et al. (2018). Given two probability distributions p^g_S and p^g_T, we have

$$\varepsilon^z_S(h, h') - \varepsilon^z_T(h, h') \leq 2K \cdot WD_1(p^g_S, p^g_T)$$

for every hypothesis h, h' in $\mathcal{H}$. Then, we have the following bound for the target error

$$\begin{aligned}
\varepsilon_T(h \circ g, f) &\leq \varepsilon_T(h \circ g, h^\star) + \varepsilon_T(h^\star \circ g, f) && (5)\\
&\leq \varepsilon_T(h \circ g, h^\star) + \varepsilon_S(h \circ g, h^\star) - \varepsilon_S(h \circ g, h^\star) + \varepsilon_T(h^\star \circ g, f) && (6)\\
&\leq \varepsilon_S(h \circ g, h^\star) + \varepsilon_T(h^\star \circ g, f) + 2K \cdot WD_1(p^g_S, p^g_T) && (7)\\
&= \varepsilon^z_S(h, h^\star) + \varepsilon^z_T(h^\star, f^g_T) + 2K \cdot WD_1(p^g_S, p^g_T) && (8)\\
&\leq \varepsilon^z_S(h, f^g_S) + \varepsilon^z_S(h^\star, f^g_S) + \varepsilon^z_T(h^\star, f^g_T) + 2K \cdot WD_1(p^g_S, p^g_T) && (9)\\
&\leq \varepsilon^z_S(h, f^g_S) + \varepsilon^z_S(h^\star, f^g_S) + \varepsilon^z_T(h^\star, f^g_S) + \varepsilon^z_T(f^g_S, f^g_T) + 2K \cdot WD_1(p^g_S, p^g_T) && (10)
\end{aligned}$$

where the lines (5), (9), (10) have been obtained using triangle inequality, Line (7) by applying Shen's et al. above inequality, Line (8) by using $\varepsilon_U(h \circ g, f) = \varepsilon^z_U(h, f^g_U)$. Now, let us analyze the term $\varepsilon^z_S(h^\star, f^g_S) + \varepsilon^z_T(h^\star, f^g_S)$. Denote as $r_S(z) = |h^\star(z) - f^g_S(z)|$. Hence, we have

$$\begin{aligned}
\varepsilon^z_S(h^\star, f^g_S) + \varepsilon^z_T(h^\star, f^g_S) &= \int r_S(z)[p^g_S(z) + p^g_T(z)]dz && (11)\\
&= \sum_k p_S(y = k) \int r_S(z) p^g_S(z|y = k)\big[1 + \frac{p_T(y = k)}{p_S(y = k)} S_k(z)\big]dz && (12)\\
&\leq \left(1 + \sup_{k,z}[w(z)S_k(z)]\right) \varepsilon_S(h^\star, f^g_S) && (13)
\end{aligned}$$

where Line (12) has been obtaine by expanding marginal distriutions. Merging the last inequality into Equation (10) concludes the proof. □

8.3 Proposition 1 and its proof

Proposition 1. *Denote as $\nu = \frac{1}{C}\sum_{j=1}^C \delta_{p^j_S}$ and $\mu = \frac{1}{C}\sum_{j=1}^C \delta_{p^j_T}$, representing respectively the class-conditional probabilities in source and target domain. 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*

$$\sum_j \mathcal{D}(p^j_S, p^j_T) \leq \sum_j \mathcal{D}(p^j_S, p^{\sigma(j)}_T)$$

then solving the optimal transport problem between ν and μ as defined in equation (3) *using $\mathcal{D}$ as the ground cost matches correctly class-conditional probabilities.*

Proof. The solution $\mathbf{P}^*$ of the OT problem lies on an extremal point of Π_C. Birkhoff's theorem Birkhoff (1946) states that the set of extremal points of Π_C is the set of permutation matrices so that there exists an optimal solution of the form $\sigma^* : [1, \cdots, C] \to [1, \cdots, C]$. The support of $\mathbf{P}^*$ is $\mathcal{D}$-cyclically monotone (Ambrosio & Gigli, 2013; Santambrogio, 2015) (Theorem 1.38), meaning that $\sum_j^C \mathcal{D}(p^j_S, p^{\sigma^*(j)}_T) \leq \sum_j^C \mathcal{D}(p^j_S, p^{\sigma(j)}_T), \forall \sigma \neq \sigma^*$. Then, by hypothesis, σ^* can be identified to the identity permutation, and solving the optimal assignment problem matches correctly class-conditional probabilities. □

