Machine Learning (2019) 108:1635–1652
https://doi.org/10.1007/s10994-019-05823-0

On the analysis of adaptability in multi-source domain adaptation

Ievgen Redko[1] · Amaury Habrard[1] · Marc Sebban[1]

Received: 21 January 2019 / Revised: 10 May 2019 / Accepted: 21 June 2019 / Published online: 27 June 2019
© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2019

Abstract
In many real-world applications, it may be desirable to benefit from a classifier trained on a given *source* task from some largely annotated dataset in order to address a different but related *target* task for which only weakly labeled data are available. Domain adaptation (DA) is the framework which aims at leveraging the statistical similarities between the source and target distributions to learn well. Current theoretical results show that the efficiency of DA algorithms depends on (i) their capacity of minimizing the divergence between the source and target domains and (ii) the existence of a good hypothesis that commits few errors in both domains. While most of the work in DA has focused on new divergence measures, the second aspect, often modeled as the *capability term*, remains surprisingly under-investigated. In this paper, we show that the problem of the best joint hypothesis estimation can be reformulated using a Wasserstein distance-based error function in the context of multi-source DA. Based on this idea, we provide a theoretical analysis of the capability term and derive inequalities allowing us to estimate it from finite samples. We empirically illustrate the proposed idea on different data sets.

Keywords Transfer learning · Domain adaptation · Learning theory

1 Introduction

Current advances in statistical learning theory offer a variety of results that study the problem of estimating the probability that a hypothesis h picked from a given hypothesis class $\mathcal{H}$ can achieve a small true risk. These results take often the form of generalization bounds on the true risk, and are derived using concentration inequalities w.r.t. $\mathcal{H}$. Classic generalization

Editors: Karsten Borgwardt, Po-Ling Loh, Evimaria Terzi, Antti Ukkonen.

✉ Ievgen Redko
Ievgen.Redko@univ-st-etienne.fr

Amaury Habrard
Amaury.Habrard@univ-st-etienne.fr

Marc Sebban
Marc.Sebban@univ-st-etienne.fr

[1] CNRS, Institut d Optique Graduate School Laboratoire Hubert Curien UMR 5516, Univ Lyon, UJM-Saint-Etienne, 42023 Saint-Étienne, France

bounds assume that training and test data follow the same distribution. This assumption, however, does not reflect the peculiarities of many real-world applications like in computer vision, language processing or speech recognition where training and test data actually often follow a related but different probability distribution. The need for algorithms addressing this problem has led to the emergence of a new machine learning area called domain adaptation (DA), a subfield of transfer learning (Pan and Yang 2010), where the *source* (training) and *target* (test) distributions are assumed to be related but different. Existing generalization guarantees for DA are expressed in the form of bounds over the target true risk of h involving (i) the source true risk, (ii) a divergence between the domain distributions and (iii) a term λ evaluating the capability of the considered class $\mathcal{H}$ to solve the problem, often expressed as a joint error of the ideal hypothesis between the two domains. The majority of theoretical investigations during the past few years aimed at introducing new divergence measures, like the $\mathcal{H}$ divergence (Ben-David et al. 2010a), the discrepancy distance (Mansour et al. 2009a; Cortes and Mohri 2014), integral probability metrics (Zhang et al. 2012), to cite the most widely used ones. Surprisingly, very few theoretical results studied the capability term λ which was often assumed to be negligibly small to allow adaptation or said differently, the two source/target labeling functions were supposed to be similar. Nevertheless, it was shown by Ben-David et al. (2010b) that in general minimizing only the divergence between the two domain distributions is not sufficient for efficient adaptation. Considering the following set of assumptions: (1) the source and target distributions are close to each other; (2) there exists a hypothesis $h \in \mathcal{H}$ with low error λ on both domains; (3) the labeling function does not change between the source and target domains, the authors concluded that neither the combination (1) + (3) nor (2) + (3) suffices for successful adaptation. Consequently, the existence of a good joint hypothesis plays a crucial role in DA in the same way as the divergence measure.

In this paper, we provide a first theoretical analysis of the λ term using ideas from the optimal transportation theory. We choose this particular mathematical framework because it provides a large variety of theoretical results that are particularly suited for DA. Similar to Crammer et al. (2008); Mansour et al. (2009b), we place our work in a more general and complex setting of multi-source DA where we possess $N \geq 2$ source domains. We motivate this particular choice by the fact that a multi-source scenario allows a more accurate estimation of the adaptability of a given DA problem. We redefine λ by expressing the error function of a hypothesis over each domain in terms of the Wasserstein distance. This choice offers us a powerful geometric tool that we use to compare probability distributions and that, as mentioned in Le Gouic and Loubes (2017), can represent more accurately the inner geometry of a large possibly high-dimensional data sample. Furthermore, using Wasserstein distance as a loss function offers several algorithmic advantages as its gradient is not vanishing on distributions with a different support compared to Kullback–Leibler and L_p distances: a property that has allowed to overcome the mode collapse problem in Generative Adversarial Networks (Goodfellow et al. 2014; Arjovsky et al. 2017). The theoretical contributions of this paper are the following: (1) we characterize the uniqueness and existence of the capability term; (2) we present inequalities that allow us to bound the true λ term by its finite-sample approximation.

The rest of this paper is organized as follows: Sect. 2 is devoted to the definition of the DA problem with multiple sources, to the introduction of the optimal transport theory and some related concentration inequalities. In Sect. 3, we present a new definition of the capability term w.r.t. a Wasserstein distance-based error function and use it in Sect. 4 to establish the uniqueness and the existence of the capability term for every distinct multi-source DA problem. We further prove finite-sample inequalities for its empirical counterpart in both one- and d-dimensional cases with two different strategies to compute it. In Sect. 5, we show the validity and the appropriateness of our estimation procedure by evaluating it on synthetic data. We conclude in Sect. 6 by drawing future research directions of this work.

2 Preliminary knowledge

In this section, we formally define a DA problem and present the general form of DA generalization bounds. Then, we introduce a brief overview of the concepts from optimal transport used in the next sections.

2.1 Domain adaptation

Let us define a domain D as a pair consisting of a distribution μ_D on the instance space Ω and a labeling function $f_D : \Omega \rightarrow [0, 1]$. We further define a hypothesis class $\mathcal{H}$ as a set of functions so that $\forall h \in \mathcal{H}, h : \Omega \rightarrow [0, 1]$. With this notations, the error function of a given domain can be defined as follows.

Definition 1 Given a convex loss-function $\ell : [0, 1] \times [0, 1] \rightarrow \mathbb{R}^+$, the true risk according to the distribution μ_D that a hypothesis $h \in \mathcal{H}$ disagrees with a labeling function f_D (which can also be a hypothesis) is defined as $\epsilon_\ell^D(h, f_D) = \underset{x\sim\mu_D}{\mathbb{E}} [\ell(h(x), f_D(x))]$.

