Optimal Transport for Multi-source Domain Adaptation under Target Shift

Ievgen Redko, Nicolas Courty, Rémi Flamary, Devis Tuia

Optimal Transport for Multi-source Domain Adaptation under Target Shift

Ievgen Redko
Univ Lyon, UJM-Saint-Etienne, CNRS
Institut d Optique Graduate School
Laboratoire Hubert Curien UMR 5516
F-42023, Saint-Etienne, France

Nicolas Courty
University of Bretagne Sud, CNRS and INRIA
IRISA, UMR 6074
F-56000, Vannes, France

Rémi Flamary
University of Côte d'Azur, OCA
Laboratoire Lagrange, CNRS UMR 7293
F-06108, Nice, France

Devis Tuia
Laboratory of Geoscience and Remote Sensing
Wageningen University and Research
6700, Wageningen, Netherlands

Abstract

In this paper, we tackle the problem of reducing discrepancies between multiple domains, i.e. *multi-source domain adaptation*, and consider it under the *target shift* assumption: in all domains we aim to solve a classification problem with the same output classes, but with different labels proportions. This problem, generally ignored in the vast majority of domain adaptation papers, is nevertheless critical in real-world applications, and we theoretically show its impact on the success of the adaptation. Our proposed method is based on optimal transport, a theory that has been successfully used to tackle adaptation problems in machine learning. The introduced approach, Joint Class Proportion and Optimal Transport (JCPOT), performs multi-source adaptation and target shift correction simultaneously by learning the class probabilities of the unlabeled target sample and the coupling allowing to align two (or more) probability distributions. Experiments on both synthetic and real-world data (satellite image pixel classification) task show the superiority of the proposed method over the state-of-the-art.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

1 INTRODUCTION

In many real-world applications, it is desirable to use models trained on largely annotated data sets (or *source domains*) to label a newly collected, and therefore unlabeled data set (or *target domain*). However, differences in the probability distributions between them hinder the success of the direct application of learned models to the latter. To overcome this problem, recent machine learning research has devised a family of techniques, called domain adaptation (DA), that deals with situations where source and target samples follow different probability distributions (Quiñonero-Candela *et al.* , 2009; Patel *et al.* , 2015). The inequality between the joint distributions can be characterized in a variety of ways depending on the assumptions made about the conditional and marginal distributions. Among those, arguably the most studied scenario called covariate shift (or sample selection bias) considers the situation where the inequality between probability density functions (pdfs) is due to the change in the marginal distributions (Zadrozny *et al.* , 2003; Bickel *et al.* , 2007; Huang *et al.* , 2007; Liu & Ziebart, 2014; Wen *et al.* , 2014; Fernando *et al.* , 2013; Courty *et al.* , 2017a).

Despite the large number of methods proposed in the literature to solve the DA problem under the covariate shift assumption, very few considered the (widely occurring) situation where the changes in the joint distribution is caused by a shift in the distribution of the outputs, a setting that has been referred to as *target shift*. In practice, the target shift assumption implies that a change in the class proportions across domains is at the base of the shift: such a situation is also known as choice-based or endogenous stratified

sampling in econometrics (Manski & Lerman, 1977) or as prior probability shift (Storkey, 2009). In the classification context, target shift was first introduced in (Japkowicz & Stephen, 2002) and referred to as the class imbalance problem. In order to solve it, several approaches were proposed. In (Lin *et al.* , 2002), authors assumed that the shift in the target distribution was known a priori, while in (Yu & Zhou, 2008), partial knowledge of the target shift was supposed to be available. In both cases, the assumption of prior knowledge about the class proportions in the target domain seems quite restrictive. More recent methods that avoid making this kind of assumptions are (Chan & Ng, 2005; Zhang *et al.* , 2013). In the former, authors used a variation of the Expectation Maximization algorithm, which relies on a computationally expensive estimation of the conditional distribution. In the latter, authors estimate the proportions in both domains directly from observations. Their approach, however, also relies on computationally expensive optimization over kernel embeddings of probability functions. This last line of work has been extended in (Zhang *et al.* , 2015) to the multi-domain setting, and as such constitutes a relevant baseline for our work. The algorithm proposed by the authors produces a set of hypotheses with one hypothesis per domain that can be combined using the theoretical study on multi-source domain adaptation presented in (Mansour *et al.* , 2009b). Despite the rather small corpus of works in the literature dealing with the subject, target shift often occurs in practice, especially in applications dealing with anomaly/novelty detection (Blanchard *et al.* , 2010; Scott *et al.* , 2013; Sanderson & Scott, 2014), or in tasks where spatially located training sets are used to classify wider areas, as in remote sensing image classification (Tuia *et al.* , 2015; Zhang *et al.* , 2015).

In this paper, we propose a new algorithm for correcting the target shift based on optimal transport (OT). OT theory is a branch of mathematics initially introduced by Gaspard Monge for the task of resource allocation (Monge, 1781). Originally, OT tackled a problem of aligning two probability measures in a way that minimizes the cost of moving a unit mass between them, while preserving the original marginals. The recent appearance of efficient formulations of OT (Cuturi, 2013) has allowed its application in DA, as OT allows to learn explicitly the transformation of a given source pdf into the pdf of the target sample. In this work, we build upon a recent work on DA (Courty *et al.* , 2017b), where authors successfully casted the DA problem as an OT problem, and extend it to deal with the target shift setting. Our motivation to propose new specific algorithms for target shift stems from the fact that many popular DA algorithms designed to tackle covariate shift cannot handle the target shift equally well. This is illustrated in Figure 1, where we show that the DA method based on OT mentioned above fails to restrict the transportation of mass across instances of different classes when the class proportions of source and target domains differ. However, as we show in the following sections, our *Joint Class Proportion and Optimal Transport (JCPOT)* model manages to do it correctly. Furthermore and contrary to the original contribution, we also consider the much more challenging case of multi-source domain adaptation, where more than one source domains with changing distributions of outputs are used for learning. To the best of our knowledge, this is the first multi-source DA algorithm that efficiently leverages the target shift and shows an increasing performance with the increasing number of source domains considered.