8.4 Proposition 2 and its proof

Proposition 2. *Denote as γ the optimal coupling plan for distributions ν and μ with balanced class-conditionals such that $\nu = \frac{1}{C}\sum_{j=1}^{C}\delta_{p_S^j}$ 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 $\gamma = \frac{1}{C}diag(\mathbb{1})$ then $\gamma' = diag(\mathbf{a})$ 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.*

Proof. By assumption and without loss of generality, the class-conditionals are arranged so that $\gamma = \frac{1}{C}\text{diag}(\mathbb{1})$. Because the weights in the marginals are not uniform anymore, γ is not a feasible solution for the OT problem with ν' and μ' but $\gamma' = \text{diag}(\mathbf{a})$ is. Let us now show that any feasible non-diagonal plan Γ has higher cost than γ' and thus is not optimal. At first, consider any permutation σ of C elements and its corresponding permutation matrix $\mathbf{P}_\sigma$, because $\gamma = \frac{1}{C}\text{diag}(\mathbb{1})$ is optimal, the cyclical monotonicity relation $\sum_i \mathcal{D}_{i,i} \leq \sum_i \mathcal{D}_{i,\sigma(i)}$ holds true $\forall\sigma$. Starting from $\gamma' = \text{diag}(\mathbf{a})$, any direction $\Delta_\sigma = -\mathbf{I} + \mathbf{P}_\sigma$ is a feasible direction (it does not violate the marginal constraints) and due to the cyclical monotonicity, any move in this direction will increase the cost. Since any non-diagonal $\gamma_z \in \Pi(\mathbf{a}, \mathbf{a})$ can be reached with a sum of displacements Δ_σ (property of the Birkhoff polytope) it means that the transport cost induced by γ_z will always be greater or equal to the cost for the diagonal γ' implying that γ' is the solution of the OT problem with marginals $\mathbf{a}$.
As a corollary, it is straightforward to show that $W(\nu', \mu') = \sum_{i=1}^{C} \mathcal{D}_{i,i}a_i = 0 \implies \mathcal{D}_{i,i} = 0$ as $a_i > 0$ by hypothesis. □

8.5 Algorithm for training the full MARS model

We present here the algorithm we have used for training the full model. It is based on a standard backpropagation strategy using stochatic gradient descent. We estimate the label proportion using Algorithm 2 and then uses 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 backpropagated the error through the parameters of the feature extractor g and the classifier f.

In practice, we first train the model without adaptation (hence only based on the classification loss with uniform weights, until reaching 0 training errors and then start adapting as detailed in Algorithm 1)

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
Ensure: $\mathbf{p}$: label proportion
Initialize feature extractor 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 $\{w_i\}_{i=1}^{C}$ based on the source proportion in the batch samples and $\mathbf{p}_T$
  for $t = 1, \cdots, n$ **do**
    $x_e^s \leftarrow g(x_B^s)$, $x_e^t \leftarrow g(x_B^t)$
    sample random points x' from the lines between x_e^s and x_e^t pairs.
    compute gradient penalty $\mathcal{L}_{\text{grad}}$ using x_e^s, x_e^t and x'
    compute empirical Wasserstein dual loss $\mathcal{L}_{wd} = \sum_i w(\mathbf{x}_i^s)v(x_i^s) - \frac{1}{B}\sum_i v(x_i^t)$
    $\theta_v \leftarrow \theta_v + \alpha_v \nabla_{\theta_v}[\mathcal{L}_{wd} - \mathcal{L}_{grad}]$
  end for
  compute the weighted classification loss $\mathcal{L}_w = \sum_i w(x_i^s)L(y_i^s, x_e^s)$
  $\theta_h \leftarrow \theta_h + \alpha_h \nabla_{\theta_h}\mathcal{L}_w$
  $\theta_g \leftarrow \theta_g + \alpha_g \nabla_{\theta_g}[\mathcal{L}_w + \mathcal{L}_{wd}]$
until fdf

9 Experimental Results

9.1 Dataset details

We have considered 4 family of domain adaptation problems based on the digits, Visda, Office-31 and Office-Home dataset. For all these datasets, we have not considered the natural train/test number of examples, in order to be able to build different label distributions at constant number of examples (suppose one class has at most 800 examples, if we want that class to represent 80% of the samples, then we are limited to 1000 samples).

For the digits problem, We have used the MNIST, USPS and the MNITSM datasets. we have learned the feature extractor from scratch and considered the following train-test number of examples setting. For MNIST-USPS, USPS-MNIST and MNIST-MNISTM, we have respectively used 60000-3000, 7291-10000, 10000-10000.

The VisDA 2017 problem is a 12-class classification problem with source and target domain being simulated and real images. We have considerd two sets of problem, a 3-class one (based on the classes *aeroplane, horse* and *truck*) and the full 12-class problem.