When the source and target risks (or error functions) are defined w.r.t. h and f_S or f_T, we use the shorthand $\epsilon_\ell^S(h, f_S) = \epsilon_\ell^S(h)$ and $\epsilon_\ell^T(h, f_T) = \epsilon_\ell^T(h)$. The ultimate goal of DA is to learn a good hypothesis h on S (given by a labeled sample of size n_S) that has a good performance on T (given by a possibly unlabeled sample of size n_T). In what follows, we consider the generalization of the DA problem, where not 1 but N source domains are available. Furthermore, we place ourselves in a semi-supervised setting where a small portion of labeled data is available from T. This setting is likely to be one of the most complicated ones because it definitely prevents the learner from only using the target examples to learn a hypothesis which would work well on target distribution. We define N different source domains (where the target domain T can either be or not a part of this set) represented by N labeled samples S_j $(j = 1, \ldots, N)$ of size $n_j = \beta_j n$ $(\sum_{j=1}^N \beta_j = 1, \sum_{j=1}^N n_j = n)$ drawn from some unknown distribution μ_{S_j} and labeled by f_{S_j}. Now, let us consider the weighted multi-source error of a hypothesis h defined for some vector $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_N\}$ as follows:

$$\epsilon_\ell^{\boldsymbol{\alpha}}(h) = \sum_{j=1}^N \alpha_j \epsilon_\ell^{S_j}(h), \tag{1}$$

where $\sum_{j=1}^N \alpha_j = 1$ and each α_j represents a weight assigned to the source domain S_j. We further denote by $\hat{\epsilon}_\ell^{\boldsymbol{\alpha}}(h)$ its empirical counterpart defined over the empirical error functions $\hat{\epsilon}_\ell^{S_j}(h)$. Denoting by $\hat{h}_{\boldsymbol{\alpha}}$ and h_T^* the minimizers of $\hat{\epsilon}_\ell^{\boldsymbol{\alpha}}(h)$ and $\epsilon_\ell^T(h)$ respectively, Ben-David et al. (2010b) have shown that generalization bounds for multi-source DA can be expressed as follows:

$$\epsilon_\ell^T(\hat{h}_{\boldsymbol{\alpha}}) \leq \epsilon_\ell^T(h_T^*) + 2\sum_{j=1}^N \alpha_j \left(\mathrm{d}(\mu_{S_j}, \mu_T) + \lambda_j\right) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right),$$

where λ_j is the combined error of the ideal hypothesis h^* that minimizes $\epsilon_\ell^{S_j}(h) + \epsilon_\ell^T(h)$ and $\mathrm{d}(\cdot, \cdot)$ is some divergence measure on the space of probability distributions. From this result, one can instantly see that both the distance between the source and target distributions and the λ term have equal impact on the success of adaptation. Furthermore, as stated in Ben-David et al. (2010a), "*when the combined error of the ideal joint hypothesis is large, then there is no classifier that performs well on both the source and target domains, so we cannot hope to find a good target hypothesis by training only on the source domain*". This statement is the core motivation of our work.

2.2 Optimal transport and Wasserstein distance

Optimal transportation theory was first introduced in Monge (1781) to study the problem of resource allocation. Assuming that we have a set of factories and a set of mines, the goal is to move the ore from mines to factories in an optimal way, i.e., by minimizing the overall transport cost. The Wasserstein metric is a distance between two probability distributions which relies on the optimization problem of the optimal transport. Here, we focus specifically on the Wassertein distance between the source and target distributions μ_S and μ_T. Let $\mathcal{P}_p(\Omega) := \{\mu \in \mathcal{P}(\Omega) : \int_\Omega \|x\|^p d\mu(x) < \infty\}$ be the space of probability measures supported on Ω with finite pth moment. The Wasserstein distance of order p between μ_S, μ_T for any $p \geq 1$ is defined as:

$$W_p^p(\mu_S, \mu_T) = \inf_{\gamma \in \Pi(\mu_S, \mu_T)} \int_{\Omega \times \Omega} c(x, y)^p d\gamma(x, y),$$

where $c : \Omega \times \Omega \rightarrow \mathbb{R}^+$ is a metric, $\Pi(\mu_S, \mu_T)$ is the collection of joint probability measures on $\Omega \times \Omega$ with marginals μ_S and μ_T, also called the set of couplings. In practice, we deal with the empirical measures $\hat{\mu}_S = \frac{1}{n_S} \sum_{i=1}^{n_S} \delta_{x_i^S}$ and $\hat{\mu}_T = \frac{1}{n_T} \sum_{i=1}^{n_T} \delta_{x_i^T}$ defined on finite samples and represented by the uniformly weighted sums of Diracs with mass at locations x_i^S and x_i^T, respectively. In such a context, the Wassertein distance $W_p^p(\hat{\mu}_S, \hat{\mu}_T)$ corresponds to the minimum cost of turning the source probability mass in the target probability mass obtained by solving the Monge–Kantorovich problem. More formally, it can be written in terms of the inner product between the coupling matrix γ and a cost matrix C as follows:

$$W_p^p(\hat{\mu}_S, \hat{\mu}_T) = \min_{\gamma \in \Pi(\hat{\mu}_S, \hat{\mu}_T)} \langle C, \gamma \rangle_F,$$

where $\langle \cdot,\cdot \rangle_F$ is the Frobenius dot product and C is a dissimilarity matrix, i.e., $C_{ij} = c(x_i^S, x_j^T)^p$, defining the energy needed to move a probability mass from x_i^S to x_j^T. In case when $p = 1$, one obtains the popular Earth mover's distance (Rubner et al. 2000) commonly used in image retrieval. Different results have been proposed in the literature regarding the convergence in expectation of the empirical measure to the true one in terms of Wasserstein distance. As these results play a major role in our work, we present the concentration inequality for measures supported on $\mathbb{R}^d$ below.

Theorem 1 (Fournier and Guillin 2015) *Let $\mu \in \mathcal{P}(\Omega)$, $\Omega \subset \mathbb{R}^d$ and let $p \in (0, d/2)$. Assume that $M_q(\mu) = \int_{\mathbb{R}^d} |x|^q \mu(dx) < \infty$ for some $q > p$, $q \neq \frac{d}{d-p}$. Then there exists a constant ς depending on p, d and q such that for all $n \geq 1$ the following bound holds*

$$\mathbb{E}\left[W_p(\mu, \hat{\mu})\right] \leq \varsigma M_q^{p/q}(\mu) \left(n^{-\frac{p}{d}} + n^{-\frac{q-p}{q}}\right).$$

This theorem shows that $W_p(\mu, \hat{\mu}) \rightarrow 0$ with probability one and the rate of convergence depends on a variety of hypotheses and properties of the distribution μ that are discussed in the following sections.

