Optimal transport for Domain adaptation

Nicolas Courty
University of Bretagne Sud - IRISA
Campus de Tohannic, 56000 Vannes, France
ncourty@irisa.fr

Rémi Flamary
Laboratoire Lagrange UMR CNRS 7293
Observatoire de la Côte d'Azur
Université de Nice, France
remi.flamary@unice.fr

Alain Rakotomamonjy
LITIS Université de Rouen
76800 Saint Etienne du Rouvray, France
alain.rakotomamonjy@univ-rouen.fr

Devis Tuia
Department of Geography
University of Zurich
8057 Zurich, Switzerland
devis.tuia@geo.uzh.ch

Abstract

Domain adaptation from one data space (or domain) to the other is one of the most challenging tasks of modern data analytics. If the adaptation is done correctly, models built on a specific data space become able to process data depicting the same semantic concepts (the classes), but observed by another observation system with its own specificities. In this paper, we propose an optimal transportation model that aligns the representations in the source and target domains. We learn a transportation plan matching both PDFs and constrain labeled samples in the source domain to remain close during transport with a non convex group regularization. This way, we exploit at the same time the labeled information in the source (the same that will be used by the classifier after adaptation) and the unlabeled distributions observed in both domains. We propose an efficient majoration-minimization algorithm to solve the resultaing optimization problem and discuss its convergence. Numerical experiments of real data show the interest of the method, that outperforms state-of-the-art approaches.

1 Introduction

The multiplication of data sources and acquisition devices provides nowadays tremendous quantities of data. In practical applications, the wealth of data available is however often counterbalanced by the lack of annotated information, which is generally necessary to run classification algorithms aiming at generalizing over new unseen examples. Moreover, classical learning methods are challenged by the plurality of sources and by the need of designing methods that are accurate when predicting in a previously unseen environment, or target domain: this is mostly due to subtle or pronounced discrepancies observed in the different data distributions, or *drifts*. In computer vision, for example, this problem is known as the visual adaptation problem, where domain to domains drifts may occur when changing lighting conditions, acquisition devices, or by considering the presence or absence of backgrounds [1]. In practice, the causes of drift are numerous and application-specific.

Several works study the generalization capabilities of a classifier allowing to transfer knowledge from a labeled source domain to an unlabeled target domain: this situation is referred to as transductive transfer learning [2]. In this work, we assume that the source and target domains are by essence different, which is usually referred to as the domain adaptation. We address the most difficult variant of this problem, where data labels are only available in the source domain. This is the **unsupervised domain adaptation** problem, whose bet is that the effects of the drifts can be

reduced if data undergo a phase of *adaptation* toward a common representation where both domains look more alike.

Several theoretical works [3, 4, 5] have emphasised the role played by the divergence between the two domains probability distribution functions, leading to a principled way of solving the domain adaptation problem: to bring closer both distributions, while using the label information available in the source domain to learn a classifier. This work follows the same intuition, by exploring the use of the optimal transport (OT) distances as a measure of divergence and transporting the samples so that theirs distribution is more similar. OT distances are also known as Wasserstein, Monge-Kantorovich or Earth Mover distances, and have very strong and important properties: *i)* they can be evaluated when only empirical measures of those distributions are observed, and without the estimation of parametrical or semi-parametrical distributions as a pre-process; *ii)* there is no particular constraint on the overlap of the support of the distributions to provide meaningful results, which is clearly not the case with other information theoretic divergences, such as the Kullback-Leibler divergence. Building on those properties, we propose an original algorithm for domain adaptation based on OT. Most of the method and some results presented in this paper were already previously published in [6]. We provide in this paper a complementary discussion on the convergence of the computation strategy, and a new experiment on real data which validates the previously established conclusions.

Related works on Domain Adaptation are presented in the next Section, while Section 3 formalizes the problem of unsupervised domain adaptation and the use of optimal transport to solve it. The originality of our approach resides in the inclusion of an additional regularization term tailored to fit the domain adaptation constraints. Their pertinency is examined in the experimental Section 5, where we demonstrate the efficiency of the new proposed framework on an optical character recognition task and a computer vision problem.

2 Related works on Domain adaptation

Domain adaptation strategies can be roughly divided in two families, depending on whether they can access labels in the target domains (semi-supervised DA) or not (unsupervised DA).

In the first family, we find methods searching for projections discriminative in both domains, either by using dot products between the source samples and the transformed target samples [1, 7, 8], by learning projections, for which labeled samples of the target domain fall on the correct side a large margin classifier trained on the source data [9] or by extracting common features under pairwise constraints [10, 11].