The Office-31 is an object categorization problem involving 31 classes with a total of 4652 samples. There exists 3 domains in the problem based on the source of the images : Amazon (A), DSLR (D) and WebCam (W). We have considered all possible pairwise source-target domains.

The Office-Home is another object categorization problem involving 65 classes with a total of 15500 samples. There exists 4 domains in the problem based on the source of the images : Art, Product, Clipart (Clip), Realworld (Real).

For the Visda and Office datasets, we have considered Imagenet pre-trained ResNet-50 features and our feature extractor (which is a fully-connected feedforword networks) aims at adapting those features. We have used pre-trained features freely available at `https://github.com/jindongwang/transferlearning/blob/master/data/dataset.md`.

9.2 Architecture details

Toy The feature extractor is a 2 layer fully connected network with 200 units and ReLU activation function. The classifier is also a 2 layer fully connected network with same number of units and activation function. Discriminators have 3 layers with same number of units.

Digits For the MNIST-USPS problem, the architecture of our feature extractor is composed of the two CNN layers with 32 and 20 filters of size 5×5 and 2-layer fully connected networks as discriminators with 100 and 10 units. The feature extractor uses a ReLU activation function and a max pooling. For he MNIST-MNISTM adaptation problem we have used the same feature extractor network and discriminators as in Ganin & Lempitsky (2015).

VisDA For the VisDA dataset, we have considered pre-trained 2048 features obtained from a ResNet-50 followed by 2 fully connected networks with 100 units and ReLU activations. The latent space is thus of dimension 100. Discriminators and classifiers are also a 2 layer Fully connected networks with 100 and respectively 1 and "number of class" units.

Office For the office datasets, we have considered pre-trained 2048 features obtained from a ResNet-50 followed by two fully connected networks with output of 100 and 50 units and ReLU activations. The latent space is thus of dimension 50. Discriminators and classifiers are also a 2 layer fully connected networks with 50 and respectively 1 and "number of class" units.

For Digits and VisDA and Office applications, all models have been trained using ADAM for 100 iterations with validated learning rate, while for the toy problem, we have used a SGD.

Table 2: Comparing Source-Only model and GMM+OT approach on the VisDA-3-mode problems. We can note that for these problems where the latent space is of dimension 100, the GMM+OT compares poorly to Source-Only. In addition, we can note that there is very high variability in the performance.

Configuration	Source	GMM+OT
Setting 1	79.28±4.3	81.22±4.7
Setting 4	80.15±5.3	76.28±9.8
Setting 2	81.47±3.5	74.79±10.4
Setting 3	78.35±3.2	69.97±10.8
Setting 5	83.52± 3.5	76.95±10.4
Setting 6	80.84±4.2	72.86±10.2
Setting 7	79.22±3.7	69.48±9.8

Table 3: Table of the dataset experimental settings. We have considered different domain adaptation problems and different configurations of the label shift in the source and target domain. For the digits and VisDA problem, we provide the ratio of samples of classes for each problem (*e.g.,* for the third setting of VisDA-3 problem, the second class accounts for the 70% of the samples in target domain). For Office datasets, because of large amount of classes, we have changed percent of samples of that class in the source or target (*e.g.,* the 10-class in Office Home uses respectively 20% and 100% of its sample for the source and target domain).