The rest of the paper is organized as follows: in Section 2, we present regularized OT and its application to DA. Section 3 details the target shift problem and provides a generalization bound for this learning scenario and a proof that minimizing the Wasserstein distance between two distributions with class imbalance leads to the optimal solution. In Section 4, we present the proposed JCPOT method for unsupervised DA when no labels are used for adaptation. In Section 5, we provide comparisons to state-of-art methods for synthetic data in the multi-source adaptation scenario and we report results for a real life case study performed in remote sensing pixel classification.

2 OPTIMAL TRANSPORT

In this section we introduce key concepts of optimal transport and some important results used in the following sections.

2.1 Basics and Notations

OT can be seen as the search for a plan that moves (transports) a probability measure μ_1 onto another measure μ_2 with a minimum cost measured by some function c. In our case, we use the squared Euclidean distance L_2^2, but other domain-specific measures, more suited to the problem at hand, could be used instead. In the relaxed formulation of Kantorovitch (Kantorovich, 1942), OT seeks for an optimal coupling that can be seen as a joint probability distribution between μ_1 and μ_2. In other words, if we define $\Pi(\mu_1, \mu_2)$ as the space of probability distributions over $\mathbb{R}^2$ with marginals μ_1 and μ_2, the optimal transport is the coupling $\gamma \in \Pi(\mu_1, \mu_2)$, which minimizes the following quantity:

$$W_c(\mu_1, \mu_2) = \inf_{\gamma \in \Pi(\mu_1, \mu_2)} \int_{\mathbb{R}^2} c(\mathbf{x}_1, \mathbf{x}_2) d\gamma(\mathbf{x}_1, \mathbf{x}_2),$$

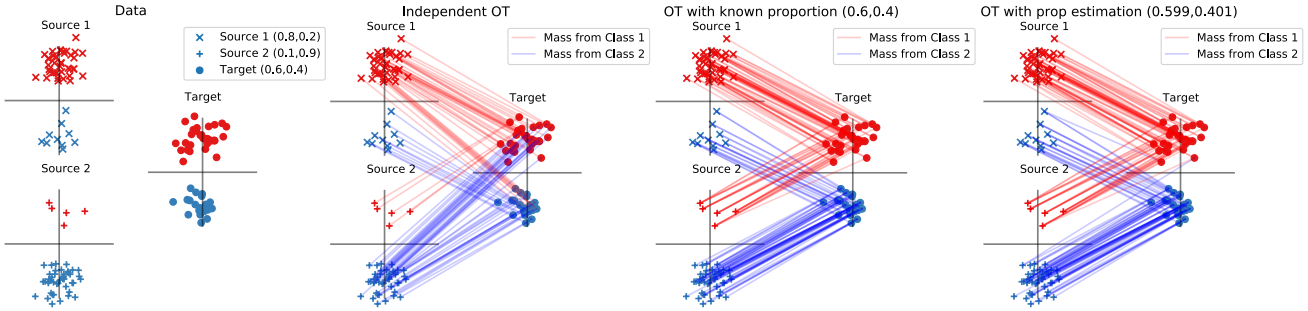

Figure 1: Illustration of the importance of proportion estimation for target shift: (a) the data of 2 source and 1 target domains with different class proportions is visualized; (b) DA method based on OT (Courty *et al.* , 2014) transports instances across different classes due to class proportions imbalance; (c) the transportation obtained when the true class proportions are used to reweigh instances; (d) transportation obtained with JCPOT that is nearly identical to the one obtained with an a priori knowledge about the class proportions.

where $c(\mathbf{x}_1, \mathbf{x}_2)$ is the cost of moving $\mathbf{x}_1$ to $\mathbf{x}_2$ (drawn from distributions μ_1 and μ_2, respectively). In the discrete versions of the problem, *i.e.* when μ_1 and μ_2 are defined as empirical measures based on vectors in $\mathbb{R}^d$, $\Pi(\mu_1, \mu_2)$ denotes the polytope of matrices γ such that $\gamma \mathbf{1} = \mu_1, \gamma^T \mathbf{1} = \mu_2$ and the previous equation reads:

$$W_{\mathbf{C}}(\mu_1, \mu_2) = \min_{\gamma \in \Pi(\mu_1, \mu_2)} \langle \gamma, \mathbf{C} \rangle_F, \tag{1}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot product, $\mathbf{C} \geq 0$ is a cost matrix $\in \mathbb{R}^{n_1 \times n_2}$, representing the pairwise costs of transporting bin i to bin j, and γ is a joint distribution given by a matrix of size $n_1 \times n_2$, with marginals defined as μ_1 and μ_2. Solving equation (1) is a simple linear programming problem with equality constraints, but its dimensions scale quadratically with the size of the sample. Alternatively, one can consider a regularized version of the problem, which has the extra benefit of being faster to solve.

2.2 Entropic Regularization

In (Cuturi, 2013), the authors added a regularization term to γ that controls the smoothness of the coupling through the entropy of γ. The entropy regularized version of the discrete OT reads:

$$W_{\mathbf{C},\epsilon}(\mu_1, \mu_2) = \min_{\gamma \in \Pi(\mu_1, \mu_2)} \langle \gamma, \mathbf{C} \rangle_F - \epsilon h(\gamma), \tag{2}$$

where $h(\gamma) = -\sum_{ij} \gamma_{ij} (\log \gamma_{ij} - 1)$ is the entropy of γ. Similarly, denoting the Kullback-Leibler divergence (KL) as $\mathrm{KL}(\gamma|\rho) = \sum_{ij} \gamma_{ij} (\log \frac{\gamma_{ij}}{\rho_{ij}} - 1) = \langle \gamma, \log \frac{\gamma}{\rho} - \mathbf{1} \rangle_F$, one can establish the following link between OT and Bregman projections.