3 Ideal joint hypothesis with the Wasserstein distance-based error function

The vast majority of DA algorithms focuses on minimizing the divergence between the source and target sets either by learning a shared representation space or by reweighting the source data (Margolis 2011). However, the impossibility theorems for DA presented in Ben-David et al. (2010b) suggest that for a successful adaptation, only minimizing the divergence is not enough. The key term that is not taken into account but that, nevertheless, remains of the huge importance is λ: the error of the best joint hypothesis over the source and target domains. The importance of λ is highlighted by its appearance in many theoretical results on DA. First introduced by Ben-David et al. (2007), it has then been taken into account in Crammer et al. (2008) under the form of disparity coefficients measuring the disagreement between labels of the source domains and in Mansour et al. (2009a) as the error defined over the target distribution between the ideal source and target hypotheses.[1] The intuition behind λ is the following: while the divergence term in the bounds encourages one to reduce the discrepancy between the available source and target samples in order to align them, λ motivates to adapt in a way that ensures the separability of classes of the aligned samples. This intuition highlights the importance of λ and prompts us to provide a complete theoretical framework that allows an estimation of λ from observable data. In the next section, we show that this can be achieved if we express the error functions of source and target domains in terms of the Wasserstein distance.[2]

To proceed, we first note that the idea of using the Wasserstein distance as a loss function has been proposed in Frogner et al. (2015) and applied successfully for multi-label and multi-class classification. For the sake of completeness, we give its definition below.

Definition 2 Let $\mathcal{K}$ denote the space of all possible outputs and $\mathcal{H}$ be the hypothesis space. $\forall h \in \mathcal{H}, h : \Omega \to \Delta^{|\mathcal{K}|}$, let $h(\kappa|x) = h(x)_\kappa$ be the predicted value at element $\kappa \in \mathcal{K}$, given the input $x \in \Omega$. Let $f(\kappa)$ be the ground truth value for κ given by the corresponding label y. Then, the Wasserstein loss is defined as

$$W_p^p(h(\cdot|x), f(\cdot)) = \min_{\gamma \in \Pi(h(x), f)} \langle C, \gamma \rangle_F.$$

For this particular loss function, the error with respect to a given domain is defined as:

$$\epsilon_{W_\mathcal{K}}^D(h, f_D) = \mathop{\mathbb{E}}_{x \sim \mu_D} \left[W_p^p(h(\cdot|x), f_D \right],$$

where for each x, $h(\cdot|x)$ and f_D yield a distribution over the output space $\mathcal{K}$ providing the information regarding the multiple possible labels of x. One may note that in case of multi-class classification, $f_D(x)$ is given by a one-hot vector, while in multi-label classification it can take non-trivial values that do not necessarily sum to 1. In both cases, the overall true error amounts to taking the expectation with respect to the marginal data distribution over the discrepancies measured by the Wasserstein distance between the desired output and the obtained one.

[1] In a less general case where the hypothesis space is restricted to the weighted combination of source hypotheses, one can obtain λ-free bounds as shown in Mansour et al. (2009b).

[2] Note that generalization bounds involving the Wasserstein distance between the source and target probability distributions have been proposed in Redko et al. (2017). This work, however, is different from the results of the mentioned paper as it does not aim at introducing new generalization bounds for domain adaptation, nor it considers the divergence between the source and target probability distributions in the provided analysis.

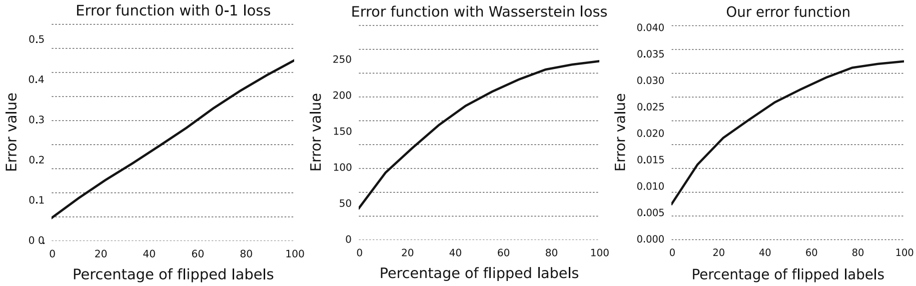

Fig. 1 Comparison of empirical error functions defined with (left) traditional 0–1 loss, (middle) original Wasserstein loss and (right) global error function from this paper. The errors were averaged based on class probabilities obtained by fitting a linear classifier on samples consisting of 500–10,500 instances with step equal to 1000 with 20 features and 2 classes. The classification problem is made gradually harder by randomly flipping a certain proportion of instances' labels (x-axis) between the two classes. For the original Wasserstein loss, we calculate point-wise Wassertein distances between the vector of probabilities and the one-hot encoding vector with the true label. For our error function, we calculate the Wasserstein distance between the normalized vectors of probabilities over all instances and that of true labels. We observe that despite the different scales, all losses behave similarly and reflect the increasing difficulty of the classification problem

In order to introduce the definition of the error function that we use in our work, we first note that in case of binary classification with $|\mathcal{K}| = 2$, Definition 2 boils down to comparing the two-dimensional vector $[h(0|x), h(1|x)]$ of class probabilities produced by h to one-hot vector given by f_D. In this particular case, the information carried by elements of $h(\cdot|x)$ becomes redundant as $h(0|x) = 1 - h(1|x)$ and knowing only $h(0|x)$ or $h(1|x)$ is enough to calculate the Wasserstein loss at x. Driven by this observation, we propose to define the error function of a given domain D in terms of the Wasserstein distance between the hypothesis h and true labeling function f_D considered as probability measures supported on Ω. In this case, the definition reads:

$$\epsilon_W^D(h, f_D) := W_p^p(h(x), f_D(x)), \; h, f_D \in \mathcal{P}(\Omega). \tag{2}$$

We note that our definition of the error function is quite different from the one used by Frogner et al. (2015) as it compares the outputs of the hypothesis and labeling functions directly over the whole input space. This means that $\epsilon_W^D(h, f_D)$ can be seen as a global measure of disagreement as it does not rely on local point-wise averaging over the instances of the input space. In order to evaluate the adequacy of the proposed error function, we illustrate the behaviour of the traditional empirical error with 0-1 loss, the empirical error calculated with the original Wasserstein loss and our proposed empirical error on a set of classification problems with increasing difficulty in Fig. 1. From this figure, we note that all error functions behave similarly thus justifying our proposed definition in case of binary classification.

