# JDCOT : an Algorithm for Transfer Learning in Incomparable Domains using Optimal Transport

Marion Jeamart [1] & Renan Bernard [1] & Nicolas Courty [1] & Chloé Friguet [1]
& Valérie Garès[2]

[1] *Univ. Bretagne-Sud, UMR 6074, IRISA, Vannes, France - prenom.nom@univ-ubs.fr*
[2] *INSA, UMR 6625, IRMAR, Rennes, France - valerie.gares@insa-rennes.fr*

**Résumé.** L'adaptation de domaine est un champ de l'apprentissage par transfert où les données d'entrainement (source) du modèle et celles utilisées pour le test (cible) proviennent de deux domaines dont les distributions sous-jacentes sont différentes, et il convient d'adapter le modèle appris pour qu'il puisse être utilisé sur les données cibles avec de bonnes performances. On présente ici un algorithme permettant de traiter l'adaptation de domaine dans le cas d'espaces source et cible de hétérogènes car représentés par des espaces de caractéristiques différents. La méthode développée utilise le transport optimal pour coupler les distributions des deux domaines et son implémentation est illustrée sur des données de référence.

**Mots-clés.** transport optimal, apprentissage par tranfert, adaptation de domaine hétérogène

**Abstract.** Domain adaptation is a field of transfer learning where the training data (source) and the test data (target) come from different domains. The data in these two domains have therefore different underlying distributions, and the learned model should be adapted so that it can be used on the target data with good performance. We present here a domain adaptation algorithm to adapt heterogeneous domains, *i.e.* described by different features. The developed method uses optimal transport to map the distributions of the two domains and its implementation is illustrated on benchmark data.

**Keywords.** optimal transport, transfer learning, heterogeneous domain adaptation

## 1 Introduction

Transfer learning aims at leveraging knowledge or models learnt on a specific task to different, but somehow related, learning tasks. In the specific case where the learning tasks (classification, regression) are the same, but the distribution of the learning and test data are different, domain adaptation (DA) is considered to slightly update, with no human supervision, a model trained on a source domain ($S$) so that it becomes more robust when being used on data of interest lying in another domain, also named target domain ($T$), describing the same phenomenon but with another point of view.

In general, most of the existing domain adaptation methods assume that data of both source and target domains are represented by the same features space, with identical dimensions. But in some applications, it does not hold so that development of novel methods must consider domain adaptation across heterogeneous feature spaces, which is referred to as heterogeneous domain adaptation (HDA).

Transfer learning in general is useful in a context where the availability of data is limited. Using knowledge acquired in a setting into an other setting helps to solve issues when data lack. Different cases for supervision are investigated in the literature, and main methods can be classified regarding two strategies: (1) project both data into a common subspace by jointly learning the common subspace and a classifier, then iteratively align the discriminative dimensions ([7], [1]), and (2) jointly perform implicit data reconstruction and learn a classifier ([4]).

In this paper, we focus on three kinds of HDA (Table 1): (1) unsupevised HDA when all the source labels are observed, but none in target, (2) semi-supervised HDA when all the source labels and some labels in target are observed, (3) partial HDA when source and target labels are both partially observed.

| | $Y^S$ | $Y^T$ |
|---|---|---|
| Unsupervised DA | observed | unobserved |
| Semi-supervised DA | observed | partially observed |
| Partial DA | partially observed | partially observed |

Table 1: Several learning cases depending on the labels availability in $S$ and/or $T$. *Note: in the partial DA setting, since labels in both domains are partially observed, source and target domains can be interchanged.*

Recently, optimal transport (OT) has been investigated as an efficient tool to deal with DA issues in a unsupervised or semi-supervised context. In the following, we first present the main principles of OT for DA in section 2. Then, we propose to extend the use of OT for DA to the case of heterogeneous transfer learning in section 3. Several numerical experiments are conducted to assess the proposed algorithm. Finally, some perspectives are discussed in section 4

# 2 Optimal transport for domain adaptation

**Notations** In the following, let $(X^S, Y^S) \in \mathbb{R}^{n^S \times d^S} \times \mathcal{C}$ be the source data and $(X^T, Y^T) \in \mathbb{R}^{n^T \times d^T} \times \mathcal{C}$ the target data. Moreover, $P(x^S, y^S) \in \mathcal{P}(\Omega^S, \mathcal{C})$ and $P(x^T, y^T) \in \mathcal{P}(\Omega^T, \mathcal{C})$ denote the joint probability distribution of the data in the source and target domains, respectively.