Proposition 1. (Benamou *et al.* , 2015, Eq. (6,7)). *For $\zeta = \exp\left(-\frac{C}{\epsilon}\right)$, the minimizer $\gamma^\star$ of (2) is the solution of the following Bregman projection*

$$\gamma^\star = \operatorname*{arg\,min}_{\gamma \in \Pi(\mu_1, \mu_2)} \mathrm{KL}(\gamma|\zeta).$$

For an undefined μ_2, $\gamma^\star$ is solely constrained by the marginal μ_1 and is the solution of the following closed-form projection:

$$\gamma^\star = \mathrm{diag}\left(\frac{\mu_1}{\zeta \mathbf{1}}\right) \zeta, \tag{3}$$

where the division has to be understood component-wise.

As it follows from this proposition, the entropy regularized version of OT can be solved with a simple algorithm based on successive projections over the two marginal constraints and admits a closed form solution. We refer the reader to (Benamou *et al.* , 2015) for more details on this subject.

2.3 Application to Domain Adaptation

A solution to the two domains adaptation problem based on OT has been proposed in (Courty *et al.* , 2014). It consists in estimating a transformation of the source domain sample that minimizes their average displacement w.r.t. target sample, *i.e.* an optimal transport solution between the discrete distributions of the two domains. The success of the proposed algorithm is due to an important advantage offered by OT metric over other distances used in DA (e.g. MMD): it preserves the topology of the data and admits a rather efficient estimation. The authors further added a regularization term used to encourage instances from the target sample to be transported to instances of the source sample of the same class, therefore promoting group sparsity in γ thanks to the $\|\cdot\|_q^p$ norm with $q = 1$ and $p = \frac{1}{2}$ (Courty *et al.* , 2014) or $q = 2$ and $p = 1$ (Courty *et al.* , 2017b).

3 DOMAIN ADAPTATION UNDER THE TARGET SHIFT

In this section, we formalize the target shift problem and provide a generalization bound that shows the key factors that have an impact when learning under it.

To this end, let us consider a binary classification problem with K source domains, each being represented by a sample of size $n^{(k)}, k = 1, \dots, K$, drawn from a probability distribution $P_S^k = (1 - \pi_S^k)P_0 + \pi_S^k P_1$. Here $0 < \pi_S^k < 1$ and P_0, P_1 are marginal distributions of the source data given the class labels 0 and 1, respectively with $P_0 \neq P_1$. We also possess a target sample of size n drawn from a probability distribution $P_T = (1 - \pi_T)P_0 + \pi_T P_1$ such that $\exists\ j \in [1, \dots, N] : \pi_S^j \neq \pi_T$. This last condition is a characterization of target shift used in previous theoretical works on the subject (Scott *et al.* , 2013).

Following (Ben-David *et al.* , 2010), we define a domain as a pair consisting of a distribution P_D on some space of inputs Ω and a labeling function $f_D : \Omega \to [0, 1]$. A hypothesis class $\mathcal{H}$ is a set of functions so that $\forall h \in \mathcal{H}, h : \Omega \to \{0, 1\}$. Given a convex loss-function l, the true risk with respect to the distribution P_D, for a labeling function f_D (which can also be a hypothesis) and a hypothesis h is defined as

$$\epsilon_D(h, f_D) = \mathbb{E}_{x \sim P_D} \left[l(h(x), f_D(x)) \right]. \tag{4}$$

In the multi-source case, when the source and target error functions are defined w.r.t. h and $f_S^{(k)}$ or f_T, we use the shorthand $\epsilon_S^{(k)}(h, f_S^{(k)}) = \epsilon_S^{(k)}(h)$ and $\epsilon_T(h, f_T) = \epsilon_T(h)$. The ultimate goal of multi-source DA then is to learn a hypothesis h on K source domains that has the best possible performance in the target one.

To this end, we define the combined error of source domains as a weighted sum of source domains error functions:

$$\epsilon_S^{\boldsymbol{\alpha}} = \sum_{k=1}^{K} \alpha_k \epsilon_S^{(k)}, \sum_{k=1}^{K} \alpha_k = 1, \alpha_k \in [0, 1]\ \forall k \in [1, \dots, K].$$

We further denote by $f_S^{\boldsymbol{\alpha}}$ the labeling function associated to the distribution mixture $P_S^{\boldsymbol{\alpha}} = \sum_{k=1}^{N} \alpha_j P_S^j$. In multi-source scenario, the combined error is minimized in order to produce a hypothesis that is used on the target domain. Here different weights α_k can be seen as measures reflecting the proximity of the corresponding source domain distribution to the target one.

For the target shift setup introduced above, we can prove the following proposition.

Proposition 2. *Let $\mathcal{H}$ denote the hypothesis space of predictors $h : \Omega \to \{0, 1\}$ and l be a convex loss function. Let $disc_l(P_S, P_T) = \max_{h,h' \in \mathcal{H}} |\epsilon_S(l(h, h')) - \epsilon_T(l(h, h'))|$ be the discrepancy distance (Mansour* et al. *, 2009a) between two probability distributions P_S and P_T. Then, for any fixed $\boldsymbol{\alpha}$ and for any $h \in \mathcal{H}$ the following holds:*

$$\epsilon_T(h) \leq \epsilon_S^{\boldsymbol{\alpha}}(h) + |\pi_T - \sum_{j=1}^{N} \alpha_j \pi_S^j| disc_l(P_0, P_1) + \lambda,$$

where $\lambda = \min_{h \in \mathcal{H}} \epsilon_S^{\boldsymbol{\alpha}}(h) + \epsilon_T(h)$ represents the joint error between the combined source error and the target one[1].