Configuration	Proportion Source	Proportion Target
MNIST-USPS balanced	$\{\frac{1}{10}, \cdots, \frac{1}{10}\}$	$\{\frac{1}{10}, \cdots, \frac{1}{10}\}$
MNIST-USPS mid	$\{\frac{1}{10}, \cdots, \frac{1}{10}\}$	$\{0, \cdots, 3, 6\} = 0.02, \{4, 5\} = 0.02, \{7, 8, 9\} = 0.1$
MNIST-USPS high	$\{\frac{1}{10}, \cdots, \frac{1}{10}\}$	$\{0\} = 0.3665, \{1\} = 0.3651, \{2, \cdots\} = 0.0335$
USPS-MNIST balanced	$\{\frac{1}{10}, \cdots, \frac{1}{10}\}$	$\{\frac{1}{10}, \cdots, \frac{1}{10}\}$
USPS-MNIST mid	$\{\frac{1}{10}, \cdots, \frac{1}{10}\}$	$\{0, \cdots, 3, 6\} = 0.02, \{4, 5\} = 0.02, \{7, 8, 9\} = 0.1$
USPS-MNIST high	$\{\frac{1}{10}, \cdots, \frac{1}{10}\}$	$\{0\} = 0.3665, \{1\} = 0.3651, \{2, \cdots\} = 0.0335$
MNIST-MNISTM (1)	$\{0-4\} = 0.05, \{5-9\} = 0.15$	$\{0, \cdots, 3, 6\} = 0.02, \{4, 5\} = 0.02, \{7, 8, 9\} = 0.1$
MNIST-MNISTM (2)	$\{0-2\} = 0.26, \{3-9\} = 0.03$	$\{0-6\} = 0.03, \{7-9\} = 0.26$
MNIST-MNISTM (3)	$\{0-5\} = 0.05, \{6-9\} = 0.175$	$\{0-3\} = 0.175, \{4-9\} = 0.05$
VisDA-3 (1)	{0.33,0.33,0.34}	{0.33,0.33,0.34}
VisDA-3 (2)	{0.4,0.2,0.4}	{0.2,0.6,0.2}
VisDA-3 (3)	{0.4,0.2,0.4}	{0.15,0.7,0.15}
VisDA-3 (4)	{0.4,0.2,0.4}	{0.1,0.8,0.1}
VisDA-3 (5)	{0.6,0.2,0.2}	{0.2,0.2,0.6}
VisDA-3 (6)	{0.6,0.2,0.2}	{0.15,0.2,0.65}
VisDA-3 (7)	{0.6,0.2,0.2}	{0.2,0.65,0.15}
VisDA-12 (1)	$\{\frac{1}{12}, \cdots, \frac{1}{12}\}$	$\{\frac{1}{12}, \cdots, \frac{1}{12}\}$
VisDA-12 (2)	$\{\frac{1}{12}, \cdots, \frac{1}{12}\}$	$\{0-3\} = 0.15, \{4-11\} = 0.05$
VisDA-12 (3)	$\{\frac{1}{12}, \cdots, \frac{1}{12}\}$	$\{0-1\} = 0.2, \{2-5\} = 0.1, \{6-11\} = 0.03$
Office-31	$\{0-15\} : 30\%$ $\{15-31\} : 80\%$	$\{0-15\} : 80\%$ $\{15-31\} : 30\%$
Office-Home	$\{0-32\} : 20\%$ $\{33-65\} : 100\%$	$\{0-32\} : 100\%$ $\{33-65\} : 20\%$

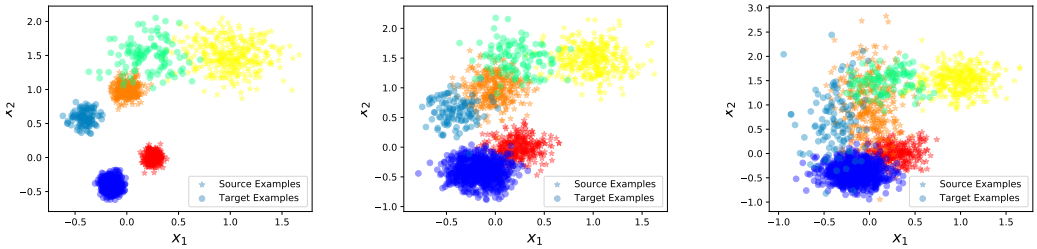

Figure 3: Examples of source and target domain examples. For each domain, data are composed of three Gaussians defining each class. In the source domain, classes are balanced whereas in the target domain, we have a ratio of $0.8, 0.1, 0.1$. The three configurations presented here vary in their covariance matrices. From left to right, we have Gaussians that are larger and larger making them difficult to classify. In the most right examples, the second class of the source domain and the third one of the target domain are mixed. This region becomes indecidable for our model as the source loss want to classify it as "Class 2" while the Wasserstein distance want to match it with "Class 3" of the source domain.

Table 4: Table of averaged balanced accuracy for the compared models and different domain adaptation models. Number of runs used 20. Reported in bold are the best performances as well as other methods which achieves performance that are statistically similar according to a Wilcoxon signrank test with $p = 0.01$.