**Optimal Transport (OT)** Firstly introduced by G. Monge in 1781 and recently more widely studied (see [5] for details), OT is an optimisation problem that allows to define a distance between two probability measures (discrete, semi-discrete or continuous) called the Wasserstein distance. This distance is actually used as a loss in many optimisation problems and approximation algorithms allow to scale this problem to large dimension settings. We consider two sets of weighted samples $X^S = \{(x_i^S, w_i^S), i = 1 \dots n^S, \sum_{i=1}^{n^S} w_i^S = 1\}$ and $X^T = \{(x_j^T, w_j^T), j = 1 \dots n^T, \sum_{j=1}^{n^T} w_j^T = 1\}$.

The Kantorovich formulation consists in finding a coupling matrix $P \in \mathbb{R}_+^{n^S \times n^T}$ that satisfies $\gamma = \underset{P \in U(w^S, w^T)}{\operatorname{argmin}} \sum_{i,j} C_{ij} P_{ij}$ where $U(w^S, w^T)$ is the set of matrices $P \in \mathbb{R}_+^{n^S \times n^T}$ which verifies $\sum_{i=1}^{n^S} P_{ij} = w_j^T$, $\forall j = 1 \dots n^T$ and $\sum_{j=1}^{n^T} P_{ij} = w_j^T$, $\forall i = 1 \dots n^S$, and $C$ is a cost matrix. In the following, we will note : $\gamma = OT(w^S, w^T, C)$.

**Optimal Transport for (unsupervised) Domain Adaptation (OTDA)[2].** Assuming that $P(y^T|X^T) = P(y^S|M(X^S))$ and $P(X^T) = P(M(X^S))$ with $M : \Omega^S \mapsto \Omega^T$ a nonlinear transformation of the input space, the transport map $\gamma$ is obtained by solving the OT problem between $X^S$ and $X^T$ : $\gamma = OT(w^S, w^T, C)$ where $w^S = \mathbf{1}_{n^S}/n^S$, $w^T = \mathbf{1}_{n^T}/n^T$ and $C \in \mathbb{R}^{n^S \times n^T}$ such that $C_{ij} = d(x_i^S, x_j^T)$, with $d$ a dissimilarity measure and $\mathbf{1}$ the indicator function.

$\gamma \in \mathbb{R}^{n^S \times n^T}$ can be considered as the empirical joint distribution of $P(X^S, X^T)$. The transport map $\gamma$ is then used to calculate the barycentric coordinates of the source samples in the target domain: $M(x_i^S) = \frac{1}{w_i^S} \sum_j \gamma_{ij} x_j^T$. Finally, a classifier $f$ is trained on the transported source data $(M(X^S), Y^S)$ and applied on the target data $X^T$ to predict target labels : $Y_{pred}^T = f(X^T)$.

**Joint Distribution Optimal Transport (JDOT) [3].** Developed to deal with changes in the marginal distributions of the features and in the conditional distributions of $Y$ that can occur with real-world data but that are not handled by OTDA, JDOT consists in simultaneously optimising the coupling matrix $\gamma$ and the classifier $f$ by minimising:

$$\min_{\gamma, f} \sum_{i,j} \Big[ \alpha \; d(x_i^S, x_j^T) + \mathcal{L}(y_i^S, f(x_j^T)) \Big] \gamma_{ij}$$

where $\alpha$ is a hyper-parameter and $\mathcal{L}$ is a loss function. A Block Coordinate Descent (BCD) is performed to alternatively estimate $\gamma$ and $f$ in practice. The authors have shown the superiority of their approach through experiments on benchmark datasets *w.r.t.* several domain adaptation state-of-the-art methods, including previous OT-based approaches, domain adversarial neural networks or transfer components.

Nevertheless, this approach does not address the heterogeneous domain adaptation problem. In the following, JDCOT is introduced to deal with this context.

