Domain adaptation with regularized optimal transport

Nicolas Courty[1], Rémi Flamary[2], and Devis Tuia[3]

[1] Université de Bretagne Sud, IRISA,Vannes, France
[2] Université de Nice, Lab. Lagrance UMR CNRS 7293, OCA, Nice, France
[3] EPFL, LASIG, Lausanne, Switzerland

Abstract. We present a new and original method to solve the domain adaptation problem using optimal transport. By searching for the best transportation plan between the probability distribution functions of a source and a target domain, a non-linear and invertible transformation of the learning samples can be estimated. Any standard machine learning method can then be applied on the transformed set, which makes our method very generic. We propose a new optimal transport algorithm that incorporates label information in the optimization: this is achieved by combining an efficient matrix scaling technique together with a majoration of a non-convex regularization term. By using the proposed optimal transport with label regularization, we obtain significant increase in performance compared to the original transport solution. The proposed algorithm is computationally efficient and effective, as illustrated by its evaluation on a toy example and a challenging real life vision dataset, against which it achieves competitive results with respect to state-of-the-art methods.

1 Introduction

While most learning methods assume that the test data $\mathbf{X}_t = (\mathbf{x}_i^t)_{i=1,\dots,N_t}$, $\mathbf{x}_i \in \mathbb{R}^d$ and the training data $\mathbf{X}_s = (\mathbf{x}_i^s)_{i=1,\dots,N_s}$ are generated from the same distributions $\mu_t = \mathcal{P}(\mathbf{X}_t)$ and $\mu_s = \mathcal{P}(\mathbf{X}_s)$, real life data often exhibit different behaviors. Many 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 [1]. In our work, we assume that the source and target domains are by nature different, which is usually referred to as domain adaptation. In the classification problem, the training data are usually associated with labels corresponding to C different classes. We consider the case where only the training data are associated with a label $\mathbf{Y}_s = (\mathbf{y}_i)_{i=1,\dots,N_s}$, $\mathbf{y}_i \in \{1,\dots,C\}$, yielding an **unsupervised domain adaptation** problem, since no labelled data is available in the target domain. In this acceptation, the training (resp. testing) domain is usually referred to as source (resp. target) distribution.

Domain adaptation methods seek to compensate for inter domain differences by exploiting the similarities between the two distributions. This compensation is usually performed by reweighing the contribution of each samples in the learning

process (*e.g.* [2]) or by means of a global data transformation that aligns the two distributions in some common feature space (*e.g.* [3]). Our work departs from these previous works by assuming that there exists a non-rigid transformation of the distribution that can account for the non-linear transformations occurring between the source and target domains. This transformation is conveniently expressed as a transportation of the underlying probability distribution functions thanks to optimal transport (OT). The OT problem has first been introduced by the French mathematician Gaspard Monge in the middle of the 19th century as the way to find a minimal effort solution to the transport of a given mass of dirt into a given hole. The problem reappeared later in the work of Kantorovitch [4], and found recently surprising new developments as a polyvalent tool for several fundamental problems [5]. In the domain of machine learning, OT has been recently used for computing distances between histograms [6] or label propagation in graphs [7].

Contributions Our contributions are twofold: *i)* First, we show how to transpose the optimal transport problem to the domain adaptation problem, and we propose experimental validations of this idea. To the best of our knowledge, this is the first time that optimal transport is considered in the domain adaptation setting. *ii)* Second, we propose an elegant group-based regularization for integrating label information, which has the effect of regularizing the transport by adding inter-class penalties. The resulting algorithm exploits a proven efficient optimization approaches and will benefit from any advances in this domain. The proposed optimal transport with label regularization (**OT-reglab**) allows to achieve competitive state-of-the-art results on challenging datasets.

2 Related Work

Two main strategies have been considered to tackle the domain adaptation problem: on the one hand, there are approaches considering the transfer of instances, mostly via sample re-weighting schemes based on density ratios between the source and target domains [2,8]. By doing so, authors compare the data distributions in the input space and try to make them more similar by weighting the samples in the source domain.

On the other hand, many works have considered finding a common feature representation for the two (or more) domains, or a latent space, where a classifier using only the labeled samples from the source domain generalize well on the target domains [9,10]. The representation transfer can be performed by matching the means of the domains in the feature space [10], aligning the domains by their correlations [11] or by using pairwise constraints [12]. 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. In computer vision, methods exploiting a gradual alignment of sets of eigenvectors have been proposed: in [13], authors start from the hypothesis that domain adaptation can

be better approached if comparing gradual distortions and therefore use intermediary projections of both domains along the Grassmannian geodesic connecting the source and target observed eigenvectors. In [14,15], 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. 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 is global and applied the same way to the whole target domain.

An approach combining the two logics is found in [3], where authors extend the sample re-weighing reasoning to similarity of the distributions in the feature space by the use of surrogate kernels. By doing so, a linear transformation of the domains is found, but, as for the feature representation approaches above, it is the same for all samples transferred.