The second family is the one considered in this paper. Many works have considered finding a common feature representation for the two (or more) domains. This representation, or *latent space*, allows to project samples from all domains in a space where a classifier using only the labeled samples from the source domain generalize well on the target domains [12, 13]. The representation transfer can be performed by matching the means of the domains in the feature space [13], aligning the domains by their correlations [14] or by using pairwise constraints [15]. In most of these works, the common latent space is found via feature extraction, where the dimensions retained summarize the information common to the domains. Recently, the unsupervised domain adaptation problem has been revisited by considering strategies based on a gradual alignment of the feature representation: in [16], authors compare gradual distortions and therefore use intermediary projections of both domains along the Grassmannian geodesic connecting the source and target observed eigenvectors. In [17, 18], authors propose to obtain all sets of transformed intermediary domains by using a geodesic-flow kernel, instead of sampling a fixed number of projections along the geodesic path. While these methods have the advantage of providing easily computable out-of-sample extensions (by projecting unseen samples onto the latent space eigenvectors), the transformation defined remains global and must be therefore applied the same way to the whole target domain.

Our proposition strongly differs from those reviewed above, as it defines a local transportation plan for each sample in the source domain. In this sense, the domain adaptation problem can be seen as a graph matching problem [19, 20] for all samples to be transported, where their final coordinates are found by mapping the source samples to coordinates matching the marginal distribution of the target domain. In the authors knowledge, this is the first attempt to use optimal transportation theory in domain adaptation problems.

3 Optimal transportation and domain adaptation

Let $\Omega \subseteq \mathbb{R}^d$ be an input measurable space of dimension d and $\mathcal{C} = \{-1, 1\}$ the set of labels. $\mathcal{P}(\Omega)$ denotes the set of all the probability measures over Ω. The standard learning paradigm assumes classically the existence of a set of data $\mathbf{X}_s = \{\mathbf{x}_i^s\}_{i=1}^{N_s}$ associated with a set of class label information $\mathbf{Y}_s = \{\mathbf{y}_i^s\}_{i=1}^{N_s}$, $\mathbf{y}_i^s \in \mathcal{C}$ (the learning set), and a data set with unknown labels $\mathbf{X}_t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$ (the testing set). In order to determine the set of labels $\mathbf{Y}_t$ associated with $\mathbf{X}_t$, one usually relies on an empirical estimate of the joint probability distribution $\mathbf{P}(\mathbf{X}, \mathbf{Y}) \in \mathcal{P}(\Omega \times \mathcal{C})$ from $(\mathbf{X}_s, \mathbf{Y}_s)$, and the assumption that $\mathbf{X}_s$ and $\mathbf{X}_t$ are drawn from the same distribution $\mu \in \mathcal{P}(\Omega)$.

In the considered adaptation problem, one assumes the existence of two distinct joint probability distributions $\mathbf{P}_s(\mathbf{X}, \mathbf{Y})$ and $\mathbf{P}_t(\mathbf{X}, \mathbf{Y})$, which correspond respectively the *source* and *target* domains. Their respective marginal distribution over $\mathbf{X}$ are μ_s and μ_t.

We are searching for a transformation between the two domains that minimizes the impact of the domain change. This intuition is motivated by theoretical generalization bound [21], which contains the divergence between the source and target distributions. Based on such bound, we propose a principled way to perform domain adaptation that reduces both this divergence and the classification error in the source domain. This method is based on the computation of OT between the empirical distributions and performing a transportation of the source samples and their label onto the target distribution, leading to a decrease in the divergence between those distributions.

3.1 Monge-Kantorovitch optimal transportation and discrete distributions

The Kantorovitch formulation of the optimal transport [22] seeks for a probabilistic coupling $\boldsymbol{\gamma} \in \mathcal{P}(\Omega_1 \times \Omega_2)$ between Ω_1 and Ω_2:

$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \int_{\Omega_1 \times \Omega_2} c(\mathbf{x}, \mathbf{y}) d\boldsymbol{\gamma}(\mathbf{x}, \mathbf{y}), \qquad \text{s.t. } \mathrm{P}^{\Omega_1} \# \boldsymbol{\gamma} = \mu_s, \mathrm{P}^{\Omega_2} \# \boldsymbol{\gamma} = \mu_t, \tag{1}$$

where P^{Ω_i} is the projection over Ω_i. In this formulation, $\boldsymbol{\gamma}$ can be understood as a joint probability measure with marginals μ_s and μ_t. $\boldsymbol{\gamma}_0$ is the unique solution to the optimal transport problem.