Throughout the rest of the paper, we assume that $h(x)$, $f_{S_j}(x)$ and $f_T(x)$ are normalized so that $\int_\Omega h(x)dx = 1$, $\int_\Omega f_T(x)dx = 1$ and $\int_\Omega f_{S_j}(x)dx = 1, \forall j = \{1, \ldots, N\}$. In case of binary classification, the values of a hypothesis and labeling functions are given by the probability of a given instance for belonging to one of the classes of interest as explained above. One may further note that in the binary setting, normalising labelling function does not change the outputted predictions. To see that, we can consider a labeling function $f = (0.2, 0.8, 0.8, 0.2)$ of 4 instances, where 1st and 4th instances belong to class 0 ($f \geq 0.5$) while 2nd and 3rd are in class 1 ($f < 0.5$). To obtain the same predictions for the normalized $f = (0.1, 0.4, 0.4, 0.1)$, one simply have to threshold it at $\frac{1}{2\sum f} = \frac{1}{4}$.

We further set $p = 2$ in (2), where the squared 2-Wasserstein distance is chosen due to the fact that it is strictly convex (Álvarez Esteban et al. 2011). Note that the Wasserstein distance belongs to a vast family of Integral-probability metrics (IPMs) (Zolotarev 1984) along with many other noticeable examples including the famous Maximum mean discrepancy distance (Smola et al. 2007) closely related to the regularized optimal transport (Genevay et al. 2018). Nevertheless, the particular choice of the Wasserstein distance made in this paper is due to the existence of a well-defined and extensively studied notion of the Wasserstein barycenter that we make use of to establish our results.

We now turn our attention to the overall weighted joint error $\lambda_{\boldsymbol{\alpha}}$ defined as the weighted sum over all λ_j as follows:

$$\lambda_{\boldsymbol{\alpha}} := \sum_{j=1}^{N} \alpha_j \lambda_j = \min_{h \in \mathcal{H}} \sum_{j=1}^{N} \alpha_j \left(\epsilon_W^S(h, f_{S_j}) + \epsilon_W^T(h, f_T) \right).$$

Let us rewrite $\lambda_{\boldsymbol{\alpha}}$ as $\lambda_{\boldsymbol{\alpha}} = J(h^*)$, where

$$\begin{aligned} J(h^*) = \min_{h \in \mathcal{H}} J(h) &= \min_{h \in \mathcal{H}} \sum_{j=1}^{N} \alpha_j \left(\epsilon_W^S(h, f_{S_j}) + \epsilon_W^T(h, f_T) \right) \\ &= \min_{h \in \mathcal{H}} \mathbb{E} \left[\sum_{j=1}^{N} \alpha_j W_2^2(h, f_{S_j}) + W_2^2(h, f_T) \right]. \end{aligned}$$

Similarly, we define the empirical value of $\lambda_{\boldsymbol{\alpha}}$ by $\hat{\lambda}_{\boldsymbol{\alpha}} := \hat{J}(h^*)$, where

$$\hat{J}(h^*) = \min_{h \in \mathcal{H}} \hat{J}(h) = \min_{h \in \mathcal{H}} \sum_{j=1}^{N} \alpha_j (\hat{\epsilon}_W^S(h, f_{S_j}) + \hat{\epsilon}_W^T(h, f_T)).$$

The goal of the next section is to study how the true $\lambda_{\boldsymbol{\alpha}}$ can be related to its empirical counterpart $\hat{\lambda}_{\boldsymbol{\alpha}}$.

4 Analysis of $\boldsymbol{\lambda_\alpha}$

The core construction proposed in this section is the expression of $\lambda_{\boldsymbol{\alpha}}$ based on the Wasserstein barycenter optimization problem defined for a parametric family of random probability measures. To proceed, we start by defining the key quantities used to achieve this goal.

4.1 Uniqueness and existence

Let us consider a parametric set of probability measures $\{f_\theta, \theta \in \Theta \subset \mathbb{R}^s, s \geq 1\}$, where for every parameter vector θ, we assume that f_θ admits a density with respect to the Lebesgue measure on Ω. Now, if $\boldsymbol{\theta} \in \Theta$ is a random vector with distribution $\mathbb{P}_\theta$ admitting a density function $g : \Theta \to \mathbb{R}_+$, then $f_{\boldsymbol{\theta}}$ is a random probability measure with distribution $\mathbb{P}_g$. Let us define the true source and target labeling functions as random probability measures parameterized by random vectors $\boldsymbol{\theta}_j, \boldsymbol{\theta}_T \in \Theta$ so that $f_{S_j} = f_{\boldsymbol{\theta}_j}, \forall j \in \{1, \ldots, N\}$ and $f_T = f_{\boldsymbol{\theta}_{N+1}}$. For a vector $\alpha = \{\alpha_1, \ldots, \alpha_{N+1}\}$ with $\alpha_i \geq 0, \forall i$, we can now define the measures $\hat{\mathbb{P}} = \sum_{i=1}^{N+1} \alpha_i \delta_{f_{\theta_i}}$ and $\hat{\mathbb{P}}_N = \sum_{i=1}^{N+1} \alpha_i \delta_{\hat{f}_{\theta_i}}$, where $\hat{f}_{\boldsymbol{\theta}_i} = \sum_{j=1}^{n_i} f_{S_i}(x_j^{S_i}) \delta_{x_j^{S_i}}$ and $n_i =$

$\beta_i n$, $\forall i \in \{1, \dots, N\}$ is the number of samples available in the i^{th} source domain and $\{x_j^{S_i}\}_{j=1}^{n_i}$ is its associated random sample drawn from μ_{S_i}. Likewise, $\hat{f}_{\boldsymbol{\theta}_{N+1}} = \sum_{j=1}^{n_T} f_T(x_j^T)\delta_{x_j^T}$, where the sample $\{x_j^T\}_{j=1}^{n_T}$ is drawn from μ_T with $n_{N+1} = n_T$. To summarize, $f_{\boldsymbol{\theta}}$ is a random probability measure that defines the underlying distribution of the source labeling functions; $\{f_{\boldsymbol{\theta}_j}\}_{j=1}^{N+1}$ are independent copies of $f_{\boldsymbol{\theta}}$ that can be seen as realizations of the underlying distribution $f_{\boldsymbol{\theta}}$ and finally $\hat{f}_{\boldsymbol{\theta}_j}$ are empirical source labeling functions defined w.r.t. the available finite samples in all source domains. We can now rewrite $\lambda_{\boldsymbol{\alpha}}$ and $\hat{\lambda}_{\boldsymbol{\alpha}}$ as follows:

$$\lambda_{\boldsymbol{\alpha}} = \min_{h \in \mathcal{P}_2(\Omega)} \sum_{i=1}^{N+1} \alpha_i W_2^2(h, f_{\boldsymbol{\theta}_i}) = \min_{h \in \mathcal{P}_2(\Omega)} \mathbb{E}_{\hat{\mathbb{P}}} \left[W_2^2(h, f_{\boldsymbol{\theta}}) \right],$$

$$\hat{\lambda}_{\boldsymbol{\alpha}} = \min_{h \in \mathcal{P}_2(\Omega)} \sum_{i=1}^{N+1} \alpha_i W_2^2(h, \hat{f}_{\boldsymbol{\theta}_i}) = \min_{h \in \mathcal{P}_2(\Omega)} \mathbb{E}_{\hat{\mathbb{P}}_N} \left[W_2^2(h, f_{\boldsymbol{\theta}}) \right].$$