Setting	Source	DANN	$\text{WD}_{\beta=0}$	$\text{WD}_{\beta=1}$	$\text{WD}_{\beta=2}$	$\text{WD}_{\beta=3}$	$\text{WD}_{\beta=4}$	IW-WD	MARSg	MARSc
MNIST-USPS 10 modes										
Balanced	76.89±3.7	79.74±3.5	93.71±0.7	74.27±4.3	51.33±4.0	76.61±3.3	71.90±5.7	95.28±0.4	**95.61±0.7**	**95.64±1.0**
Mid	80.41±3.1	78.65±3.0	94.30±0.7	75.36±3.4	55.55±4.3	78.98±3.1	72.32±4.2	**95.60±0.5**	89.70±2.3	90.39±2.6
High	78.13±4.9	81.79±4.0	**93.86±1.1**	87.44±1.7	83.83±5.2	85.65±2.5	83.65±3.0	**94.08±1.0**	88.30±1.5	89.65±2.3
USPS-MNIST 10 modes										
Balanced	77.04±2.6	80.49±2.2	73.35±2.8	66.70±2.9	49.86±2.8	55.83±2.9	52.12±3.5	80.52±2.2	**84.59±1.7**	**85.50±2.1**
Mid	**79.54±2.8**	**78.88±1.8**	75.85±1.6	63.33±2.3	53.22±2.8	47.20±2.4	48.29±2.9	**78.36±3.5**	**79.73±3.6**	**78.49±2.5**
High	**78.48±2.4**	**77.79±2.0**	**76.14±2.7**	63.00±3.3	57.56±4.8	51.19±4.4	49.31±3.3	71.53±4.7	75.62±1.8	**77.14±2.4**
MNIST-MNISTM 10 modes										
Setting 1	58.34±1.3	**61.22±1.1**	57.44±1.7	50.20±4.4	47.01±2.0	57.85±1.1	55.95±1.3	**63.10±3.1**	58.08±2.3	56.58±4.6
Setting 2	59.94±1.1	61.09±1.0	58.08±1.4	53.39±3.5	48.61±2.4	59.74±0.7	58.14±0.8	65.03±3.5	57.69±2.3	55.64±2.1
Setting 3	58.14±1.2	**60.39±1.4**	57.68±1.2	47.72±4.9	42.15±7.3	57.09±1.0	53.52±1.1	52.46±14.8	53.68±7.2	53.72±3.3
VisdDA 3 modes										
setting 1	79.28±4.3	78.83±9.1	91.83±0.7	73.78±2.0	61.65±2.2	65.62±2.7	58.58±2.6	**94.11±0.6**	92.47±1.2	92.13±1.8
setting 4	80.15±5.3	75.46±9.3	72.75±1.2	86.86±7.5	86.82±1.2	80.16±6.9	75.71±2.0	85.88±5.7	87.69±3.0	**91.29±4.8**
setting 2	81.47±3.5	68.46±14.7	68.81±1.3	84.45±1.2	**93.15±0.4**	73.65±14.2	60.67±0.9	78.73±10.8	84.04±4.3	**91.80±3.4**
setting 3	78.35±3.2	58.93±15.9	64.13±1.9	79.17±0.8	77.12±10.3	89.93±0.5	**94.38±0.3**	77.96±9.3	75.68±4.1	73.81±13.2
setting 5	83.52±3.5	80.83±14.5	63.82±0.6	73.70±7.3	50.91±1.1	76.52±6.7	59.28±1.0	**90.40±3.6**	**89.01±0.9**	**89.03±3.5**
setting 6	80.84±4.2	54.76±19.8	45.27±2.4	63.70±5.1	67.05±6.1	42.86±10.8	62.21±1.4	**94.36±1.0**	**93.70±0.4**	**93.86±1.0**
setting 7	79.22±3.7	42.94±2.5	57.51±1.5	55.39±2.0	50.22±4.3	43.66±8.3	62.47±0.8	**88.52±4.9**	78.56±3.2	**82.33±7.5**
VisdDA 12 modes										
setting 1	41.90±1.5	52.79±2.1	45.81±4.3	44.23±3.0	35.45±4.6	40.96±3.0	37.59±3.4	50.35±2.3	53.31±0.9	**55.05±1.6**
setting 2	41.75±1.5	50.82±1.6	45.72±8.9	40.49±4.8	36.21±5.0	36.12±4.6	31.86±5.7	48.59±1.8	53.09±1.6	**55.33±1.6**