Since one does not have a direct access to μ_s or μ_t, but rather to collections of samples from those distributions, the optimal transport problem is generally adapted to the discrete case. In this case, the two empirical distributions can be expressed as

$$\mu_s = \sum_{i=1}^{n_s} p_i^s \delta_{\mathbf{x}_i^s}, \quad \mu_t = \sum_{i=1}^{n_t} p_i^t \delta_{\mathbf{x}_i^t} \tag{2}$$

where $\delta_{\mathbf{x}_i}$ is the Dirac at location $\mathbf{x}_i \in \mathbb{R}^d$. p_i^s and p_i^t are probability masses associated to the i-th sample, and belonging to the probability simplex, *i.e.* $\sum_{i=1}^{n_s} p_i^s = \sum_{i=1}^{n_t} p_i^t = 1$. The set of probabilistic coupling between those two distributions is the set of doubly stochastic matrices $\mathcal{P}$ defined as

$$\mathcal{P} \quad = \quad \left\{ \boldsymbol{\gamma} \in (\mathbb{R}^+)^{n_s \times n_t} \middle| \, \boldsymbol{\gamma} \mathbf{1}_{n_t} = \mu_s, \boldsymbol{\gamma}^T \mathbf{1}_{n_s} = \mu_t \right\} \tag{3}$$

where $\mathbf{1}_d$ is a d-dimensional vector of ones. The Kantorovitch formulation of the optimal transport [22] becomes:

$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\operatorname{argmin}} \left\langle \boldsymbol{\gamma}, \mathbf{C} \right\rangle_F \tag{4}$$

where $\left\langle ., . \right\rangle_F$ is the Frobenius dot product and $\mathbf{C} \geq 0$ is a cost matrix, whose terms $C(i, j)$ depict the energy needed to move a probability mass from $\mathbf{x}_i^s$ to $\mathbf{x}_j^t$. In our setting, this cost was chosen as the Euclidian distance between the two locations, i.e. $C(i, j) = ||\mathbf{x}_i^s - \mathbf{x}_j^t||^2$, but alternative metrics could be interestingly explored.

Once the transport $\boldsymbol{\gamma}_0$ has been computed, the source samples must be transported in the target domain using their transportation plan. In our approach, we suggest to compute directly the image of the source samples as the result of this transport, *i.e.* for $t = 1$. Those images can be expressed through $\boldsymbol{\gamma}_0$ as center of masses of the weighted target samples. Let $\mathbf{T}_{\boldsymbol{\gamma}_0} : \mathbb{R}^d \to \mathbb{R}^d$ be the mapping induced by the optimal transport coupling. This map transforms the source elements $\mathbf{X}_s$ in a their corresponding elements in the target domain, $\hat{\mathbf{X}}_s$. The mapping $\mathbf{T}_{\boldsymbol{\gamma}_0}$ can be expressed as:

$$\hat{\mathbf{X}}_s = \mathbf{T}_{\boldsymbol{\gamma}_0}(\mathbf{X}_s) = \mathrm{diag}((\boldsymbol{\gamma}_0 \mathbf{1}_{n_t})^{-1}) \boldsymbol{\gamma}_0 \mathbf{X}_t. \tag{5}$$

Since $\mathbf{T}_{\boldsymbol{\gamma}_0}^{-1} = \mathbf{T}_{\boldsymbol{\gamma}_0^T}$, we note that $\mathbf{T}_{\boldsymbol{\gamma}_0}$ is fully invertible and can be also used to compute an adaptation from the target domain to the source domain.

3.2 Regularized optimal transport

Regularization is a classical approach used to prevent overfitting when only few samples are available, or even in presence of outliers. While it is always possible to enforce *a posteriori* a given regularity in the transport result, a more theoretically convincing solution is to regularize the transport by relaxing some of the constraints in the problem formulation of Eq.(4). This possibility has been explored in recent papers [23, 24].

More specifically, in [24], Cuturi proposes to regularize the expression of the transport by the entropy of the probabilistic coupling. The regularized version of the transport $\boldsymbol{\gamma}_0^\lambda$ is then the solution of the following minimization problem:

$$\boldsymbol{\gamma}_0^\lambda = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\operatorname{argmin}} \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F - \frac{1}{\lambda} h(\boldsymbol{\gamma}), \tag{6}$$

where $h(\boldsymbol{\gamma}) = -\sum_{i,j} \boldsymbol{\gamma}(i,j) \log \boldsymbol{\gamma}(i,j)$ computes the entropy of $\boldsymbol{\gamma}$. The intuition behind this form of regularization is the following: since most of the elements of $\boldsymbol{\gamma}_0$ should be zero with high probability, one can look for a smoother version of the transport by relaxing this sparsity through an entropy term. As a result, and contrary to the previous approach, more couplings with non-null weights are allowed, leading to a denser coupling between the distributions. An appealing result of this formulation is the possibility to derive a computationally very efficient algorithm, which uses the scaling matrix approach of Sinkhorn-Knopp [25].

4 Domain Adaptation with Label Regularized Optimal Transport

From the definitions above, the use of optimal transport for domain adaptation is rather straightforward: by computing the optimal transport from the source distribution μ_s to the target distribution μ_t, one defines a local transformation of the source domain to the target domain. This transformation can be used to adapt the training distribution by means of a simple interpolation. Once the source labeled samples have been transported, any classifier can be used to predict in the target domain. In this section, we present our optimal transport with label regularization algorithm (**OT-labreg**) and present an efficient algorithm to solve the problem. We finally discuss how to interpolate the training set from this regularized transport.