# 3 JDCOT: Optimal transport for heterogeneous transfer learning

The aim of the proposed method called Joint Distribution Co-Optimal Transport (JDCOT) is to deal with heterogeneous domain adaptation with different dimension data. The way OT maps two domains makes it impractical if they are on different spaces. Co-Optimal Transport (COOT) [6] has therefore been developed to deal with incomparable spaces, by simultaneously optimising two transport maps between both samples and features with a BCD. We propose to use the principle of COOT to adapt JDOT to the HDA framework.

**Method** The JDCOT method consists in simultaneously solving the OT problem on the samples ($\gamma^s$), the OT problem on the variables ($\gamma^v$) and the classification problem ($f^T$), by minimising with a BCD, with $i \in [1; n^S], j \in [1; n^T], k \in [1; d^S], \ell \in [1; d^T]$:

$$\min_{\substack{\gamma^s, \gamma^v, f^S, f^T \\ \gamma^s \in U(w^S, w^T) \\ \gamma^v \in U(v^S, v^T)}} \sum_{i,j,k,\ell} \left[ \alpha \; d(X^S_{i,k}, X^T_{j,\ell}) + \mathcal{L}\Big( \tilde{f}^S(x^S_i), \tilde{f}^T(x^T_j) \Big) \right] \gamma^s_{ij} \gamma^v_{k\ell} \tag{1}$$

where $v^S$ and $v^T$ are the weights of the points $(X')^S$ and $(X')^T$ respectively, $\alpha$ is the JDOT hyper-parameter, and $\tilde{f}^S$ and $\tilde{f}^T$ are classifiers, respectively on source and target domains, such that $\tilde{f}(x) = y$ if $y$ is observed and $\tilde{f}(x) = f(x)$ otherwise.

**Experiments: Settings and datasets** We have conducted extensive numerical experiments to illustrate the proposed method considering the three settings reported in Table 1 (unsupervised DA, semi-supervised DA and partial DA), using the MNIST and USPS datasets. These datasets contain gray level images of handwritten digits ($K = 10$ classes) sized $16 \times 16$ for USPS (source) and $28 \times 28$ for MNIST (target), so that the two domains do not share the same size, as expected in the heterogeneous setting. The experimental protocol consists in randomly selecting 300 images per class ($n^S = n^T = 300 \times 10$) or 30 images per class ($n^S = n^T = 30 \times 10$) in each dataset for the training set. $n^S_*$ (resp. $n^T_*$) images are labelled in the source (resp. target) domain. The test set is made up of 200 target images per class and is fixed such as it's the same for all the experiments. A convolutional neural network composed of 2 convolutional and 2 dense layers (CNN) is considered as classifiers.

**Unsupervised and semi-supervised HDA** At first, JDCOT and COOT are compared in the unsupervised ($n^S_* = n^S$ and $n^T_* = 0$) and semi-supervised ($n^S_* = n^S$ and $\forall k \in [1; K], n^T_{k,*} \in \{1; 3; 10\}$) cases. In these settings, all source labels are known such that $\forall x^S_i \in X^S, \tilde{f}^S(x^S_i) = y^S_i$. In the semi-supervised case, $f^T$ is initialised with the $n^T_*$

available target labels. Its performance on the test set is reported as a baseline score in Table 2. Table 2 also reports the correct classification rate on the test set (test accuracy) over 10 random samplings for the training sets. The test accuracy of COOT is computed with the CNNt model trained with the estimated target labels given by label propagation through $\gamma^s$. The test accuracy of JDCOT is the one given by $f^T$ at the end of the BCD procedure. In the unsupervised case, since no target label is available to initialise $f^T$, a first estimate of the target labels is given by label propagation after executing the unsupervised COOT method. Both JDCOT and COOT allow to improve the baseline score and JDCOT's performance is growing along with the number of known target labels. The 3 000-sized sample is big enough to allow a good unsupervised performance for COOT so that introducing some target labels does not improve it. In the 300-sized sample case, introducing semi-supervision allows to considerably improve both COOT and JDCOT's performances. Finally, JDCOT outperforms COOT in all cases and is generally more stable.