setting 3	40.64±4.3	49.17±1.3	47.12±1.6	42.10±3.0	36.32±4.4	37.26±3.5	34.96±5.4	46.59±1.3	50.78±1.6	**52.08±1.2**
Office 31										
A - D	73.73±1.4	74.26±1.8	**77.22±0.7**	65.10±2.0	62.65±2.6	71.47±1.2	63.89±1.1	75.74±1.6	76.07±0.9	**78.20±1.3**
D - W	83.64±1.1	81.89±1.5	82.61±0.6	83.53±0.8	82.80±0.7	80.10±0.5	**87.09±0.9**	78.93±1.5	**86.32±0.6**	86.20±0.8
W - A	54.05±0.9	52.16±1.0	48.94±0.4	56.81±0.4	53.02±0.5	58.83±0.4	54.93±0.5	52.23±0.7	**60.68±0.8**	55.18±0.8
W - D	92.76±0.9	87.64±1.4	95.07±0.3	93.13±0.5	87.60±0.9	94.69±0.6	91.18±0.6	**97.04±0.9**	95.14±0.8	93.80±0.6
D - A	52.51±0.9	48.06±1.2	49.78±0.4	48.75±0.5	50.13±0.4	50.28±0.7	50.75±0.5	41.39±1.8	**54.65±0.9**	**54.95±0.9**
A - W	67.45±1.5	70.15±1.0	67.07±0.6	60.62±2.1	52.92±1.4	63.98±1.3	59.73±0.8	68.76±1.6	**73.09±1.5**	**71.90±1.2**
Office Home										
Art - Clip	37.66±0.7	36.85±0.6	33.42±1.2	31.43±1.6	27.13±1.6	31.63±5.2	29.30±6.6	37.65±0.6	37.58±0.5	**38.65±0.5**
Art - Product	49.72±0.9	49.98±0.9	39.43±3.6	38.82±2.3	35.05±2.3	35.09±3.4	32.85±3.6	48.98±0.3	**55.27±0.7**	52.18±0.4
Art - Real	58.22±1.0	53.68±0.5	51.09±2.3	50.35±1.8	46.40±2.4	51.52±4.5	45.34±11.0	57.74±0.7	**63.88±0.5**	58.75±0.7
Clip - Art	35.29±1.4	35.70±1.5	28.92±2.9	23.13±2.0	18.37±1.5	21.95±3.1	20.44±2.3	28.74±1.2	**41.15±0.6**	**40.73±0.8**
Clip - Product	**51.94±1.3**	**52.06±0.8**	39.17±7.9	39.26±2.6	34.73±1.9	39.58±2.8	39.46±2.9	34.46±2.1	51.69±0.5	**52.12±0.5**
Clip - Real	50.65±1.2	51.42±1.0	43.24±2.2	40.06±2.1	32.71±1.4	39.22±2.4	35.78±2.8	35.72±1.1	53.97±0.3	**56.63±0.5**
Product - Art	**39.59±1.6**	39.47±1.5	39.17±1.0	36.11±1.0	38.77±1.1	39.50±0.6	38.24±0.6	33.95±1.4	37.77±1.1	**39.31±1.3**
Product - Clip	32.71±0.9	**37.18±1.0**	33.82±0.5	28.38±0.7	28.40±0.6	29.72±0.5	31.76±0.8	24.89±1.0	30.86±0.8	29.25±0.9
Product - Real	**62.12±1.3**	**62.52±1.2**	**62.56±0.7**	58.09±0.5	57.58±0.6	59.33±0.6	57.11±0.8	59.22±0.9	60.48±0.6	**62.20±0.7**
Real - Product	68.30±1.0	**70.39±0.8**	**70.19±0.5**	61.72±0.8	63.40±0.9	61.51±1.0	65.45±0.6	64.47±1.5	64.79±3.6	66.49±1.1
Real - Art	**40.25±0.9**	**41.31±1.0**	39.16±0.7	33.46±1.3	31.61±1.5	36.90±0.9	36.14±0.9	36.93±1.9	39.90±1.4	39.17±1.6
Real - Clip	**42.74±1.1**	40.86±1.0	40.42±0.5	35.59±0.8	34.90±0.9	40.42±0.5	35.64±0.8	35.60±2.0	38.69±2.1	38.82±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