4.1 Regularizing the transport with class labels

As it was presented in te previous section, optimal transport does not include any information about the particular nature of the elements of the source domain (*i.e.* the fact that those samples belong to different classes). However, this information is generally available, as labeled samples are used in the classification step following the adaptation. Our proposition is to penalize couplings that match together source samples with known different labels. This is illustrated in Figure 1.c, where samples belonging to the same classes are only associated to points associated to the same class, contrarily to the standard and regularized versions of the transport (Figures 1.a and 1.b).

Principles of the label regularization. We want to concentrate the transport information on elements of the same class c for each column of $\boldsymbol{\gamma}$. This is usually achieved by using $\ell_p - \ell_q$ mixed-norm regularization. The main idea is that, even if we do not know the class of the samples in the target distribution, we can promote group sparsity in the columns of $\boldsymbol{\gamma}$ such that a given target point will be associated with only one of the classes observed in the source domain.

Promoting group sparsity leads to a new term in the cost function (6), which now reads:

$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\operatorname{argmin}} \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F - \frac{1}{\lambda} h(\boldsymbol{\gamma}) + \eta \sum_j \sum_c ||\boldsymbol{\gamma}(\mathcal{I}_c, j)||_q^p, \tag{7}$$

where $\mathcal{I}_c$ contains the index of the lines such that the class of the element is c, $\boldsymbol{\gamma}(\mathcal{I}_c, j)$ is a vector containing coefficients of the jth column of $\boldsymbol{\gamma}$ associated to class c and $||\cdot||_q^p$ denotes the ℓ_q

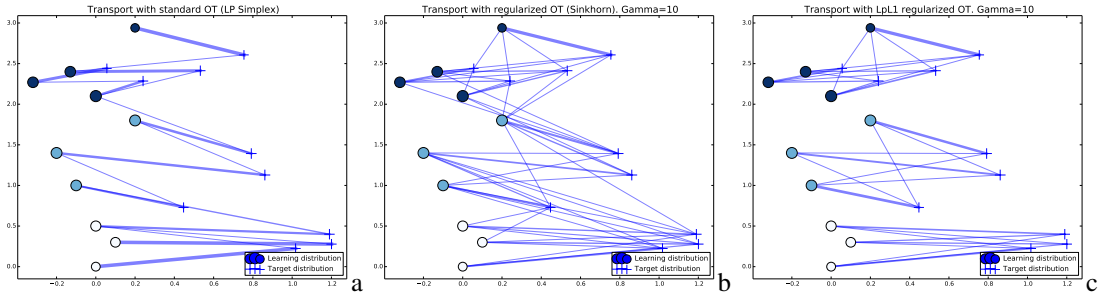


Figure 1: Optimal transport for two simple distributions ($\circ$ and $+$). The colored circles represent 3 different classes. The optimal transport solution is depicted as blue lines, whose thickness represents the strength of the coupling. (a) original optimal transport (**OT-ori**); (b) Sinkhorn transport (**OT-reg** [24]); (c) proposed class-wise regularization term (**OT-reglab**).

norm to the power of p. η is a regularization parameter that weights the impact of the supervised regularization.

Despite the fact that the choice $p = 1, q = 2$ is extremely popular choice for promoting group sparsity, due in part to the convexity of the regularization term, we decided to choose in this work $q = 1$ and $p = 1/2$. The use of the square root ($p = 1/2$) will enforce sparsity per group thanks to its non-dependability in 0, and this approach has been recently used for promoting non-grouped sparsity in compressed sensing [26]. Finally, this choice leads to a non-convex optimization problem but is a perfect fit for our application since majoration of the non-convex term leads to a linear loss that can be efficiently solved using the algorithm form [24] as discussed in the following.

4.2 Majoration Minimization strategy

The optimization problem with a $\ell_p - \ell_1$ regularization boils down to optimizing

$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\operatorname{argmin}} \, J(\boldsymbol{\gamma}) + \eta \Omega(\boldsymbol{\gamma}), \tag{8}$$

with $J(\boldsymbol{\gamma}) = \langle \boldsymbol{\gamma}, C \rangle_F - \frac{1}{\lambda} h(\boldsymbol{\gamma})$ and the regularization term that can be expressed as

$$\Omega(\boldsymbol{\gamma}) = \sum_j \sum_c ||\boldsymbol{\gamma}(\mathcal{I}_c, j)||_1^p = \sum_j \sum_c g(||\boldsymbol{\gamma}(\mathcal{I}_c, j)||_1) \tag{9}$$

where $g(\cdot) = (\cdot)^p$ is a concave function of a positive variable ($\forall \boldsymbol{\gamma} \geq 0$). A classical approach to address this problem is to perform what is called Majorization-minimization [27]. This can be done because the $\ell_p - \ell_1$ regularization term is concave in the positive orthant. For a given group of variables, one can use the concavity of g to majorize it around a given vector $\hat{\mathbf{w}} > 0$