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 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 problem

3 Optimal transportation

In this Section, we introduce the original formulation of optimal transport through the Monge-Kantorovitch problem and its discrete formulation. Then, regularized versions of the optimal transport are exposed.

3.1 The Monge-Kantorovitch problem and Wasserstein space

Let us first consider two domains Ω_1 and Ω_2 (in the following, we will assume without further indication that $\Omega_1 = \Omega_2 = \mathbb{R}^d$). Let $\mathcal{P}(\Omega_i)$ be the set of all the probability measures over Ω_i. Let $\mu \in \mathcal{P}(\Omega_1)$, and $\mathbf{T}$ be an application from $\Omega_1 \to \Omega_2$. The *image measure* of μ by $\mathbf{T}$, noted $\mathbf{T}\#\mu$, is a probability measure over Ω_2 which verifies:

$$\mathbf{T}\#\mu(y) = \mu(\mathbf{T}^{-1}(\mathbf{y})), \quad \forall \mathbf{y} \in \Omega_2. \tag{1}$$

Let $\mu_s = \mathcal{P}(\Omega_1)$ and $\mu_t = \mathcal{P}(\Omega_2)$ be two probability measures from the two domains. $\mathbf{T}$ is said to be a transport if $\mathbf{T}\#\mu_s = \mu_t$. The cost associated to this transport is

$$C(\mathbf{T}) = \int_{\Omega_1} c(\mathbf{x}, \mathbf{T}(\mathbf{x}))d\mu(\mathbf{x}), \tag{2}$$

where the cost function $c : \Omega_1 \times \Omega_2 \to \mathbb{R}^+$ can be understood as a regular distance function, but also as the energy required to move a mass $\mu(\mathbf{x})$ from $\mathbf{x}$

to $\mathbf{y}$. It is now possible to define the **optimal transport** $\mathbf{T}_0$ as the solution of the following minimization problem:

$$\mathbf{T}_0 = \arg\min_{\mathbf{T}} \int_{\Omega_1} c(\mathbf{x}, \mathbf{T}(\mathbf{x}))d\mu(\mathbf{x}), \quad \text{s.t. } \mathbf{T}\#\mu_s = \mu_t \tag{3}$$

which is the original Monge transportation problem. The equivalent Kantorovitch formulation of the optimal transport [4] seeks for a probabilistic coupling $\boldsymbol{\gamma} \in \mathcal{P}(\Omega_1 \times \Omega_2)$ between Ω_1 and Ω_2:

$$\boldsymbol{\gamma}_0 = \arg\min_{\boldsymbol{\gamma}} \int_{\Omega_1 \times \Omega_2} c(\mathbf{x}, \mathbf{y})d\boldsymbol{\gamma}(\mathbf{x}, \mathbf{y}), \quad \text{s.t. } \mathrm{P}^{\Omega_1}\#\boldsymbol{\gamma} = \mu_s, \mathrm{P}^{\Omega_2}\#\boldsymbol{\gamma} = \mu_t, \tag{4}$$

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. It allows to define the **Wasserstein distance** between μ_s and μ_t as:

$$\mathbf{W}_2(\mu_s, \mu_t) = \inf_{\gamma} \int_{\Omega_1 \times \Omega_2} c(\mathbf{x}, \mathbf{y})d\boldsymbol{\gamma}(\mathbf{x}, \mathbf{y}), \text{ s.t. } \mathrm{P}^{\Omega_1}\#\boldsymbol{\gamma} = \mu_s, \mathrm{P}^{\Omega_2}\#\boldsymbol{\gamma} = \mu_t, \tag{5}$$

This distance, also known as the Earth Mover Distance in computer vision community [16], defines a metric over the space of integrable squared probability measure.

3.2 Optimal transport of discrete distributions

Usually one does not have a direct access to μ_s or μ_t but rather to collections of samples from those distributions. It is then straightforward to adapt the optimal transport problem to the discrete case. The two distributions can be written as

$$\mu_s = \sum_{i=1}^{n_s} p_i^s \delta_{\mathbf{x}_i^s}, \quad \mu_t = \sum_{i=1}^{n_t} p_i^t \delta_{\mathbf{x}_i^t} \tag{6}$$

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 belong 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 then the set of doubly stochastic matrices $\mathcal{P}$ defined as

$$\mathcal{P} = \left\{\boldsymbol{\gamma} \in (\mathbb{R}^+)^{n_s \times n_t} | \, \boldsymbol{\gamma}\mathbf{1}_{n_t} = \mu_s, \boldsymbol{\gamma}^T\mathbf{1}_{n_s} = \mu_t\right\} \tag{7}$$

where $\mathbf{1}_d$ is a d-dimensional vector of ones. The Kantorovitch formulation of the optimal transport [4] reads:

$$\boldsymbol{\gamma}_0 = \arg\min_{\boldsymbol{\gamma} \in \mathcal{P}} \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F \tag{8}$$

where $\langle ., . \rangle_F$ is the Frobenius dot product and $\mathbf{C} \geq 0$ is the cost function matrix of term $C(i, j)$ related to the energy needed to move a probability mass from $\mathbf{x}_i^s$ to $\mathbf{x}_j^t$. This cost can be chosen for instance as the Euclidian distance between the two locations, i.e. $C(i, j) = ||\mathbf{x}_i^s - \mathbf{x}_j^t||^2$, but other types of metric could be considered, such as Riemannian distances over a manifold [5].