The second term in the bound can be minimized for any α_k when $\pi_T = \pi_S^k,\ \forall k$. This can be achieved by using a proper reweighting of the class distributions in the source domains, but requires to have access to the target proportion which is assumed to be unknown. In the next section, we propose to estimate optimal proportions by minimizing the sum of the Wasserstein distances between all reweighted sources and the target distribution. In order to justify this idea, we prove below that the minimization of the Wasserstein distance between a weighted source distribution and a target distribution yields the optimal proportion estimation. To proceed, let us consider the multi-class problem with C classes, where the target distribution is defined as

$$P_T = \sum_{i=1}^{C} \pi_i^T P_i,$$

with P_i being a distribution of class $i \in \{1, \dots, C\}$. As before, the source distribution with weighted classes can be then defined as

$$P_S^{\pi} = \sum_i \pi_i P_i,$$

where $\pi \in \Delta_C$ are coefficients lying in the probability simplex $\Delta_C \overset{\text{def}}{=} \{\alpha \in \mathbb{R}_+^C : \sum_{i=1}^{C} \alpha_i = 1\}$ that reweigh the corresponding classes.

As the proportions of classes in the target distribution are unknown, our goal is to reweigh source classes distributions by solving the following optimization problem:

$$\pi^\star = \arg\min_{\pi \in \Delta_C}\ W(P_S^{\pi}, P_T). \tag{5}$$

We can now state the following proposition.

Proposition 3. *Assume that $\forall i, \nexists \alpha \in \{\Delta_C | \alpha_i = 0, P_i = \sum_j \alpha_j P_j\}$. Then, for any distribution P_T, the unique solution π^* minimizing (5) is given by π^T.*

Note that this result extends straightforwardly to the multi-source case where the optimal solution of minimizing the sum of the Wasserstein distance for all

[1] Proofs of several theoretical results of this paper can be found in the Supplementary material.

source distributions is the target domain proportions. As real distributions are accessible only through available finite samples, in practice, we propose to minimize the Wasserstein distance between the empirical target distribution $\hat{P}_T$ and the empirical source distributions $\hat{P}_S^k$. The convergence of the exact solution of this problem with empirical measures can be characterized using the concentration inequalities established for Wasserstein distance in (Bobkov & Ledoux, 2016; Fournier & Guillin, 2015) where the rate of convergence is inversely proportional to the number of available instances in source domains and consequently, to the number of source domains.

4 JOINT CLASS PROPORTION AND OPTIMAL TRANSPORT (JCPOT)

In this section, we introduce the proposed JCPOT method, that aims at finding the optimal transportation plan and estimating class proportions jointly. The main underlying idea behind JCPOT is to reweigh instances in the source domains in order to compensate for the discrepancy between the source and target domains class proportions.

4.1 Data and Class-Based Weighting

We assume to have access to several data sets corresponding to K different domains $\mathbf{X}^{(k)}$, $k = 1, \dots, K$. These domains are formed by $n^{(k)}$ instances $\mathbf{x}_i^{(k)} \in \mathbb{R}^d$ with each instance being associated with one of the C classes of interest. In the following, we use the superscript (k) when referring to quantities in one of the source domains (e.g $\mu^{(k)}$) and the equivalent without superscript when referring to the same quantity in the (single) target domain (e.g. μ). Let $y_i^{(k)}$ be the corresponding class, *i.e.* $y_i^{(k)} \in \{1, \dots, C\}$. We are also given a target domain $\mathbf{X}$, populated by n instances defined in $\mathbb{R}^d$. The goal of unsupervised multi-source adaptation is to recover the classes y_i of the target domain samples, which are all unknown.

JCPOT works under the target shift assumption presented in Section 3. For every source domain, we assume that its data points follow a probability distribution function or probability measure $\mu^{(k)}$ ($\int \mu^{(k)} = 1$). In real-world situations, $\mu^{(k)}$ is only accessible through the instances $\mathbf{x}_i^{(k)}$ that we can use to define a distribution $\mu^{(k)} = \sum_{i=1}^{n^{(k)}} m_i^{(k)} \delta_{\mathbf{x}_i^{(k)}}$, where $\delta_{\mathbf{x}_i^{(k)}}$ are Dirac measures located at $\mathbf{x}_i^{(k)}$, and $m_i^{(k)}$ is an associated probability mass. By denoting the corresponding vector of mass as $\mathbf{m}^{(k)}$, *i.e.* $\mathbf{m}^{(k)} = [m_i^{(k)}]_{i=\{1,\dots,n^{(k)}\}}$, and $\delta_{\mathbf{X}^{(k)}}$ the corresponding vector of Dirac measures, one can write $\mu^{(k)} = (\mathbf{m}^{(k)})^T \delta_{\mathbf{X}^{(k)}}$. Note that when the data set is a collection of independent data points, the weights of all instances in the sample are usually set to be equal. In this work, however, we use different weights for each class of the source domain so that we can adapt the proportions of classes w.r.t. the target domain. To this end, we note that the measures can be decomposed among the C classes as $\mu^{(k)} = \sum_{c=1}^{C} \mu_c^{(k)}$. We denote by $h_c^{(k)} = \int \mu_c^{(k)}$ the proportion of class c in $\mathbf{X}^{(k)}$. By construction, we have $h_c^{(k)} = \sum_{i=1}^{n^{(k)}} \delta(y_i^{(k)} = c) m_i^{(k)}$.