$$g(\mathbf{w}) \leq g(\|\hat{\mathbf{w}}\|_1) + g'(\|\hat{\mathbf{w}}\|_1)^\top (\mathbf{w} - \hat{\mathbf{w}}) \tag{10}$$

with $g'(\|\hat{\mathbf{w}}\|_1) = p(\|\hat{\mathbf{w}}\|_1)^{p-1}$ for $\hat{\mathbf{w}} > 0$. For each group, the regularization term can be majorized by a linear approximation. In other words, for a fixed $\hat{\boldsymbol{\gamma}}$

$$\Omega(\boldsymbol{\gamma}) \leq \tilde{\Omega}(\boldsymbol{\gamma}) = \langle \boldsymbol{\gamma}, \mathbf{G} \rangle_F + cst \tag{11}$$

where the matrix $\mathbf{G}$ has components

$$\mathbf{G}(\mathcal{I}_c, j) = p(\|\hat{\boldsymbol{\gamma}}(\mathcal{I}_c, j)\| + \epsilon)^{p-1}, \quad \forall c, j \tag{12}$$

We added a small $\epsilon > 0$ that helps avoiding numerical instabilities, as discussed in [26]. Finally, solving problem (7) can be performed by iterating the two steps illustrated in Algorithm 1. This iterative algorithm is of particular interest in our case as it consists in iteratively using an efficient Sinkhorn-Knopp matrix scaling approach [24]. Moreover this kind of MM algorithm is known to converge in a small number of iterations.

In what follows, we provide some hints on the convergence of Algorithm 1. Remind that we want to solve the following group-sparse entropy-regularized optimal transport problem

$$\min_{\gamma \in \mathcal{P}} f(\gamma) + h(\gamma) \tag{13}$$

Algorithm 1 Majoration Minimization for $\ell_p - \ell_1$ regularized Optimal Transport

Initialize $\mathbf{G} = \mathbf{0}$
Initialize $\mathbf{C}_0$ as in Equation (4)
repeat
  $\mathbf{C} \leftarrow \mathbf{C}_0 + \mathbf{G}$
  $\boldsymbol{\gamma} \leftarrow$ Solve problem (6) with $\mathbf{C}$
  $\mathbf{G} \leftarrow$ Update $\mathbf{G}$ with Equation (12)
until Convergence

where $f(\gamma) = \langle \gamma, \mathbf{C} \rangle + \frac{1}{\lambda} \sum_{i,j} \gamma_{i,j} \log \gamma_{i,j}$, $h(\gamma) = \sum_k \left(\sum_{i,j \in A_k} \gamma_{i,j} + \epsilon \right)^p$, with $\epsilon > 0$ and $0 < p < 1$. The constraint set $\mathcal{P}$ over γ is defined as $\{\gamma \in \mathbb{R}^{n_s \times n_r} : \gamma_{i,j} \geq 0, \gamma \mathbf{1} = \mu_s, \gamma^\top \mathbf{1} = \mu_r\}$. Note that $f(\gamma)$ is a strictly convex and smooth function, but it is not gradient Lipschitz. Similarly, $h(\gamma)$ is a strictly concave and smooth function.

The existence of a minimizer of this problem naturally derives from the facts that the objective function is continuous and the set $\mathcal{P}$ can be included in a bounded subset of $\mathbb{R}^{n_s \times n_r}$. Hence, the objective function reaches its minimum over the constraint set.

Remark 1 *Because our problem is a classical constrained smooth optimization problem, from textbook results, we can say that $\gamma^\star$ is a critical point of our problem if the following holds*

$$\nabla f(\gamma^\star) + \nabla h(\gamma^\star) \in \mathcal{N}_{\mathcal{P}}(\gamma^\star) \tag{14}$$

where $\mathcal{N}_{\mathcal{P}}(\gamma)$ denotes the normal cone of $\mathcal{P}$ at γ. Because our problem is non-convex, this condition is only a necessary condition for local optimality.

The algorithm we used for solving the above problem is based on an iterative approach where at each iteration, the concave function h is linearly majorized by its first-order approximation.