We further introduce the combined error defined over the unknown distribution $\mathbb{P}_g$ such that

$$\begin{aligned} \lambda_{\mathbb{P}_g} &= \min_{h \in \mathcal{P}_2(\Omega)} J_{\mathbb{P}_g}(h) = \min_{h \in \mathcal{P}_2(\Omega)} \mathbb{E}_{\mathbb{P}_g} \left[W_2^2(h, f_{\boldsymbol{\theta}}) \right] \\ &= \min_{h \in \mathcal{P}_2(\Omega)} \int_{\Theta} W_2^2(h, f_\theta) g(\theta) d\theta. \end{aligned}$$

This functional is defined over the distribution that generated a sequence of random densities $f_{\boldsymbol{\theta}_j}, \forall j$. The proposed reformulation allows us now to relate the definition of the ideal joint hypothesis to the Wasserstein barycenter optimization problem (Agueh and Carlier 2011) defined as follows.

Definition 3 For a set of N probability measures $\{\mu_1, \mu_2, \dots, \mu_N\} \in \mathcal{P}_p(\Omega)$, the empirical Wasserstein barycenter $\hat{\nu}^*$ is defined as

$$\hat{\nu}^* = \operatorname*{arg\,min}_{\nu \in \mathcal{P}_p(\Omega)} \sum_{i=1}^{N} \alpha_i W_p^p(\nu, \mu_i),$$

where for all i, $\alpha_i \geq 0$, $\sum_{i=1}^N \alpha_i = 1$. In a more general case, the population Wasserstein barycenter ν^* of a random measure $\boldsymbol{\mu}$ with distribution $\mathbb{P}$ can be defined as

$$\nu^* = \operatorname*{arg\,min}_{\nu \in \mathcal{P}_p(\Omega)} \mathbb{E} \left[W_p^p(\nu, \boldsymbol{\mu}) \right] = \int_{\mathcal{P}_p(\Omega)} W_p^p(\nu, \mu) d\mathbb{P}(\mu).$$

Given this definition, we note that the minimizers $\hat{h}^*$, h^* and $h^*_{\mathbb{P}_g}$ of $\hat{J}(h)$, $J(h)$ and $J_{\mathbb{P}_g}(h)$, respectively are all Wasserstein barycenters defined w.r.t. different sets of random measures. Now, the following result for $\hat{J}(h)$ $J(h)$, and $J_{\mathbb{P}_g}(h)$ can be obtained.

Theorem 2 *Let (Ω, c) be a separable locally compact geodesic space. Assume that for any $\boldsymbol{\theta} \in \Theta$, $f_{\boldsymbol{\theta}} \in \mathcal{P}_2(\Omega)$ has the probability distribution $\mathbb{P}_g$ admitting a density function $g : \Theta \to \mathbb{R}_+$. Let us denote by $\hat{h}^*$, h^* and $h^*_{\mathbb{P}_g}$ the minimizers of $\hat{J}(h)$, $J(h)$ and $J_{\mathbb{P}_g}(h)$, respectively. Then, for any $\boldsymbol{\alpha} \in \Delta^{N+1}$ the following statements hold:*

1. *the minimizers $\hat{h}^*$, h^* and $h^*_{\mathbb{P}_g}$ of $\hat{J}(h)$, $J(h)$ and $J_{\mathbb{P}_g}(h)$ always exist and they are unique;*
2. $\lim_{N \to \infty} W_2^2(h^*, h^*_{\mathbb{P}_g}) = 0$, $\lim_{\forall i, n_i \to \infty} W_2^2(\hat{h}^*, h^*) = 0$.

Proof From the theorem statement, we may write $\hat{h}^*$, h^* and $h^*_{\mathbb{P}_g}$ as

$$\hat{h}^* = \arg\min_h \mathbb{E}_{\hat{\mathbb{P}}_N}\left[W_2^2(h, f_{\boldsymbol{\theta}})\right],$$
$$h^* = \arg\min_h \mathbb{E}_{\hat{\mathbb{P}}}\left[W_2^2(h, f_{\boldsymbol{\theta}})\right],$$
$$h^*_{\mathbb{P}_g} = \arg\min_h \mathbb{E}_{\mathbb{P}_g}\left[W_2^2(h, f_{\boldsymbol{\theta}})\right].$$

From Definition 3, it immediately follows that $\hat{h}^*$, h^* and $h^*_{\mathbb{P}_g}$ are Wasserstein barycenters of random probability measure $f_{\boldsymbol{\theta}}$ with respect to the distributions $\hat{\mathbb{P}}_N$, $\mathbb{P}_N$ and $\mathbb{P}_g$, respectively. The existence of Wasserstein barycenters with respect to any probability distribution for locally compact geodesic space was proved in Le Gouic and Loubes (2017, Theorem 2) and thus it ensures the existence of $\hat{h}^*$, h^* and $h^*_{\mathbb{P}_g}$. The uniqueness follows from the fact that $f_{\boldsymbol{\theta}}$ is assumed to be absolutely continuous with respect to the Lebesgue measure and thus satisfies the uniqueness condition as proved in Boissard et al. (2015, Theorem 3.1). Finally, the consistency of the Wasserstein barycenters follows from Le Gouic and Loubes (2017, Theorem 3). □

This theorem has two important consequences. First, it states that for any probability measures $\mu_{S_j}, \mu_T \in \mathcal{P}_2(\Omega)$ and any labeling functions $f_{S_j}, f_T : \Omega \to [0, 1]$ that are absolutely continuous with respect to the Lebesque measure on Ω, the true and empirical ideal joint errors $\lambda_{\boldsymbol{\alpha}}$ and $\hat{\lambda}_{\boldsymbol{\alpha}}$ calculated based on the available source and target samples are unique. Secondly, it establishes the convergence of $\hat{\lambda}_{\boldsymbol{\alpha}}$ to $\lambda_{\boldsymbol{\alpha}}$ and $\lambda_{\boldsymbol{\alpha}}$ to $\lambda_{\mathbb{P}_g}$ with the increasing number of available sources and the increasing size of available source samples, respectively. The first consequence shows that each adaptation problem given by a set of source domains and a target one can be uniquely characterized by the adaptability term $\lambda_{\boldsymbol{\alpha}}$ which determines the a priori success of DA. Furthermore, the second consequence implies that with the increasing number of source domains, the estimation of the $\lambda_{\boldsymbol{\alpha}}$ term becomes more reliable meaning that $\hat{\lambda}_{\boldsymbol{\alpha}}$ can be explicitly learned when the number of available labeling functions grows to infinity.