Remark 1. When $n_s = n_t = n$ and when $\forall i, j \;\; p_i^s = p_j^t = 1/n$, the $\boldsymbol{\gamma}_0$ is simply a permutation matrix

Remark 2. In the general case, it can be shown that $\boldsymbol{\gamma}_0$ is a sparse matrix with at most $n_s + n_t - 1$ non zero entries (rank of constraints matrix).

This problem can be solved by linear programming, with combinatorial algorithms such as the simplex methods and its network variants (transport simplex, network simplex, etc.). Yet, the computational complexity was shown to be $\mathcal{O}(n^2)$ in practical situations [17] for the network simplex (while being $\mathcal{O}(n^3)$ in theory) which leverages the utility of the method to handle big data. However, the recent regularization of Cuturi [6] allows a very fast transport computation as discussed in the next Section.

3.3 Regularized optimal transport

When the target and source distributions are high-dimensional, or even in presence of numerous outliers, the optimal transportation plan may exhibit some irregularities, and lead to incorrect transport of points. 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.(8). This possibility has been explored in recent papers [18,6].

In [18], Ferradans and colleagues have explored the possibility of relaxing the mass conservation constraints of the transport, *i.e.* slightly distorting the marginals of the coupling $\boldsymbol{\gamma}_0$. Technically, this boils down to solving the same minimization problem but with inequality constraints on the marginals in Eq.(7). As a result, elements of the source and target distributions can remain still. Yet, one major problem of this approach is that it converts the original linear program into more computationally demanding optimizations impractical for large sets.

In a recent paper [6], 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:

$$\boldsymbol{\gamma}_0^\lambda = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\arg\min} \, \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F - \frac{1}{\lambda} h(\boldsymbol{\gamma}), \tag{9}$$

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-nul 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 [19]. The optimal regularized transportation plan is found by iteratively computing two scaling vectors $\mathbf{u}$ and $\mathbf{v}$ such that:

$$\boldsymbol{\gamma}_0^{\lambda} = \text{diag}(\mathbf{u}) \exp(-\lambda \mathbf{C}) \text{diag}(\mathbf{v}), \tag{10}$$

where the exponential exp(.) operator should be understood element-wise.

Note that while these regularizations allow the inclusion of additional priors in the optimization problem, they do not take into account the fact that the elements of the source distribution belong to different classes. This idea is the core of our regularization strategy.

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 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 derive a new 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

Optimal transport aims at minimizing a transport cost linked to a metric between distributions. It 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 adaptation. Our proposition to take advantage of label information is to penalize couplings that match together samples with different labels. This is illustrated in Figure 1.c, where one can see that 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 Over each column of $\boldsymbol{\gamma}$, we want to concentrate the transport information on elements of the same class c. This is usually done by using $\ell_p - \ell_q$ mixed-norm regularization, among which the $\ell_1 - \ell_2$ known as as group-lasso is a favorite. The main idea is that, even if we do not know the class of 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.

Promoting group sparsity leads to a new term in the cost function (9), which now reads:

$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\arg\min} \left\langle \boldsymbol{\gamma}, \mathbf{C} \right\rangle_F - \frac{1}{\lambda} h(\boldsymbol{\gamma}) + \eta \sum_j \sum_c ||\boldsymbol{\gamma}(\mathcal{I}_c, j)||_q^p, \tag{11}$$

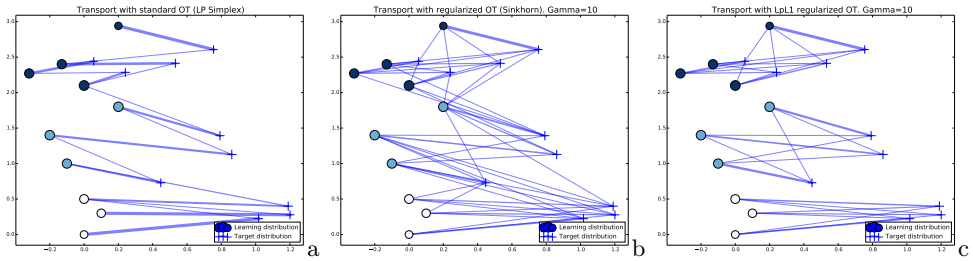

Fig. 1. Illustration of the transport for two simple distributions depicted in the image. The colored disks represent 3 different classes. The transport solution is depicted as blue lines whose thickness relate to the strength of the coupling. (a) Solution of the original optimal transport solution (**OT-ori**); (b) using the Sinkhorn transport (**OT-reg** [6]); (c) using our class-wise regularization term (**OT-reglab**).