Theorem 1 *Let $\{\gamma^k\}$ be the sequence generated by Algorithm 1. The following statements hold :*

1. *$\{f(\gamma^k) + h(\gamma^k)\}_k$ is a monotone non-increasing sequence.*
2. *assume that $\gamma^\star$ is a limit point of $\{\gamma^k\}$ then $\gamma^\star$ is a critical point of problem (13).*

Sketch of Proof: For a sake of clarify, we will note $g(\gamma) = f(\gamma) + h(\gamma)$ and $\tilde{g}(\gamma; \gamma^k) = f(\gamma) + h(\gamma^k) + \langle \nabla h(\gamma^k), \gamma - \gamma^k \rangle$. The first statement naturally comes from the approximation of the concave function h. Indeed from this approximation, we can deduce that

$$g(\gamma^k) = \tilde{g}(\gamma^k; \gamma^k) \geq \tilde{g}(\gamma^{k+1}; \gamma^k) \geq g(\gamma^{k+1}) \tag{15}$$

which proves the first statement. For the second statement, by using continuity arguments and the first statement we get

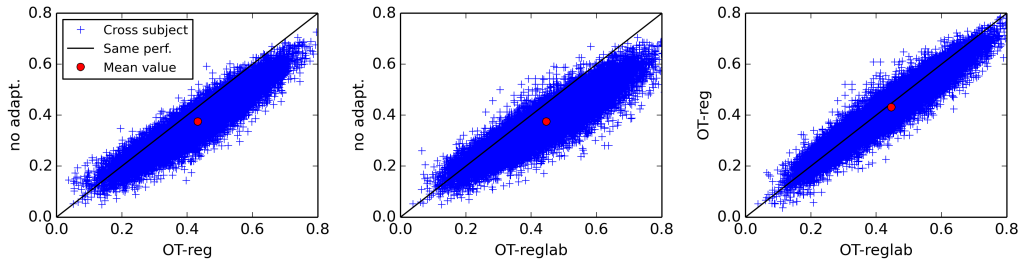
$$\tilde{g}(\gamma^\star; \gamma^\star) \leq \tilde{g}(\gamma; \gamma^\star) \quad \forall \gamma \in \mathcal{P}$$

Hence, since $\gamma^\star$ is a locally optimal for $\tilde{g}(\gamma; \gamma^\star)$, it has to satisfy the critical point condition of the above problem which is exactly Equation (14). Thus $\gamma^\star$ is also a critical point of problem (13).

5 Numerical experiments

5.1 OCR writer adaptation dataset

In order to illustrate the ability of optimal transport to perform domain adaptation, we apply our approach on an OCR problem. The dataset consists in 51935 images of handwritten letters (of size 8×16 pixels) with 26 classes (the letters of the alphabet) written by 158 different writers [28, 29]. By considering each writer as a domains, we study a large variety of inter-subjects adaptation problems (158^2-158=24806 source-target problems). Finally, note that the problem is extremely difficult due



(a) (b) (c)

Figure 2: Comparison of the accuracy of all OCR adaptation problems for different methods (**no adapt.**, **OT-reg** and **OT-reglab**). When the points are situated below the diagonal black line, it means that the performance of the method on the x-axis is better.

Method	Mean ACC	p-value
no adapt.	0.378	$< 10^{-300}$
OT-reg	0.434	$< 10^{-300}$
OT-reglab	0.448	-

Table 1: Mean accuracy and p-value for a Wilcoxon Signrank test comparing the methods to the best performing OT Class on the OCR dataset.

to the small number of samples per subject and the high variability of the writing patterns between subjects.

As classifier, we use a KNN with $k = 5$ and the regularization parameters $\lambda = 10$ and $\eta = 10$, selected for a good average performance across all transport. Results in Table 1 show the highest performances for the class regularized optimal transport (**OT-reglab**). The average improvement is also confirmed statistically with a Wilcoxon signrank test. The extremely small value of the p-value is due to the large number of performance sampling.

Fig 2 illustrates the single pairwise adaptation problems for the three methods considered: for each pair of methods, we report the accuracy of one approach as a function of the accuracy of another one. Interestingly, the figure shows that , on average, both **OT-reg** and **OT-reglab** perform better than a direct classification, while **OT-reglab** performs slightly better than the transport with Sinkhorn regularization only (**OT-reg**). Moreover, we can also observe that stronger gains are obtained for easier classification problem (upper right corner of the graphs). This can be explained by the fact that easier classification problems require a lighter adaptation and the probability of having a permutation of the classes along transport is smaller.

5.2 Visual adaptation dataset

We now evaluate our method on a challenging real world dataset coming from the computer vision community[1]. The dataset tackles a visual recognition task of several categories of objects, studied in the following papers [16, 17, 18]. The dataset contains images coming from four different domains: *Amazon* (online merchant), the *Caltech-256* image collection [30], *Webcam* (images taken from a webcam) and *DSLR* (images taken from a high resolution digital SLR camera). The domains are respectively noted in the remainder as A, C, W and D. As preprocessing, SURF description were computed, which allows to transform each image into a 800 bins histogram, subsequently normalized and reduced to standard scores. We followed the experimental protocol exposed in [17]: each dataset is considered in turn as the source domain and used to predict the others. Within those datasets, 10 classes of interest are extracted. The source domain are formed by picking 20 elements per class for domains A,C and W, and 8 for D. The training set is then formed by adapting these samples to the target domain. The latter is composed of all the elements in the test domain. The classification is conducted using a 1-NN classifier, which avoids cross-validation of hyper-parameters.