| $n^T_{k,*}$ | baseline | $n^S = n^T = 3\ 000$ | | $n^S = n^T = 300$ | |
|---|---|---|---|---|---|
| | | COOT | JDCOT | COOT | JDCOT |
| 0 | - | 72.96 ±8.2 | 77.27 ±9.1 | 57.27 ±16.2 | 58.08 ±17.2 |
| 1 | 39.59 ±6.0 | 75.81 ±4.9 | 78.45 ±1.1 | 61.74 ±14.5 | 69.98 ±2.8 |
| 3 | 56.82 ±4.4 | 75.35 ±6.5 | 79.02 ±0.9 | 69.71 ±7.2 | 73.19 ±2.4 |
| 10 | 80.49 ±3.1 | 75.75 ±6.8 | 88.34 ±1.7 | 77.25 ±1.7 | 85.67 ±1.7 |

Table 2: Mean and standard deviation of the test accuracy (%) over 10 random samplings for the training sets, considering two sample sizes. $n^T_{*,k}$ denotes the number of known labels in each class $k$, in target domain.

**Partial HDA** The partial HDA setting (semi-supervision in both domains) is then considered, with $n^S = n^T = 3\ 000$ and $n^S_{k,*} = n^T_{k,*} \in \{3; 5; 25; 100\}$ labelled samples per class in each domain. Table 3 reports the test accuracy after initialising the model with the available target labels (init) and after the whole process (final), along with the number of known labels per class. JDCOT increases the accuracy both on source and target domains which shows that the information transfer established by solving the OT problem from a domain to another works positively.

| JDCOT | $n_{*,k}$ | 3 | 5 | 25 | 100 |
|---|---|---|---|---|---|
| source | init | 70.9 ±4.3 | 77.9 ±2.3 | 92 ±0.7 | 97.6 ±0.4 |
| | final | 73.5 ±5.3 | 84.6 ±2.5 | 94.6 ±0.9 | 98 ±0.2 |
| target | init | 62.7 ±3.2 | 70.5 ±3.4 | 90.2 ±0.9 | 96.1 ±0.7 |
| | final | 68.7 ±5.5 | 79 ± 3.2 | 90.3 ±0.5 | 97 ± 0.2 |

Table 3: Mean and standard deviation of the test accuracy (%) over 10 random samplings for the training set. $n^S = n^T = 3\,000$. $n_{*,k}$ denotes the number of known labels in each class $k$, in each domain.

# 4 Conclusion and discussion

Joint Distribution Co-Optimal Transport (JDCOT) is a method to deal with heterogeneous transfer learning using optimal transport: it performs domain adaptation in the case of source and target spaces of different features and different dimensions, matching both samples and features with transport maps. The method can deal with unsupervised, semi-supervised and partial domain adaptation.

An extension of JDCOT to a deep learning setting, for example to study image datasets, has also been defined to enable dealing with larger data dimensions by simultaneously optimising 2 transport plans and 2 features extractors $g^S$ and $g^T$ (CNNs). The vector representations of the data are compared instead of the data themselves, and optimisation is done with minibatch stochastic gradient descent.

# References


[1] W.-Y. Chen, T.-M. H. Hsu, Y.-H. H. Tsai, M.-S. Chen, and Y.-C. F. Wang. Transfer neural trees: Semi-supervised heterogeneous domain adaptation and beyond. *IEEE Transactions on Image Processing*, 28(9):4620–4633, 2019. doi: 10.1109/TIP.2019.2912126.

[2] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2016.

[3] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. volume 30, 2017.

[4] H. Li, S. J. Pan, R. Wan, and A. C. Kot. Heterogeneous transfer learning via deep matrix completion with adversarial kernel embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (01):8602–8609, 2019.

[5] G. Peyré and M. Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–206, 2018.

[6] I. Redko, T. Vayer, R. Flamary, and N. Courty. CO-Optimal Transport. In *Neural Information Processing Systems (NeurIPS)*, Online, Canada, Dec. 2020.

[7] Y. Yao, Y. Zhang, X. Li, and Y. Ye. Heterogeneous domain adaptation via soft transfer network. *CoRR*, abs/1908.10552, 2019.