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

The choice of the p, q parameters is particularly sensitive. For $p \geq 1$ and $q \geq 1$ the regularization term is convex. The parameters $p = 1, q = 2$ lead to the classical group-lasso that is used, for instance, for joint features selection in multitask learning. The main problem of using the group-lasso in this case is that it makes the optimization problem much more difficult. Indeed, when using an ℓ_2 norm in the objective function, the efficient optimization procedure proposed in [6] cannot be used anymore. Moreover there is no particular reason to choose the ℓ_2 norm for regularizing coefficients of a transport matrix. Those coefficients being all positive and associated to probabilities, we propose to use $q = 1$ that will basically sum the probabilities in the groups. When $q = 1$, one needs to carefully chose the p coefficient in order to promote group sparsity. In this work we propose to use $p = 1/2 < 1$. This parameter is a common choice for promoting sparsity, as the square root is non-differentiable in zero and has been used recently for promoting non-grouped sparsity in compressed sensing [20].An additional advantage of our proposal is that, despite the fact that the proposed regularization is non-convex, a simple approach known as reweighted ℓ_1 can be performed for its optimization, as detailed below.

4.2 Majoration Minimization strategy

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

$$\boldsymbol{\gamma}_0 = \arg\min_{\boldsymbol{\gamma} \in \mathcal{P}} J(\boldsymbol{\gamma}) + \eta \Omega(\boldsymbol{\gamma}), \tag{12}$$

with $J(\boldsymbol{\gamma}) = \langle \boldsymbol{\gamma}, C \rangle_F - \frac{1}{\lambda} h(\boldsymbol{\gamma})$ and $\Omega(\boldsymbol{\gamma}) = \sum_j \sum_c ||\boldsymbol{\gamma}(\mathcal{I}_c, j)||_1^p$. We want to be able to use the optimization in [6] to solve the left term, as it is very efficient.

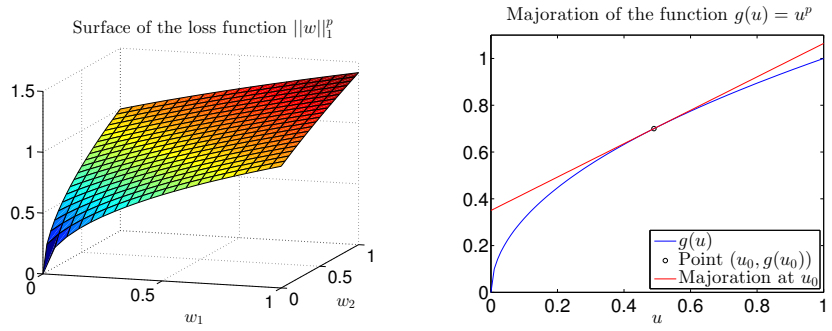

Fig. 2. Illustration of the regularization term loss for a 2D group (left). Illustration of the convexity of $g(\cdot)$ and its linear majoration (right)

First, note that the regularization term can be reformulated as

$$\Omega(\boldsymbol{\gamma}) = \sum_j \sum_c g(||\boldsymbol{\gamma}(\mathcal{I}_c, j)||_1) \tag{13}$$

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 [21]. This can be done because the $\ell_p - \ell_1$ regularization term is concave in the positive orthant as illustrated in the left part of figure 2. It is clear from this Figure that the surface can be majored by an hyperplane. For a given group of variable, one can use the concavity of g to majorize it around a given vector $\hat{\mathbf{w}} > 0$

$$g(\mathbf{w}) \leq g(\|\hat{\mathbf{w}}\|_1) + \nabla g(\|\hat{\mathbf{w}}\|_1)^\top (\mathbf{w} - \hat{\mathbf{w}}) \tag{14}$$

with $\nabla g(\|\hat{\mathbf{w}}\|_1) = p(\|\hat{\mathbf{w}}\|_1)^{p-1}$ for $\hat{\mathbf{w}} > 0$. An illustration of the majoration of $g(\cdot)$ can be seen in the right part of Figure 2. For each group, the regularization term can be majorized by a linear approximation. In other words, for a fixed $\hat{\boldsymbol{\gamma}}$

$$\Omega(\boldsymbol{\gamma}) \leq \tilde{\Omega}(\boldsymbol{\gamma}) = \langle \boldsymbol{\gamma}, \mathbf{G} \rangle_F + cst \tag{15}$$

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

$$\mathbf{G}(\mathcal{I}_c, j) = p(\|\hat{\boldsymbol{\gamma}}(\mathcal{I}_c, j)\| + \epsilon)^{p-1}, \quad \forall c, j \tag{16}$$

Note that we added a small $\epsilon > 0$ that helps avoiding numerical instabilities, as discussed in [20]. Finally, solving problem (11) 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. Moreover this kind of MM algorithm is known to converge in a small number of iterations.