[1]Results reported from [6].

	Methods							
	without labels						with	
	no adapt.	**SuK** [31]	**SGF** [16]	**GFK** [17]	**OT-ori**	**OT-reg**	**GFK-lab** [17]	**OT-reglab**
C→A	20.8 ± 0.4	32.1 ± 1.7	36.8 ± 0.5	36.9 ± 0.4	30.6 ± 1.6	41.2 ± 2.9	40.4 ± 0.7	$\mathbf{43.5 \pm 2.1}$
C→D	22.0 ± 0.6	31.8 ± 2.7	32.6 ± 0.7	35.2 ± 1.0	27.7 ± 3.7	36.0 ± 4.1	$\mathbf{41.1 \pm 1.3}$	$\mathbf{41.8 \pm 2.8}$
A→C	22.6 ± 0.3	29.5 ± 1.9	35.3 ± 0.5	35.6 ± 0.4	30.1 ± 1.2	32.6 ± 1.3	$\mathbf{37.9 \pm 0.4}$	35.2 ± 0.8
A→W	23.5 ± 0.6	26.7 ± 1.9	31.0 ± 0.7	34.4 ± 0.9	28.0 ± 2.0	34.7 ± 6.3	35.7 ± 0.9	$\mathbf{38.4 \pm 5.4}$
W→C	16.1 ± 0.4	24.2 ± 0.9	21.7 ± 0.4	27.2 ± 0.5	26.7 ± 2.3	32.8 ± 1.2	29.3 ± 0.4	$\mathbf{35.5 \pm 0.9}$
W→A	20.7 ± 0.6	26.7 ± 1.1	27.5 ± 0.5	31.1 ± 0.7	29.0 ± 1.2	38.7 ± 0.7	35.5 ± 0.7	$\mathbf{40.0 \pm 1.0}$
D→A	27.7 ± 0.4	28.8 ± 1.5	32.0 ± 0.4	32.5 ± 0.5	29.2 ± 0.8	32.5 ± 0.9	$\mathbf{36.1 \pm 0.4}$	34.9 ± 1.3
D→W	53.1 ± 0.6	71.5 ± 2.1	66.0 ± 0.5	74.9 ± 0.6	69.8 ± 2.0	81.5 ± 1.0	79.1 ± 0.7	$\mathbf{84.2 \pm 1.0}$
mean	25.8	33.9	35.4	38.5	33.9	41.3	41.9	**44.2**

Table 2: Overall recognition accuracies in % and standard deviation on the domain adaptation of visual features

We repeat each experiment 20 times and report the overall classification accuracy and the associated standard deviation. We compare the results of the three transport models (**OT-ori**, **OT-reg** and **OT-reglab**) against both a classification conducted without adaptation (**no adapt.**) and 3 state-of-the-art methods: 1) the surrogate kernel approach (**SuK**), which in [31] was shown to outperform both the Transfer Component Analysis method [13] and the samples reweighing scheme of [32]; 2) the (**SGF**) method proposed in [16] and 3) the Geodesic Flow Kernel (**GFK**) approach proposed in [17]. Note that this last method can also efficiently incorporate label information: therefore we make a distinctions between methods, which do not incorporate label information (**no adapt**, **SuK**, **SGF**, **GFK**, **OT-ori** and **OT-reg**) and those that do (**GFK-lab** and **OT-reglab**). For each setting we used the recommended parameters to tune the competing methods. Results are reported in Table. 5.2.

When no label information is used, (**OT-reg**) usually performs best. In some cases (notably when considering the adaptation from (W→A or D→W), it can even surpass the (**GFK-lab**) method, which uses labels information. **OT-ori** usually enhances the result obtained without adaptation, but remains less accurate than the competing methods (except in the case of W→A where it surpasses **SGF** and **SuK**. Among all the methods, **OT-reglab** usually performs best, and with a significant increase in the classification performances for some cases (W→C or D→W). Yet, our method does not reach state-of-the-art performance in two cases: A→C and D→A. Finally, the overall mean value (last line of the table) shows a consistent increase of the performances with the proposed **OT-reglab**, which outperforms in average **GFK-lab** by 2%. Also note that the regularized unsupervised version **OT-reg** outperforms all the competing methods by at least 3%.

6 Conclusion and discussion

We have presented a new method for unsupervised domain adaptation based on the optimal transport of discrete distributions from a source to a target domain. While the classical optimal transport provide satisfying results, it fails in some cases to provide state-of-the-art performances in the tested classification approaches. We proposed to regularize the transport by relaxing some sparsity constraints in the probabilistic coupling of the source and target distributions, and to incorporate the label information by penalizing couplings that would mix samples issued from different classes. This was performed by a Majoration Minimization strategy that exploits a $\ell_p - \ell_1$ norm, which promotes sparsity among the different classes. The corresponding algorithm is fast, and allows to work efficiently with sets of several thousand samples. With this regularization, competitive results were achieved on challenging domain adaptation datasets thanks to the ability of our approach to express both class relationship and non-linear transformations of the domains.