4.2 Finite-sample inequalities

Even though Theorem 2 gives a first analysis of $\lambda_{\boldsymbol{\alpha}}$, it does not provide a way to estimate it based on the available finite samples. In order to bridge this gap, our next result establishes an inequality that bounds $\lambda_{\boldsymbol{\alpha}}$ by $\hat{\lambda}_{\boldsymbol{\alpha}}$ and a term depicting the Wasserstein distance between the true and empirical ideal joint hypotheses.

Theorem 3 *With the assumption of Theorem 2, let $\delta^2(\Omega) := \sup_{x\in\Omega}\left\{|x|^2\right\}$, and let h^*, $\hat{h}^*$ denote the minimizers of $J(h)$ and $\hat{J}(h)$, respectively. Then, the following inequality holds*

$$\lambda_{\boldsymbol{\alpha}} \le \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\mathbb{E}\left[W_2(h^*, \hat{h}^*)\right].$$

Proof

$$\begin{aligned}\lambda_{\boldsymbol{\alpha}} &= \lambda_{\boldsymbol{\alpha}} + \hat{\lambda}_{\boldsymbol{\alpha}} - \hat{\lambda}_{\boldsymbol{\alpha}}\\ &= \hat{\lambda}_{\boldsymbol{\alpha}} + \mathbb{E}\left[W_2^2(h^*, f_\theta)\right] - \mathbb{E}\left[W_2^2(\hat{h}^*, f_\theta)\right]\end{aligned}$$

$$
\begin{aligned}
&\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \left| \mathbb{E}\left[W_2^2(h^*, f_\theta)\right] - \mathbb{E}\left[W_2^2(\hat{h}^*, f_\theta)\right] \right| \\
&\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \mathbb{E}\left[\left| W_2^2(h^*, f_\theta) - W_2^2(\hat{h}^*, f_\theta)\right|\right] \qquad (3)\\
&\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\mathbb{E}\left[W_2(h^*, \hat{h}^*)\right]. \qquad (4)
\end{aligned}
$$

Here (3) is obtained using Jensen inequality for expected value taken over the convex absolute value function. (4) is due to the reverse triangle inequality and the uniform boundedness of the class of functions $\mathcal{F} = \{W_2^2(\mu, \nu) | \mu \in \mathcal{P}(\Omega)\}$ for some $\nu \in \mathcal{P}(\Omega)$ in the supremum norm. □

This result shows that the convergence of $\hat{\lambda}_{\boldsymbol{\alpha}}$ to $\lambda_{\boldsymbol{\alpha}}$ is controlled by the convergence of $\hat{h}^*$ to h^* w.r.t. the Wasserstein distance. As mentioned in Sect. 2, this convergence can be characterized in a variety of ways depending on the support of $\{f_{\boldsymbol{\theta}_i}\}_{i=1}^{N+1}$ and on the algorithm used to calculate the barycenter.

Measures supported on $\mathbb{R}$ For our first result, we assume that the source and target labeling functions are supported on the interval $\Omega \subset \mathbb{R}$. In this one-dimensional case, computing the Wasserstein barycenter simply amounts to averaging (in the usual way) their quantile functions. This setting, known as quantile synchronization (Zhang and Müller 2011), leads to the following theorem.

Theorem 4 *With the assumptions of Theorem 3, let us suppose that $h^*_{\mathbb{P}_g}$ and $f_{\boldsymbol{\theta}_i} \in \mathcal{P}_2(\Omega \subset \mathbb{R})$, $\boldsymbol{\theta}_i \in \mathbb{R}^s, s \geq 1$ for all $i = \{1, \ldots, N+1\}$ are absolutely continuous w.r.t. the Lebesgue measure dx on $\mathbb{R}$. Denote by F^- a quantile function of $f_{\boldsymbol{\theta}}$ and let $J_2(\mu) = \int_0^1 \left[(F^{-1})'(x)\sqrt{x(1-x)}\right]^2 d(x)$ for some probability measure $\mu \in \mathcal{P}(\Omega)$ with cumulative function F such that F^{-1} is absolutely continuous. Then for any $n_i \geq 1$, $i = \{1, \ldots, N+1\}$ and $N \geq 1$ the following holds*

$$
\begin{aligned}
\lambda_{\boldsymbol{\alpha}} \leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\Bigg(&\frac{10 J_2^{\frac{1}{2}}(h^*_{\mathbb{P}_g})}{\sqrt{N+1}} + \frac{\sqrt{2\mathbb{E}J_2(f_{\boldsymbol{\theta}})}}{N+1}\sum_{i=1}^{N+1}(n_i)^{-\frac{1}{2}} \\
&+ \sqrt{\frac{1}{N+1}\int_0^1 \mathrm{Var}(F^-(\tau))d\tau}\Bigg).
\end{aligned}
$$

Proof

$$
\begin{aligned}
\lambda_{\boldsymbol{\alpha}} &\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\mathbb{E}\left[W_2(h^*, \hat{h}^*)\right] \qquad (5)\\
&\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\left(\mathbb{E}\left[W_2(h^*, h^*_{\mathbb{P}_g}) + W_2(h^*_{\mathbb{P}_g}, \hat{h}^*)\right]\right) \qquad (6)\\
&\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\frac{10 J_2^{\frac{1}{2}}(h^*_{\mathbb{P}_g})}{\sqrt{N+1}} + \mathbb{E}\left[W_2(h^*_{\mathbb{P}_g}, \hat{h}^*)\right] \qquad (7)\\
&\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\Bigg(\frac{10 J_2^{\frac{1}{2}}(h^*_{\mathbb{P}_g})}{\sqrt{N+1}} + \frac{\sqrt{2\mathbb{E}J_2(f_{\boldsymbol{\theta}})}}{N+1}\sum_{i=1}^{N+1}(n_i)^{-\frac{1}{2}}
\end{aligned}
$$

$$+\sqrt{\frac{1}{N+1}\int_0^1 \mathrm{Var}(F^-(\tau))d\tau}\Bigg). \tag{8}$$

Equation (5) follows from Theorem 3; (6) is the application of the triangle inequality with $W_2(h^*, h^*_{\mathbb{P}_g})$ standing for the Wasserstein distance between the true and the population barycenters, while $W_2(h^*_{\mathbb{P}_g}, \hat{h}^*)$ is the distance between the true and the empirical barycenters considered above. (7) is the application of the concentration inequality given in the theorem above. Finally, (8) is due to Bigot et al. (2018, Theorem 3.2). □