Since we chose to have equal weights in the classes, we define two linear operators $\mathbf{D}_1^{(k)} \in \mathbb{R}^{C \times n^{(k)}}$ and $\mathbf{D}_2^{(k)} \in \mathbb{R}^{n^{(k)} \times C}$ that allow to express the transformation from the vector of mass $\mathbf{m}^{(k)}$ to the class proportions $\mathbf{h}^{(k)}$ and back:

$$\mathbf{D}_1^{(k)}(c, i) = \begin{cases} 1 & \text{if } y_i^{(k)} = c, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mathbf{D}_2^{(k)}(i, c) = \begin{cases} \frac{1}{\#\{y_i^{(k)} = c\}_{i=\{1,\dots,n^{(k)}\}}} & \text{if } y_i^{(k)} = c, \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbf{D}_1^{(k)}$ allows to retrieve the class proportions with $\mathbf{h}^{(k)} = \mathbf{D}_1^{(k)} \mathbf{m}^{(k)}$ and $\mathbf{D}_2^{(k)}$ returns weights for all instances for a given vector of class proportions with $\mathbf{m}^{(k)} = \mathbf{D}_2^{(k)} \mathbf{h}^{(k)}$, where the masses are distributed equiproportionnally among all the data points associated to one class. For example, for a source domain with 5 elements from which first 3 belong to class 1 and the other to class 2,

$$\mathbf{D}_1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix},$$

$\mathbf{m} = [\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}]^T$ so that $\mathbf{h} = \mathbf{D}_1 \mathbf{m} = [\frac{3}{5}, \frac{2}{5}]$. These are the class proportions of this source domain. On the other hand,

$$\mathbf{D}_2 = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}^T,$$

so that $\mathbf{D}_2 \mathbf{h} = \mathbf{m}$ and $\mathbf{D}_1 \mathbf{D}_2 = \boldsymbol{I}$.

4.2 Multi-Source Domain Adaptation with JCPOT

As illustrated in Section 1, having matching proportions between the source and the target domains helps in finding better couplings, and, as shown in Section 3 it also enhances the adaptation results.

To this end, we propose to estimate the class proportions in the target domain by solving a constrained Wasserstein barycenter problem (Benamou *et al.* , 2015)

for which we use the operators defined above to match the proportions to the uniformly weighted target distribution. The corresponding optimization problem can be written as follows:

$$\underset{\mathbf{h}\in\Delta_C}{\arg\min} \quad \sum_{k=1}^{K} \lambda_k W_{\epsilon,C^{(k)}}\left((\mathbf{D}_2^{(k)}\mathbf{h})^T \delta_{\mathbf{X}^{(k)}}, \mu\right), \qquad (6)$$

where regularized Wasserstein distances are defined as

$$W_{\epsilon,C^{(k)}}(\mu^{(k)},\mu) \stackrel{\text{def}}{=} \min_{\gamma^{(k)}\in\Pi(\mu^{(k)},\mu)} \mathrm{KL}(\gamma^{(k)}|\zeta^{(k)}),$$

provided that $\zeta^{(k)} = \exp\left(-\frac{C^{(k)}}{\epsilon}\right)$ with λ_k being convex coefficients ($\sum_k \lambda_k = 1$) accounting for the relative importance of each domain. Here, we define the set $\Gamma = \{\gamma^{(k)}\}_{k=1\dots K} \in (\mathbb{R}^{n^{(k)}\times n})^K$ as the set of couplings between each source and the target domains. This problems leads to K marginal constraints $\gamma_k^T \mathbf{1}_n = \mathbf{1}_n/n$ w.r.t. the uniform target distribution, and K marginal constraints $\mathbf{D}_1^{(k)}\gamma_k \mathbf{1}_n = \mathbf{h}$ related to the unknown proportions $\mathbf{h}$.

Optimizing for the first K marginal constraints can be done independently for each k by solving the problem expressed in Equation 3. On the contrary, the remaining K constraints require to solve the proposed optimization problem for Γ and $\mathbf{h}$, simultaneously. To do so, we formulate the problem as a Bregman projection with prescribed row sum ($\forall k \;\; \mathbf{D}_1^{(k)}\gamma^{(k)}\mathbf{1}_n = \mathbf{h}$), i.e.,

$$\begin{aligned} \mathbf{h}^\star = \underset{\mathbf{h}\in\Delta_C,\Gamma}{\arg\min} & \sum_{k=1}^{K} \lambda_k \,\mathrm{KL}(\gamma^{(k)}|\zeta^{(k)}) \\ \text{s.t.} \quad & \forall k \;\; \mathbf{D}_1^{(k)}\gamma^{(k)}\mathbf{1}_n = \mathbf{h}. \end{aligned} \qquad (7)$$

This problem admits a closed form solution that we establish in the following result.

Proposition 4. *The solution of the projection defined in Equation 7 is given by:*

$$\forall k, \gamma_k = \mathrm{diag}\left(\frac{\mathbf{D}_2^{(k)}\mathbf{h}}{\zeta^{(k)}\mathbf{1}_n}\right)\zeta^{(k)}, \mathbf{h} = \Pi_{k=1}^{K}(\mathbf{D}_1^{(k)}(\zeta^{(k)}\mathbf{1}_n))^{\lambda_k}.$$

The initial problem can now be solved through an Iterative Bregman projections scheme summarized in Algorithm 1. Note that the updates for coupling matrix in lines 5 and 7 of the algorithm can be computed in parallel for each domain.

4.3 Classification in the Target Domain

When both the class proportions and the corresponding coupling matrices are obtained, we need to adapt source and target samples and classify unlabeled target instances. Below, we provide two possible ways that can be used to perform these tasks.

Algorithm 1 Joint Class Proportion and Optimal Transport (JCPOT)

1: **Input:** ϵ, maxIter, $\forall k$ ($\mathbf{C}^{(k)}$ and $\lambda^{(k)}$)
2: $cpt \leftarrow 0$,
3: $err \leftarrow \infty$
4: **for all** $k = 1, \dots, K$ **do**
5: $\quad \zeta^{(k)} \leftarrow \exp\left(-\frac{\mathbf{C}^{(k)}}{\epsilon}\right), \quad \forall k$
6: **while** $cpt <$ maxIter **and** $err >$ threshold **do**
7: $\quad$ **for all** $k = 1, \dots, K$ **do**
8: $\quad\quad \zeta^{(k)} \leftarrow \mathrm{diag}(\frac{\mathbf{m}}{\zeta^{(k)}\mathbf{1}})\zeta^{(k)}, \quad \forall k$
9: $\quad \mathbf{h}^{(cpt)} \leftarrow \exp\left(\sum_{k=1}^{K} \lambda^{(k)} \log\left(\mathbf{D}_1^{(k)}\zeta^{(k)}\mathbf{1}\right)\right)$
10: $\quad$ **for all** $k = 1, \dots, K$ **do**
11: $\quad\quad \zeta^{(k)} \leftarrow \zeta^{(k)}\,\mathrm{diag}(\frac{\mathbf{D}_2^{(k)}\mathbf{h}}{\zeta^{(k)}\mathbf{1}}), \quad \forall k$
12: $\quad err \leftarrow ||\mathbf{h}^{(cpt)} - \mathbf{h}^{(cpt-1)}||_2$,
13: $\quad$ cpt $\leftarrow$ cpt + 1
14: **return** $\mathbf{h}$, $\forall k$ $\zeta^{(k)}$