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

Initialize $\mathbf{G} = \mathbf{0}$
Initialize $\mathbf{C}_0$ as in Equation (8)
repeat
  $\mathbf{C} \leftarrow \mathbf{C}_0 + \mathbf{G}$
  $\boldsymbol{\gamma} \leftarrow$ Solve problem (9) with $\mathbf{C}$
  $\mathbf{G} \leftarrow$ Update $\mathbf{G}$ with Equation (16)
until Convergence

4.3 Interpolation of the source domain

Once the transport $\boldsymbol{\gamma}_0$ has been defined using either Equations (8), (9) or (11), the source samples must be transported in the target domain using their transportation plan. One can seek the interpolation of the two distributions by following the geodesics of the Wasserstein metric [5] (parameterized by t). This allows to define a new distribution μ_t such that:

$$\mu_t = \arg\min_{\mu} (1-t) W_2(\mu_s, \mu)^2 + t W_2(\mu_t, \mu)^2. \tag{17}$$

One can show that this distribution is:

$$\mu_t = \sum_{i,j} \boldsymbol{\gamma}_0(i,j) \delta_{(1-t)\mathbf{x}_i^s + t\mathbf{x}_j^t}. \tag{18}$$

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 barycenters of the 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 target domain dependent version $\hat{\mathbf{X}}_s$. The mapping $\mathbf{T}_{\boldsymbol{\gamma}_0}$ can be conveniently expressed as:

$$\hat{\mathbf{X}}_s = \mathbf{T}_{\boldsymbol{\gamma}_0}(\mathbf{X}_s) = \mathrm{diag}((\boldsymbol{\gamma}_0 \mathbf{1}_{n_t})^{-1}) \boldsymbol{\gamma}_0 \mathbf{X}_t. \tag{19}$$

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 by observing that $\mathbf{T}_{\boldsymbol{\gamma}_0}^{-1} = \mathbf{T}_{\boldsymbol{\gamma}_0^T}$. Let us finally remark that similar interpolation methods were used in the domain of color transfer [18].

5 Experimental validation

In this Section, we validate the proposed algorithm in two domain adaptation examples. On the first one, we study the behavior of our approach on a simple toy dataset. The second one considers a challenging computer vision dataset, used for a comparison with state-of-the-art methods. In every experiment, the original optimal transport (**OT-ori**) is computed with a network simplex approach [17].

The Sinkhorn transport, which corresponds to the regularized version of the optimal transport (**OT-reg**) described in Section 3.3, was implemented following the algorithm proposed in [6]. Our approach, **OT-reglab**, follows the Algorithm 1. As expected, these last two methods are generally one order of magnitude faster than the network simplex approach.
As for the choice of the weights of Eq. (6), the problem can be cast as an estimation of a probability mass function of a discrete variable on the sample space of the source and target distributions. A direct and reasonable choice is to take an uniform distribution, *i.e.* $p_i^s = \frac{1}{n_s}$ and $p_t^s = \frac{1}{n_t}$. This choice gives the same value for every samples in the two discrete distributions. Alternatively, one can seek to strengthen the weights of samples that are in a high density region, and lower weights for samples in low density regions. This way, outliers should be associated with lower masses. A possible solution relies on a discrete variant of the Nadaraya-Watson estimator [22] where one enforces the sum-to-1 property:

$$p_i^s = \frac{\sum_{j=1}^{n_s} k_\sigma(\mathbf{x}_i^s, \mathbf{x}_j^s)}{\sum_{j=1}^{n_s}\sum_{i=1}^{n_s} k_\sigma(\mathbf{x}_i^s, \mathbf{x}_j^s)} \tag{20}$$

where $k_\sigma(\cdot,\cdot)$ is a gaussian kernel of bandwidth σ. The drawback of such an estimator is that it adds an hyper parameter to the method. Yet, while standard approaches [22] can be used to estimate this parameter, we observed in our experiments, and for large number of samples, that this parameter exerts little influence over the final result (less than a standard deviation) for a large range of values.

5.1 Toy dataset

In this first experiment, the behavior of the optimal transport is examined on a simple two-dimensional dataset. We consider a two-class distribution by sampling independently for each class $c1$ and $c2$ following the normal distributions $\mathcal{N}_1^s$ and $\mathcal{N}_2^s$. The set of all those samples constitute the source domain. The target domain samples are then obtained by sampling the mixture $\mathcal{N}_1^t + \mathcal{N}_2^t$. The target distributions $\mathcal{N}_i^t,\ (i = 1, 2)$ are deduced from $\mathcal{N}_i^s,\ (i = 1, 2)$ by changing both the scale and translating the distribution mean. The produced domain transformation is thus non-linear and cannot be expressed by a simple $2D$ transformation of the input space. This makes the problem particularly interesting with respect to our initial assumptions on the nature of the domain change. We then sample randomly from these distributions n_1^s,n_2^s,n_1^t and n_2^t samples from $\mathcal{N}_1^s$, $\mathcal{N}_2^s$, $\mathcal{N}_1^t$ and $\mathcal{N}_2^t$ to form the corresponding learning and test sets. An illustration of this toy dataset is given in Figure 3.a for $n_1^s + n_2^s = 100$ samples in the source distribution (red and white circles) and $n_1^t + n_2^t = 200$ samples in the target one (blue crosses). Note that the size of the points in the Figure is proportional to its weight p_i and reflects the density of the distribution.