The obtained inequality implies the convergence of $\hat{\lambda}$ term to its true value with the increasing number of source domains. However, when the number of source domains is fixed and only the size of samples available in each source domain tends to infinity, the inequality suggests the existence of a bias introduced by $J_2^{\frac{1}{2}}(h^*_{\mathbb{P}_g})$ and $\int_0^1 \mathrm{Var}(F^-(\tau))d\tau$, where $\int_0^1 Var(F^-(\tau))d\tau$ is always finite for any square-integrable measure $h^*_{\mathbb{P}_g}$. In cases where $J_2(h^*_{\mathbb{P}_g}) = \infty$ or $J_2(f_{\boldsymbol{\theta}}) = \infty$, the convergence requires both the number of source domains and the size of source domain samples to tend towards infinity. Surprisingly, this would be the case for Gaussian distributions that are commonly used as a toy example in many experimental evaluations. Finally, we also note the presence of the variance of the quantile function in the bound. Given the probabilistic interpretation of the labeling functions introduced above, this term reflects the variability that exists between instances of different classes across the source domains and the target one. This is meaningful in the context of the ideal joint hypothesis that is common to source and target domains but should also perform well on each of them.

We also note that when $n_i = p, \forall i$, the barycenter can be calculated as $\hat{h}^* = \frac{1}{p}\sum_{j=1}^{p}\delta_{\bar{X}^*_j}$, where $\bar{X}^*_j = \frac{1}{N+1}\sum_{i=1}^{N+1} X^*_{i,j}$ and $X^*_{i,j}$ are order statistics of the i^{th} sample of observations $\left\{x_j^{S_i}\right\}_{j=1\ldots p}$ for source domains and $\left\{x_j^{T}\right\}_{j=1\ldots p}$ for the target one. By definition, the order statistic of a set of random variables is obtained by sorting them in the increasing order, i.e., $X^*_{i,1} = \min\{x_1^{S_i}, x_1^T\}$ and $X^*_{i,p} = \max\{x_p^{S_i}, x_p^T\}$. In this case, the inequality simplifies to the following result (Bigot et al. 2018, Theorem 3.1):

$$\lambda_{\boldsymbol{\alpha}} \leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\left(\frac{10J_2^{\frac{1}{2}}(h^*_{\mathbb{P}_g})}{\sqrt{N+1}} + \sqrt{\frac{1}{N+1}\int_0^1 Var(F^-(\tau))d\tau + \frac{2}{p+1}J_2(h^*_{\mathbb{P}_g})}\right).$$

Measures supported on $\mathbb{R}^d$ In order to prove our next result, we assume that the source and target labeling functions are supported on a subspace of $\mathbb{R}^d$. Furthermore, we consider a barycenter construction, where h^*_E is the minimizer of the Wasserstein barycenter problem with entropy regularization of h_E given as follows:

$$h^*_E = \operatorname*{arg\,min}_{h\in\mathcal{P}_2(\Omega)} \frac{1}{N+1}\sum_{i=1}^{N+1} W_2^2(h, f_{\boldsymbol{\theta}_i}) + \gamma E(h),$$

where $E(h) = \int_{\mathbb{R}^d} g_h(x)\log(g_h(x))dx$ assuming that h admits a density g_h on Ω. This particular choice is made for practical reasons as entropic regularization was proved to provide smooth barycenters especially when the input probability measures are irregular (Bigot et al.

2018b). Bearing in mind that we aim to calculate the $\hat{\lambda}_{\boldsymbol{\alpha}}$ term from finite samples, this choice appears natural and justified. The following theorem gives a finite-sample inequality for $\lambda_{\boldsymbol{\alpha}}$ when the barycenter is calculated with entropic regularization.

Theorem 5 *With the assumptions of Theorem 3, let $M_q(\mu) = \int_{\mathbb{R}^d} |x|^q \mu(dx)$ for $q > 0$. Assume that $h^*, \hat{h}^*$ are strictly log-concave probability measures supported on $\Omega \subset \mathbb{R}^d$. Then for a constant ς depending on d and $N_f = n_i^{-\frac{2}{d}} + n_i^{-\frac{q-2}{q}}$ the following holds:*

$$\lambda_{\boldsymbol{\alpha}} \leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\sqrt{\frac{2\varsigma}{\gamma(N+1)}\sum_{i=1}^{N+1} M_q^{2/q}(f_{\boldsymbol{\theta}_i})N_f}.$$

Proof

$$\begin{aligned}
\lambda_{\boldsymbol{\alpha}} &\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\mathbb{E}\left[W_2(h^*, \hat{h}^*)\right] \\
&\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\mathbb{E}\left[\sqrt{\mathrm{KL}(h^*, \hat{h}^*)}\right] && (9)\\
&\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\mathbb{E}\left[\sqrt{\mathrm{KL}(h^*, \hat{h}^*) + \mathrm{KL}(\hat{h}^*, h^*)}\right] && (10)\\
&= \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\mathbb{E}\left[\sqrt{d_E(h^*, \hat{h}^*)}\right] && (11)\\
&\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\mathbb{E}\left[\sqrt{\frac{2}{\gamma(N+1)}\sum_{i=1}^{N+1} W_2(\hat{f}_{\boldsymbol{\theta}_i}, f_{\boldsymbol{\theta}_i})}\right] && (12)\\
&\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\sqrt{\frac{2}{\gamma(N+1)}\sum_{i=1}^{N+1} \mathbb{E}\left[W_2(\hat{f}_{\boldsymbol{\theta}_i}, f_{\boldsymbol{\theta}_i})\right]} && (13)\\
&\leq \hat{\lambda}_{\boldsymbol{\alpha}} + \sqrt{2}\delta(\Omega)\sqrt{\frac{2\varsigma}{\gamma(N+1)}\sum_{i=1}^{N+1} M_q^{2/q} f_{\boldsymbol{\theta}_i} N_f}. && (14)
\end{aligned}$$