Barycentric mapping In (Courty *et al.* , 2017b), the authors proposed to use the OT matrices to estimate the position of each source instance as the barycenter of the target instances, weighted by the mass from the source sample. This approach extends to multi-source setting and naturally provides a target-aligned position for each point from each source domain. These adapted source samples can then be used to learn a classifier and apply it directly on the target sample. In the sequel, we denote the variations of JCPOT that use the barycenter mapping as `JCPOT-PT`. For this approach, (Courty *et al.* , 2017b) noted that too much regularization has a shrinkage effect on the new positions, since the mass spreads to all target points in this configuration. Also, it requires the estimation of a target classifier, trained on the transported source samples, to provide predictions for the target sample.

Label propagation We propose alternatively to use the OT matrices to perform label propagation onto the target sample. Since we have access to the labels in the source domains and since the OT matrices provide the transportation of mass, we can measure for each target instance the proportion of mass coming from every class. Therefore, we propose to estimate the label proportions for the target sample with $\mathbf{L} = \sum_{k=1}^{K} \lambda_k \mathbf{D}_1^{(k)}\gamma^{(k)}$ where the component $l_{c,i}$ in $\mathbf{L}$ contains the probability estimate of target sample i to belong to class c. Note that this label propagation technique can be seen as boosting, since the expression of $\mathbf{L}$ corresponds to a linear combination of weak classifiers from each source domain. To the best of our knowledge, this is the first time such type of approach is proposed in DA. In the following, we denote it by `JCPOT-LP` where LP stands for label propagation.

5 EXPERIMENTAL RESULTS

In this section, we present the results of our algorithm for both synthetic and real-world data from the task of remote sensing classification.

Baseline and state-of-the-art methods We compare the proposed JCPOT algorithm to three other methods designed to tackle target shift, namely `betaEM`, the variation of the EM algorithm proposed in (Chan & Ng, 2005) `betaKMM`, an algorithm based on the kernel embeddings proposed in (Zhang *et al.* , 2013)[2] , and `MDA Causal`, a multi-domain adaptation strategy with a causal view (Zhang *et al.* , 2015) [3]. Note that despite the existence of several deep learning methods that deal with covariate shift, e.g. (Ganin & Lempitsky, 2015), to the best of our knowledge, none of them tackle specifically the problem of target shift.

As explained in 4.3, our algorithms can obtain the target labels in two different ways, either based on label propagation (`JCPOT-LP`) or based on transporting points and applying a standard classification algorithm after transformation (`JCPOT-PT`). Furthermore, we also consider two additional DA algorithms that use OT (Courty *et al.* , 2014): `OTDA-LP` and `OTDA-PT` that align the domains based on OT but without considering the discrepancies in class proportions.

5.1 Synthetic Data

Data generation In the multi-source setup, we sample 20 source domains, each consisting of 500 instances and a target domain with 400 instances. We vary the source domains' class proportions randomly while keeping the target ones equal to $[0.2; 0.8]$. For more details on the generative process and some additional empirical results regarding the sensitivity of `JCPOT` to hyperparameters tuning and the running times comparison, we refer the reader to the Supplementary material.

Results Table 2 gives average performances over five runs for each domain adaptation task when the number of source domains varies from 2 to 20. As `betaEM`, `betaKMM` and `OTDA` are not designed to work in the multi-source scenario, we fusion the data from all source domains and use it as a single source domain. From the results, we can see that the algorithm with label propagation (`JCPOT-LP`) provides the best results and outperforms other state-of-the-art DA methods, except for 20 source domains, where `MDA Causal` slightly surpasses our method. It is worth noting that, all methods addressing specifically the target shift problem perform better that the `OTDA` method designed for covariate shift. This result justifies our claim about the necessity of specially designed algorithms that take into account the shifting class proportions in DA.

	Number of source domains						
	2	5	8	11	14	17	20
JCPOT	0.039	0.045	**0.027**	**0.029**	**0.035**	**0.033**	**0.034**
Scott *et al.* (2013)	**0.01**	**0.044**	0.06	0.10	0.22	**0.033**	0.14

Table 1: Accuracy of proportions estimation for simulated data.

On the other hand, we also evaluate the accuracy of proportion estimation of our algorithm and compare it with the results obtained by the algorithm proposed in Scott *et al.* (2013)[4]. As this latter was designed to deal with binary classification, we restrict ourselves to the comparison on simulated data only and present the deviation of the estimated proportions from their true value in terms of the L^1 distance in Table 1. From this Table, we can see that our method gives comparable or better results in most of the cases. Furthermore, our algorithm provides coupling matrices that allow to align the source and target domains samples and to directly classify target instanes using the label propogation described above.

5.2 Real-World Data From Remote Sensing Application

Data set We consider the task of classifying superpixels from satellite images at very high resolution into a set of land cover/land use classes (Tuia *et al.* , 2015). We use the 'Zurich Summer' data set[5], composed of 20 images issued from a large image acquired by the QuickBird satellite over the city of Zurich, Switzerland in August 2002 where the features are extracted as described in (Tuia *et al.* , 2018, Section 3.B). For this data set, we consider a multi-class classification task corresponding to the classes Roads, Buildings, Trees and Grass shared by all images. The number of superpixels per class is imbalanced and varies across images: thus it represents a real target shift problem. We consider 18 out of the 20 images, since two images exhibit a very scarce ground truth, making a reliable estimation of the true classes proportions difficult. We use each image as the target domain (average class proportions with standard deviation are $[0.25 \pm 0.07,\ 0.4 \pm 0.13,\ 0.22 \pm 0.11,\ 0.13 \pm 0.11]$) while considering remaining 17 images as source domains.