Figure 3.b presents the result of the optimal transport **OT-ori** coupling as a set of non-nul connections (red and black arcs) between the source and the target distributions. The color of those connections is related to the magnitude

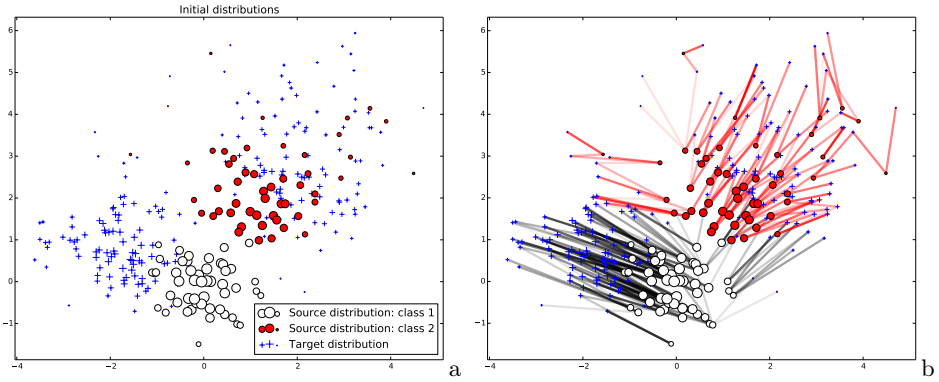

Fig. 3. Illustration of the transport **OT-ori** on a simple toy dataset. The initial distributions are depicted in the right image (a) The source distribution is depicted in white and red for respectively class 1 and 2, the target distributions are in blue. In image (b), we show the optimal transport couplings, depicted as links colored with respect to the source class label.

of the coupling (up to a global scaling factor). As expected, the coupling matrix $\boldsymbol{\gamma}_0$ contains less than $100 + 200 - 1 = 299$ non-nul entries. One can see that some white and red elements are clearly misled by the transport, but the overall adaptation remains coherent with the test distribution.

Figure 4 illustrates the results obtained on this dataset with the regularized versions of the transport **OT-reg** and **OT-reglab** for a regularization parameter value of $\lambda = 1$. The $\boldsymbol{\gamma}_0$ matrix of **OT-reg**, on the left of the first row of Figure 4, is indeed sparse, but much less than the corresponding one in **OT-ori**. This can be assessed by comparing the denser connections issued from **OT-reg** (left panel of the second row of Figure 4) with respect to those observed for **OT-ori** (right panel of Figure 3). In the proposed **OT-reglab** (right column of Figure 4), the sparsity is clearly enforced per class (the rows of the coupling matrix are sorted by class), which yields a sparser coupling matrix with block structure. In the last row of Figure 4 we show the result of the adaptation of the source distribution following the procedure described in Section 4.3. Two additional interesting behaviors are observed in the regions highlighted by red squares, where some of the incoherencies observed in **OT-ori** and **OT-reg** of the transport are resolved by the label regularization proposed with **OT-reglab**.

Classification measures. We now consider performances of a classifier trained on the source samples adapted to the target distribution. In those experiments, we use a SVM classifier with a Gaussian kernel. The hyperparameters of the classifier are computed for each trial by a 2-fold cross validation over a grid of potential values. For every setting considered, the data generation / adaptation / classification was conducted 20 times to leverage the importance of the sampling. When informative, we provide the standard deviation of the result. In the first experiment, we examine the importance of the regularization