Possible improvements of our work are numerous, and include: *i)* extension to a multi-domain setting, by finding simultaneously the best minimal transport among several domains, *ii)* extension to semi-supervised problems, where several unlabelled samples in the source domain, or labelled samples in the target domain are also available. In this last case, the group sparsity constraint should not only operate over the columns but also the lines of the coupling matrix, which makes the underlying optimization problem challenging. *iii)* Definition of the transport in a RKHS, in order to exploit the manifold structure of the data.

Acknowledgements This work has been partially funded by a visiting professor grant from EPFL, and by the Swiss National Science Foundation (grants 136827 and 150593).

References

[1] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proceedings of the 11th European conference on Computer vision: Part IV*, ECCV'10, pages 213–226, Berlin, Heidelberg, 2010. Springer-Verlag.

[2] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010.

[3] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT'09*, pages 19–30, 2009.

[4] S. Ben-David, T. Lu, T. Luu, and D. Pl. Impossibility theorems for domain adaptation. In *AISTATS*, pages 129–136, 2010.

[5] P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers. In *ICML*, pages 738–746, Atlanta, USA, 2013.

[6] N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2014.

[7] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: domain adaptation using asymmetric kernel transforms. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, 2011.

[8] I-H. Jhuo, D. Liu, D. T. Lee, and S.-F. Chang. Robust visual domain adaptation with low-rank reconstruction. In *CVPR*, pages 2168–2175, 2012.

[9] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell. Efficient learning of domain invariant image representations. In *ICLR*, 2013.

[10] Jihun Ham, Daniel Lee, and Lawrence Saul. Semisupervised alignment of manifolds. In Robert G. Cowell and Zoubin Ghahramani, editors, *10th International Workshop on Artificial Intelligence and Statistics*, pages 120–127, 2005.

[11] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*, pages 1541–1546. AAAI Press, 2011.

[12] H. Daumé III. Frustratingly easy domain adaptation. In *Ann. Meeting of the Assoc. Computational Linguistics*, 2007.

[13] S. J. Pan and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Networks*, 22:199–210, 2011.

[14] A. Kumar, H. Daumé III, and D. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012.

[15] C. Wang and S. Mahadevan. Manifold alignment without correspondence. In *International Joint Conference on Artificial Intelligence*, Pasadena, CA, 2009.

[16] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, pages 999–1006. IEEE, 2011.

[17] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073. IEEE, 2012.

[18] J. Zheng, M.-Y. Liu, R. Chellappa, and P.J. Phillips. A grassmann manifold-based domain adaptation approach. In *ICPR*, pages 2095–2099, Nov 2012.

[19] T. S. Caetano, T. Caelli, D. Schuurmans, and D.A.C. Barone. Grapihcal models and point pattern matching. *IEEE Trans. Pattern Anal. Mach, Intell.*, 28(10):1646–1663, 2006.

[20] T. S Caetano, J. J. McAuley, L. Cheng, Q. V. Le, and A. J. Smola. Learning graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(6):1048–1058, 2009.

[21] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, May 2010.

[22] L. Kantorovich. On the translocation of masses. *C.R. (Doklady) Acad. Sci. URSS (N.S.)*, 37:199–201, 1942.

[23] S. Ferradans, N. Papadakis, J. Rabin, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. In *Scale Space and Variational Methods in Computer Vision, SSVM*, pages 428–439, 2013.

[24] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation. In *NIPS*, pages 2292–2300. 2013.

[25] P. Knight. The sinkhorn-knopp algorithm: Convergence and applications. *SIAM J. Matrix Anal. Appl.*, 30(1):261–275, March 2008.

[26] E.J. Candes, M.B. Wakin, and S.P. Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

[27] D.R. Hunter and K. Lange. A Tutorial on MM Algorithms. *The American Statistician*, 58(1):30–38, 2004.

[28] Guillaume Obozinski, Ben Taskar, and Michael I Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.

[29] A. Rakotomamonjy, R. Flamary, G. Gasso, and S. Canu. lp-lq penalty for sparse linear and sparse multiple kernel multi-task learning. *IEEE Transactions on Neural Networks*, 22(8):1307–1320, 2011.

[30] G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. Technical Report CNS-TR-2007-001, California Institute of Technology, 2007.

[31] K. Zhang, V. W. Zheng, Q. Wang, J. T. Kwok, Q. Yang, and I. Marsic. Covariate shift in Hilbert space: A solution via surrogate kernels. In *ICML*, 2013.

[32] M. Sugiyama, S. Nakajima, H. Kashima, P.V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, 2008.