[2]code available at `http://people.tuebingen.mpg.de/kzhang/Code-TarS.zip`.

[3]code available at `https://mgong2.github.io/papers/MDAC.zip`

[4]code available at `http://web.eecs.umich.edu/~cscott/code/mpe_v2.zip`

[5]https://sites.google.com/site/michelevolpiresearch/data/zurich-dataset

# of source domains	Average class proportions	# of source instances	No adaptation	OTDA PT	OTDA LP	beta EM	beta KMM	MDA Causal	JCPOT PT	JCPOT LP	Target only
Multi-source simulated data											
2	[0.64 0.36]	1000	0.839	0.75	0.69	0.82	0.86	0.86	0.78	**0.87**	0.854
5	[0.5 0.5]	2'500	0.80	0.63	0.74	0.84	0.85	0.866	0.813	**0.878**	0.854
8	[0.47 0.53]	4'000	0.79	0.75	0.65	0.85	0.85	0.866	0.78	**0.88**	0.854
11	[0.48 0.52]	5'500	0.81	0.53	0.76	0.83	0.85	0.867	0.8	**0.874**	0.854
14	[0.53 0.47]	7'000	0.83	0.70	0.75	0.87	0.86	0.85	0.77	**0.88**	0.854
17	[0.52 0.48]	8'500	0.82	0.75	0.76	0.86	0.86	0.86	0.79	**0.878**	0.854
20	[0.51 0.49]	10'000	0.80	0.77	0.79	0.87	0.854	**0.877**	0.86	0.874	0.854
Zurich data set											
2	[0.168 0.397 0.161 0.273]	2'936	0.61	0.52	0.57	0.59	0.61	0.65	0.59	**0.66**	0.65
5	[0.222 0.385 0.181 0.212]	6'716	0.62	0.55	0.6	0.58	0.6	0.66	0.58	**0.68**	0.64
8	[0.248 0.462 0.172 0.118]	16'448	0.63	0.54	0.59	0.59	0.61	0.67	0.63	**0.71**	0.65
11	[0.261 0.478 0.164 0.097]	21'223	0.63	0.54	0.58	0.59	0.62	0.67	0.58	**0.72**	0.673
14	[0.256 0.448 0.192 0.103]	27'875	0.63	0.52	0.58	0.59	0.62	0.67	0.59	**0.72**	0.65
17	[0.25 0.415 0.207 0.129]	32'660	0.63	0.5	0.59	0.59	0.63	0.67	0.6	**0.73**	0.61

Table 2: Results on multi-source simulated data and pixel classification results obtained on the Zurich Summer data set. The underline numbers and those in bold correspond to the second and the best performances obtained for each configuration, respectively.

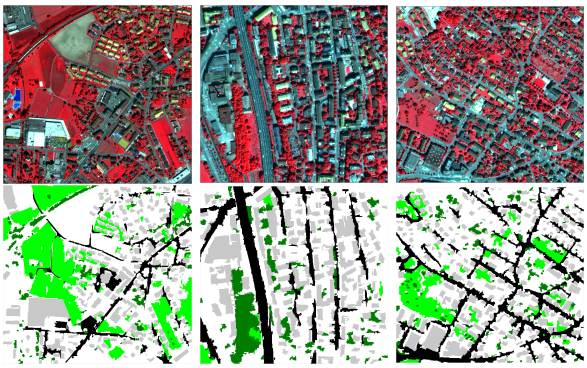

Figure 2: Original (top row) and ground truths (bottom row) images from Zurich data set. Class proportions are highly imbalanced between the images. Color legend for ground truths: black: roads, gray: buildings, green: grass; dark green: trees.

Figure 2 presents both the original and the ground truths of several images from the considered data set. One can observe that classes of all three images have very unequal proportions compared to each other.

Results The results over 5 trials obtained on this data set are reported in the lower part of Table 2. The proposed `JCPOT` method based on label propagation significantly improves the classification accuracy over the other baselines. The results show an important improvement over the "No adaptation" case, with an increase reaching 10% for `JCPOT-LP`. We also note that the results obtained by `JCPOT-LP` outperform the "Target only" baseline. This shows the benefit brought by multiple source domains as once properly adapted, they represent a much larger annotated sample that the target domain sample alone. This claim is also confirmed by an increasing performance of our approach with the increasing number of source domains. Overall, the obtained results show that the proposed method handles the adaptation problem quite well and thus allows to avoid manual labeling in real-world applications.

6 CONCLUSIONS

In this paper we proposed JCPOT, a novel method dealing with target shift: a particular and largely understudied DA scenario occurring when the difference in source and target distributions is induced by differences in their class proportions. To justify the necessity of accounting for target shift explicitly, we presented a theoretical result showing that unmatched proportions between source and target domains lead to inefficient adaptation. Our proposed method addresses the target shift problem by tackling the estimation of class proportions and the alignment of domain distributions jointly in optimal transportation framework. We used the idea of Wasserstein barycenters to extend our model to the multi-source case in the unsupervised DA scenario. In our experiments on both synthetic and real-world data, JCPOT method outperforms current state-of-the-art methods and provides a computationally attractive and reliable estimation of proportions in the unlabeled target sample. In the future, we plan to extend JCPOT to estimate proportions in deep learning-based DA methods suited to for larger datasets.

Acknowledgements. This work was partly funded through the projects OATMIL ANR-17-CE23- 0012 and LIVES ANR-15-CE23-0026 of the French National Research Agency (ANR).