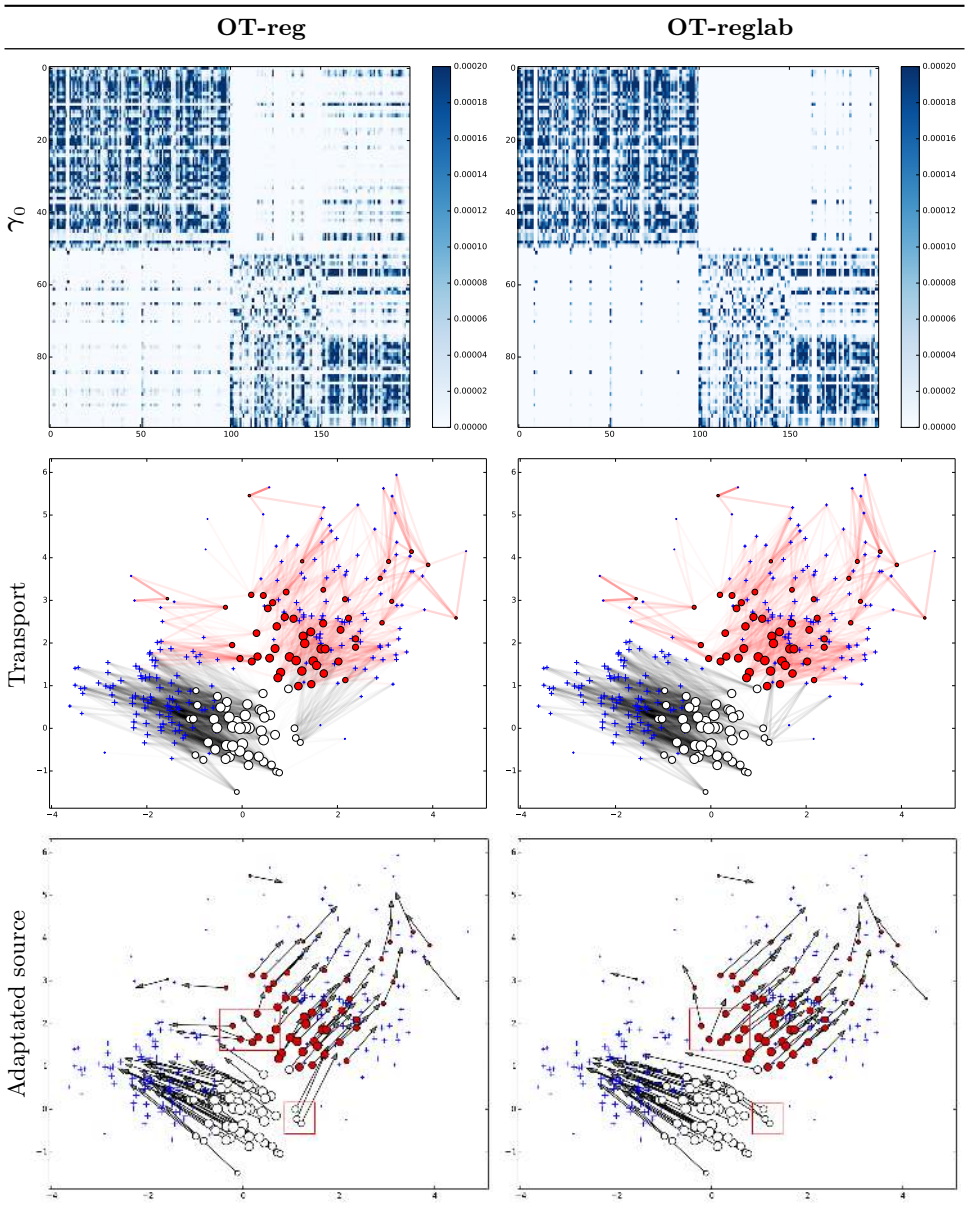

Fig. 4. Comparisons of two versions of the regularized transport: Sinkhorn transport (**OT-reg**, left column) and Sinkhorn transport with the label regularization (**OT-reglab**, right column). The first row shows the transport coupling matrices $\boldsymbol{\gamma}_0$, the second row their equivalent graphical representations, with connections colored by the source node label. The third row is the adaptation of the source samples induced by $\boldsymbol{\gamma}_0$ using Equation (19).

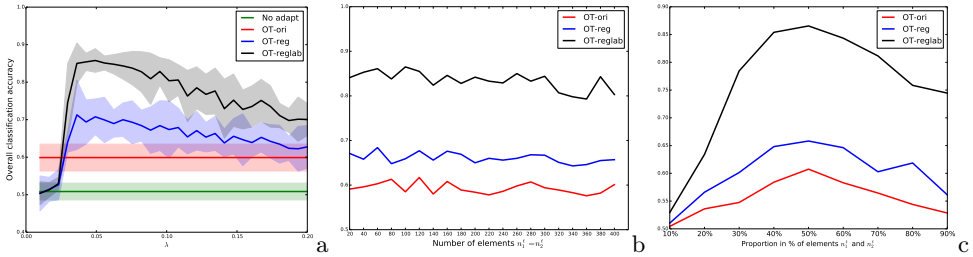

Fig. 5. Classification results for the toy dataset example: (a) influence of the regularization parameter λ; (b) influence of the proportions of samples between the source and the target distributions; (c) influence of the balance of classes on the overall performance of the adaptation.