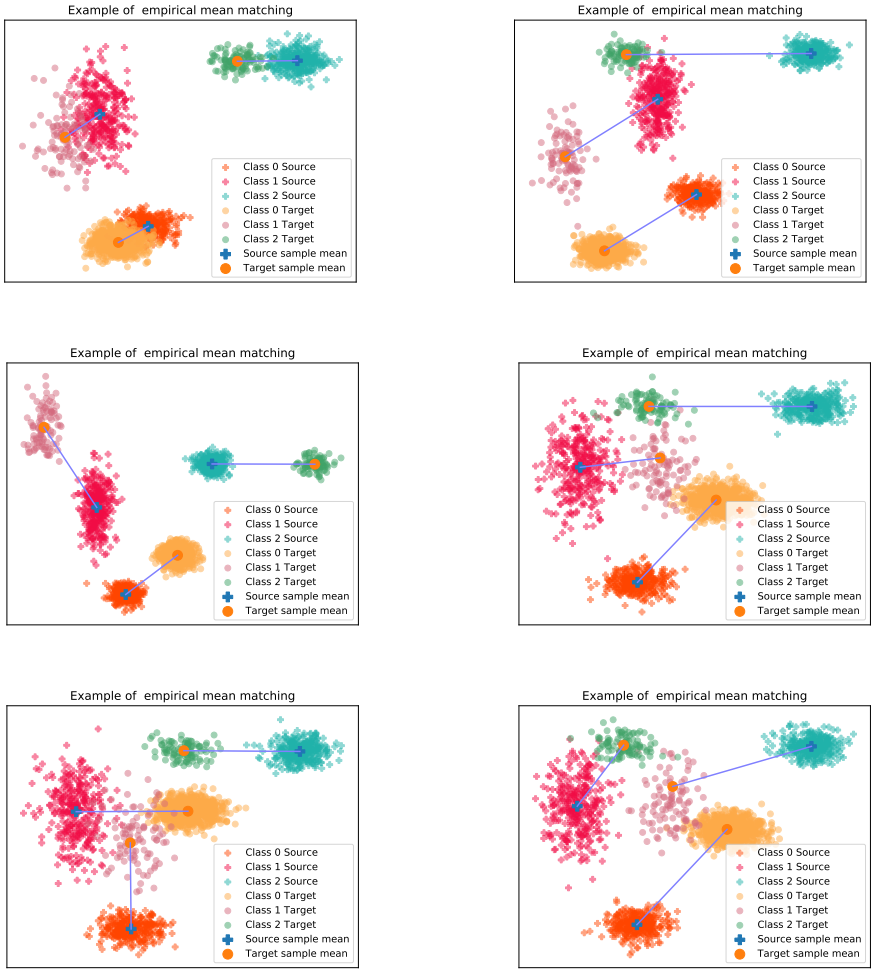

Figure 4: Example of geometrical arrangments of the source and target class-conditional distributions that allows correct and incorrect matching of classes by optimal transport of empirical means (assuming correct estimation of these means). Blue lines denote the matching. (top-left) In this setting, 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, they have been mapped through an operator such that optimal transport allows their matchings (based on their means). (middle) We have illustrated two other examples of distribution arrangments that allow class matching. (right) 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.

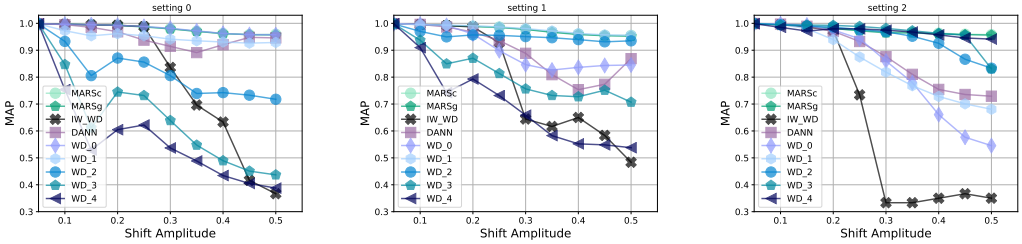

Figure 5: Performance of the compared algorithms in different label shift setting and for increasing shift between means of class-conditionals. In source domain, label distributions are uniform and shift occurs due to change only in the target domain. (left) $p_T(y=1)=0.33$, $p_T(y=2)=0.33$, $p_T(y=3)=0.34$. (middle) $p_T(y=1)=0.5$, $p_T(y=2)=0.2$, $p_T(y=3)=0.2$, (right) $p_T(y=1)=0.8$, $p_T(y=2)=0.1$, $p_T(y=3)=0.1$. For balanced problems, we note that best methods are $\text{WD}_{\beta=\{0,1\}}$, DANN and our approaches either using GMM or clustering for estimating label proportion. As expected, a too heavy reweighting yields to poor performance for $\text{WD}_{\beta=\{2,3,4\}}$. Then for a mild imbalance, $\text{WD}_{\beta=\{1,2\}}$ performs better than the other competitors while for higher imbalance, $\text{WD}_{\beta=\{3,4\}}$ works better. For all settings, our methods are competitive as they are adaptive to the imbalance through the estimation fo $p_T(y)$. The IW-WD of Combes et al. (2020) performs well until the distance between class-conditionals is too large. This is justified by theory as their estimator of the ratio $p_T(y)/p_S(y)$ is tailored for situations where class-conditionals are equal.

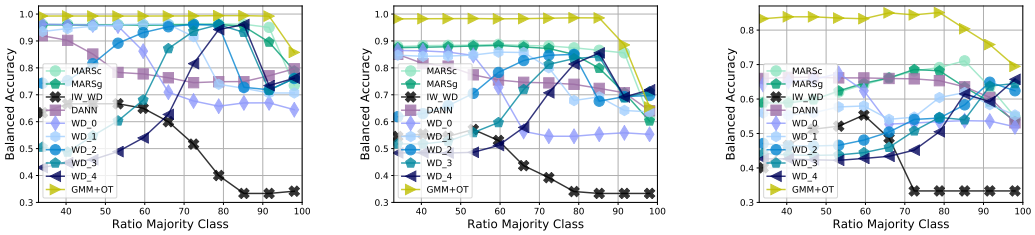

Figure 6: Performance of the compared algorithms, including **GMM+OT** for three different covariance matrices of the Gaussians composing the toy dataset with respect to the imbalance. The shift between the class-conditionals has been fixed and yields to samples similar to those presented in Figure 3. Our method is referred as **MARS**. The x-axis is given with respect to the ratio of majority class which is the class 1. (left) Low-error setting. (middle) mid-error setting. (right) high-error setting. material. We note that this toy problem can be easily solved using a GMM and a optimal transport-based label assignment. We can also remark that again as soon as the class-conditionals do not match anymore, the IW-WD of Combes et al. (2020) fails due to its inability to estimate correctly the importance weight w. .