References

Ben-David, Shai, Blitzer, John, Crammer, Koby, Kulesza, Alex, Pereira, Fernando, & Vaughan, Jennifer. 2010. A theory of learning from different domains. *Machine Learning*, **79**, 151–175.

Benamou, Jean-David, Carlier, Guillaume, Cuturi, Marco, Nenna, Luca, & Peyré, Gabriel. 2015. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, **37**(2), A1111–A1138.

Bickel, Steffen, Brückner, Michael, & Scheffer, Tobias. 2007. Discriminative Learning for Differing Training and Test Distributions. *Pages 81–88 of: ICML*.

Blanchard, Gilles, Lee, Gyemin, & Scott, Clayton. 2010. Semi-Supervised Novelty Detection. *Journal of Machine Learning Research*, **11**, 2973–3009.

Bobkov, S., & Ledoux, M. 2016. One-dimensional empirical measures, order statistics and Kantorovich transport distances. *To appear in: Memoirs of the AMS*.

Chan, Yee Seng, & Ng, Hwee Tou. 2005. Word Sense Disambiguation with Distribution Estimation. *Pages 1010–1015 of: IJCAI*.

Courty, N., Flamary, R., & Tuia, D. 2014. Domain adaptation with regularized optimal transport. *Pages 1–16 of: ECML/PKDD*.

Courty, Nicolas, Flamary, Rémi, Habrard, Amaury, & Rakotomamonjy, Alain. 2017a. Joint distribution optimal transportation for domain adaptation. *Pages 3733–3742 of: NIPS*.

Courty, Nicolas, Flamary, Rémi, Tuia, Devis, & Rakotomamonjy, Alain. 2017b. Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**(9), 1853–1865.

Cuturi, Marco. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Pages 2292–2300 of: NIPS*.

Fernando, Basura, Habrard, Amaury, Sebban, Marc, & Tuytelaars, Tinne. 2013. Unsupervised Visual Domain Adaptation Using Subspace Alignment. *Pages 2960–2967 of: ICCV*.

Fournier, Nicolas, & Guillin, Arnaud. 2015. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, **162**(3-4), 707.

Ganin, Yaroslav, & Lempitsky, Victor S. 2015. Unsupervised Domain Adaptation by Backpropagation. *Pages 1180–1189 of: ICML*, vol. 37.

Huang, J., Smola, A.J., Gretton, A., Borgwardt, K., & Schölkopf, B. 2007. Correcting Sample Selection Bias by Unlabeled Data. *In: NIPS*, vol. 19.

Japkowicz, Nathalie, & Stephen, Shaju. 2002. The Class Imbalance Problem: A Systematic Study. *Pages 429–449 of: IDA*, vol. 6.

Kantorovich, L. 1942. On the translocation of masses. *Doklady of the Academy of Sciences of the USSR*, **37**, 199–201.

Lin, Yi, Lee, Yoonkyung, & Wahba, Grace. 2002. Support Vector Machines for Classification in Nonstandard Situations. *Machine Learning*, **46**(1-3), 191–202.

Liu, Anqi, & Ziebart, Brian D. 2014. Robust Classification Under Sample Selection Bias. *Pages 37–45 of: NIPS*.

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

Mansour, Yishay, Mohri, Mehryar, & Rostamizadeh, Afshin. 2009a. Domain Adaptation: Learning Bounds and Algorithms. *In: COLT*.

Mansour, Yishay, Mohri, Mehryar, & Rostamizadeh, Afshin. 2009b. Domain adaptation with multiple sources. *Pages 1041–1048 of: NIPS*.

Monge, Gaspard. 1781. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Academie Royale des Sciences*, 666–704.

Patel, V. M., Gopalan, R., Li, R., & Chellappa, R. 2015. Visual domain adaptation: a survey of recent advances. *IEEE Signal Processing Magazine*, **32**(3), 53–69.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. 2009. *Dataset Shift in Machine Learning*. MIT Press.

Sanderson, Tyler, & Scott, Clayton. 2014. Class Proportion Estimation with Application to Multiclass Anomaly Rejection. *Pages 850–858 of: AISTATS*, vol. 33.

Scott, Clayton, Blanchard, Gilles, & Handy, Gregory. 2013. Classification with Asymmetric Label Noise: Consistency and Maximal Denoising. *Pages 489–511 of: COLT*, vol. 30.

Storkey, Amos J. 2009. When training and test sets are different: characterising learning transfer. *Pages 3–28 of: In Dataset Shift in Machine Learning*. MIT Press.

Tuia, D., Flamary, R., Rakotomamonjy, A., & Courty, N. 2015. Multitemporal classification without new labels: a solution with optimal transport. *In: 8th International Workshop on the Analysis of Multitemporal Remote Sensing Images*.

Tuia, Devis, Volpi, Michele, & Moser, Gabriele. 2018. Decision Fusion With Multiple Spatial Supports by

Conditional Random Fields. *IEEE Transactions on Geoscience and Remote Sensing*, 1–13.

Wen, Junfeng, Yu, Chun-Nam, & Greiner, Russell. 2014. Robust Learning under Uncertain Test Distributions: Relating Covariate Shift to Model Misspecification. *Pages 631–639 of: ICML*.

Yu, Yang, & Zhou, Zhi-Hua. 2008. A Framework for Modeling Positive Class Expansion with Single Snapshot. *Pages 429–440 of: PAKDD*.

Zadrozny, B., Langford, J., & Abe, N. 2003. Cost-Sensitive Learning by Cost-Proportionate Example Weighting. *Page 435 of: ICDM*.

Zhang, Kun, Schölkopf, Bernhard, Muandet, Krikamol, & Wang, Zhikun. 2013. Domain Adaptation under Target and Conditional Shift. *Pages 819–827 of: ICML*, vol. 28.

Zhang, Kun, Gong, Mingming, & Schölkopf, Bernhard. 2015. Multi-Source Domain Adaptation: A Causal View. *In: AAAI*.