parameter λ over the overall classification accuracy (Figure 5.a). In this case, we set $n_1^s = n_2^s = n_1^t = n_2^t = 100$. We confirm that the use of the transport for domain adaptation increases the performances significantly (by 8%) over a classification conducted directly with the source distribution as learning set. When varying the λ regularization parameter and using **OT-reglab**, another very significant increase is achieved (up to 25% for $\lambda = 0.04$), which demonstrates the relevance of our transport regularization. In the second experiment, we set $n_1^s = n_2^s = 100$ and we increase the number of elements in the source target n_1^t and n_2^t equivalently. For this experiment and for the next one, λ is set by a standard cross-validation method. In this case, the standard deviation is omitted as it is constant over the experiments and no informative. One can observe that the performances of the classification are *i)* consistent with the first experiment and *ii)* constant over the volume of samples in the target domain as long as the proportions are conserved. In the third experiment, we set n_1^t and n_2^t to the value of 100 samples each and we vary the proportion of the classes through a parameter $p \in [0, 1]$ with $n_1^s = p * 100$ and $n_2^s = (1 - p) * 100$. This parameter allows to control the proportion of elements in class 1 and in class 2 in the source distribution. As shown in Figure 5.c, the best result is achieved when the proportion of each class samples is similar in the source and target distributions (at 50%). This somehow highlights one limit of the method: the mass equivalent to each class should match in proportions for both distributions to get the best adaption result. Nevertheless, we can see from Figure 5.c that a variation of $\pm 15\%$ between the source and target distribution still leads to significant performance improvements.

5.2 Visual adaptation dataset

We now evaluate our method on a challenging real world dataset coming from the computer vision community. The objective is now a visual recognition task of several categories of objects, studied in the following papers [23,13,14,15]. The dataset contains images coming from four different domains: *Amazon* (online

merchant), the *Caltech-256* image collection [24], *Webcam* (images taken from a webcam) and *DSLR* (images taken from a high resolution digital SLR camera). Those domains are respectively noted in the remainder as A, C, W and D. A feature extraction method is used to preprocess those images; it namely consists in computing SURF descriptors [23], which allows to transform each image into a 800 bins histogram, which are then subsequently normalized and reduced to standard scores. We followed the experimental protocol exposed in [14]: 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-Nearest Neighbor classifier, which avoids cross-validation of hyper-parameters. As for the toy example above, 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 [3] was shown to outperform both the Transfer Component Analysis method [10] and the reweighing scheme of [2]; 2) the (**SGF**) method proposed in [13] and 3) the Geodesic Flow Kernel (**GFK**) approach proposed in [14]. 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. 1.

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 efficient 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 in this paper 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 re-

sults, 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 mix samples issued from different classes. This was made possible 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 unlabeled 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 We thank the anonymous reviewers for their critics and suggestions. This work has been partially funded by a visiting professor grant from EPFL, by the French ANR under reference ANR-13-JS02-0005-01 (Asterix project). and by the Swiss National Science Foundation (grant 136827, http://p3.snf.ch/project-136827).

References

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

Table 1. Overall recognition accuracies in % and standard deviation on the domain adaptation of visual features

	Methods							
	without labels						with	
	no adapt.	**SuK** [3]	**SGF** [13]	**GFK** [14]	**OT-ori**	**OT-reg**	**GFK-lab** [14]	**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	**43.5 ± 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	**41.1 ± 1.3**	**41.8 ± 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	**37.9 ± 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	**38.4 ± 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	**35.5 ± 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	**40.0 ± 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	**36.1 ± 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	**84.2 ± 1.0**
mean	25.8	33.9	35.4	38.5	33.9	41.3	41.9	**44.2**

2. 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.
3. 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.
4. L. Kantorovich. On the translocation of masses. *C.R. (Doklady) Acad. Sci. URSS (N.S.)*, 37:199–201, 1942.
5. C. Villani. *Optimal transport: old and new*. Grundlehren der mathematischen Wissenschaften. Springer, 2009.
6. M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation. In *NIPS*, pages 2292–2300. 2013.
7. Justin Solomon, Raif Rustamov, Guibas Leonidas, and Adrian Butscher. Wasserstein propagation for semi-supervised learning. In *Proceedings of The 31st International Conference on Machine Learning*, pages 306–314, 2014.
8. S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *ICML*, 2007.
9. H. Daumé III. Frustratingly easy domain adaptation. In *Ann. Meeting of the Assoc. Computational Linguistics*, 2007.
10. S. J. Pan and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Networks*, 22:199–210, 2011.
11. A. Kumar, H. Daumé III, and D. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012.
12. C. Wang and S. Mahadevan. Manifold alignment without correspondence. In *International Joint Conference on Artificial Intelligence*, Pasadena, CA, 2009.
13. R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, pages 999–1006. IEEE, 2011.
14. B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073. IEEE, 2012.
15. 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.
16. Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *ICCV*, pages 59–66, Jan 1998.
17. N. Bonneel, M. van de Panne, S. Paris, and W. Heidrich. Displacement interpolation using lagrangian mass transport. *ACM Transaction on Graphics*, 30(6):158:1–158:12, December 2011.
18. 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.
19. P. Knight. The sinkhorn-knopp algorithm: Convergence and applications. *SIAM J. Matrix Anal. Appl.*, 30(1):261–275, March 2008.
20. 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.
21. D.R. Hunter and K. Lange. A Tutorial on MM Algorithms. *The American Statistician*, 58(1):30–38, 2004.
22. A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.
23. K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, LNCS, pages 213–226, 2010.
24. G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. Technical Report CNS-TR-2007-001, California Institute of Technology, 2007.