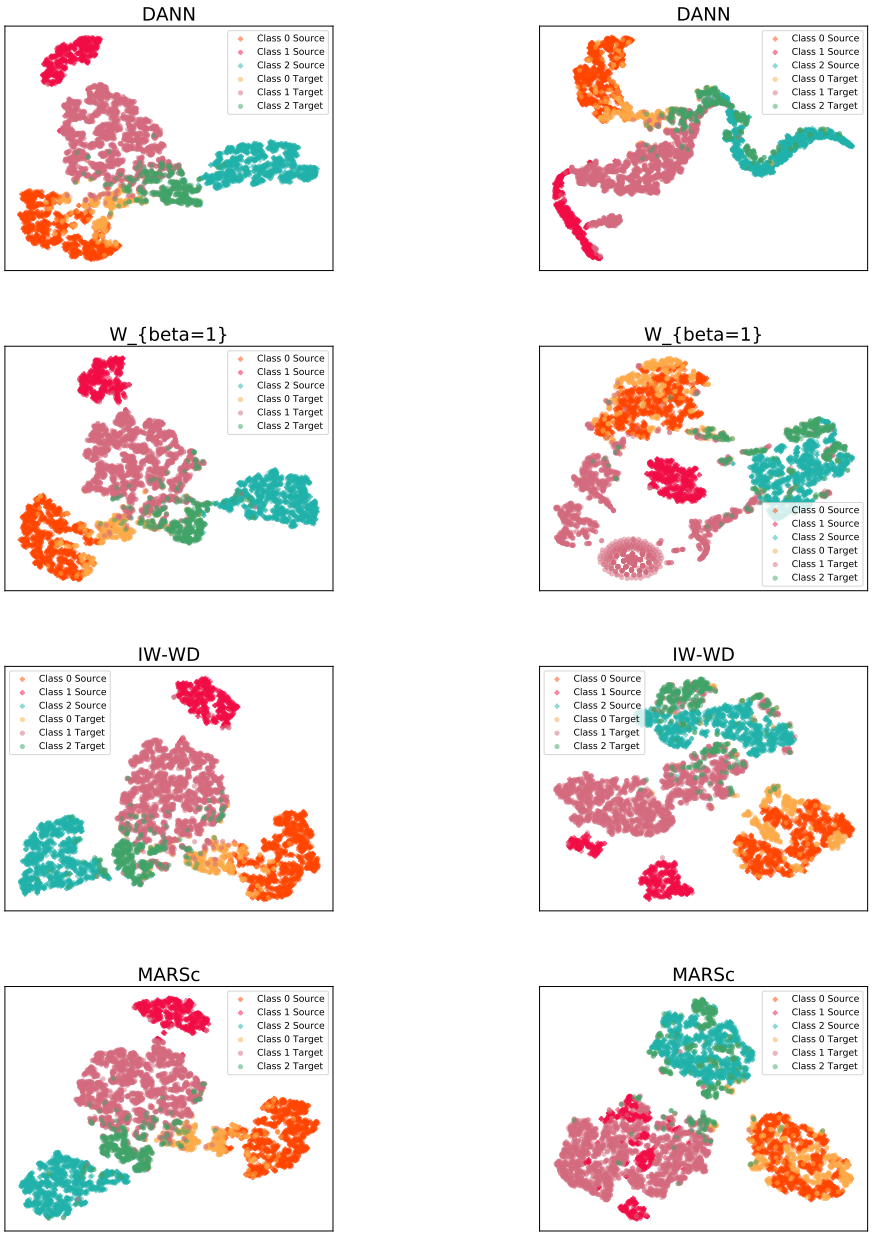

Figure 7: *t-sne* embeddings of the target sample for the VisDA-3 problem and imbalance setting 2 ($\mathbf{p}_S = [0.4, 0.2, 0.4]$ and $\mathbf{p}_T = [0.2, 0.6, 0.2]$). The columns depict the embeddings obtained (left) after training on the source data without adaptation for about 10 iterations, which 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) $\text{WD}_{\beta=1}$, (third-row), IW-WD (last row) MARSc. From the right column, we note how DANN and $\text{WD}_{\beta=1}$ struggles 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.