Here (9) is due to the Talagrand inequality for strictly log-concave measures; (10) comes from the fact that Kullback–Leibler divergence is always nonnegative; (11) introduces d_E which is the symmetrized Kullback–Leibler divergence; (12) is obtained from Bigot et al. (2018b, Theorem 3.4); (13) is due to the Jensen inequality for the concave square root function. Finally, (14) is a consequence of Theorem 1. □

This theorem shows the convergence of the $\hat{\lambda}_{\boldsymbol{\alpha}}$ to $\lambda_{\boldsymbol{\alpha}}$ when the number of source domains or the number of samples available in the source domains goes to infinity under the condition that M_q exists and thus is finite. Note that the assumption of strict log-concavity imposed on $h^*, \hat{h}^*$ follows from the use of the Talagrand inequality that was proved to hold in the case $p = 2$ only for Gaussian and strictly log-concave distributions. In the case of measures supported on a subspace of $\mathbb{R}^d$, this assumption means that h^* and $\hat{h}^*$ can be written as $a(x)cb(cx)$, where c is some positive constant, $a(x)$ is a log-concave measure and $b(cx)$ is the normal density $\mathcal{N}(0, c\boldsymbol{I}_d)$. Regarding $a(x)$, note that popular densities are log-concave, e.g. Gaussian and uniform densities on the compact and convex subsets of $\mathbb{R}^d$. Note also that Theorem 5 is proved for uniform weights $\boldsymbol{\alpha}$. This choice is dictated by the sake of simplicity in order to keep the proofs and theorem statements as simple as possible. Nevertheless, the same inequalities can be proved with non uniform weights using the analysis of unbalanced optimal transport (Chizat et al. 2015).

Finally, another popular way of obtaining smooth barycenters is to add a convex regularization on optimal transport plans as it was done in Cuturi and Doucet (2014). This kind of regularization relies on the Sinkhorn divergence introduced in Cuturi (2013) and leads to an optimization problem that can be solved efficiently using the Sinkhorn–Knopp algorithm (Sinkhorn and Knopp 1967). The analysis proposed in our paper can be also established for this particular penalization scheme using the concentration equalities for Sinkhorn barycenters provided in Bigot et al. (2018a).

5 Empirical results

In this section, we evaluate the usefulness of λ estimation in accesing the intrinsic difficulty of a domain adaptation problem. To this end, we first consider an illustrative example with 2 domains that shows how our estimation of a priori adaptability takes into account not only the geometric information but also the labeling of the considered source and target domains points. We further extend our empirical evaluation to a data set with increasing adaptation difficulty with 5 source domains.

5.1 Simulated data with two source domains

We present here experiments that aim at illustrating the appropriateness of our estimation procedure in quantifying the a priori difficulty of a DA problem. We consider a binary classification task with two source and one target domains, where source and target samples are composed of 600 and only 40 labeled points, respectively. Note that this imbalance between the sizes of the source and target domains is an approximation of a real-world adaptation problem where only a handful of target domain's labeled instances is assumed to be available. We consider two different scenarios that vary in terms of the intrinsic difficulty of the underlying DA problem. For the first case, we generate the classes of the source and target data according to Gaussian distributions having the same means and a slightly varying variance leading to highly similar labeling functions across the source and target domains. Figure 2(right) shows the generated data, where the class label $+1$ (resp. -1) is represented in light color (resp. dark color). In the second case, we flip the labels of the target examples to obtain a much more difficult adaptation scenario with completely different source and target labeling functions [Fig. 2(left)]. In this case, we expect a complete failure of the classifier learned on the source sample when applied to the target one. In order to illustrate our approach, we calculate the best joint hypothesis for each problem as a solution of the entropy regularized Wasserstein barycenter problem between the source and target labeling functions with equal weights. In this case, $\hat{\lambda}$ (indicated on top of each plot in Fig. 2) corresponds to the final value of the loss function of this optimization procedure.

We can make the following comments regarding the empirical estimation of $\hat{\lambda}$. First, we get a smaller value (0.008 versus 0.056) on the easier adaptation problem. This confirms that the value of λ indeed helps to access the a priori adaptation difficulty. Second, the distances between the marginal distributions of the source and target data stay almost unchanged for both setups, while we could have naively expected an increase of the discrepancy in the second case. This is an evidence that the divergence is not sufficient to properly reflect the

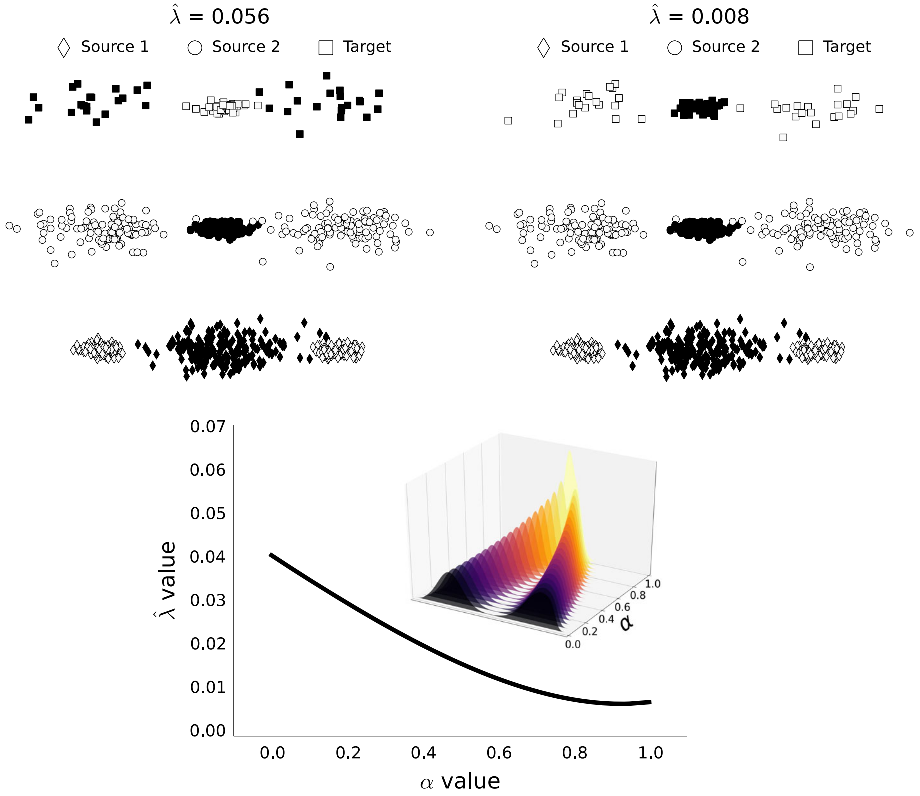

Fig. 2 Two different DA scenarios: (left) the target function is very dissimilar compared to the source ones; (right) the target and source labeling functions are very similar; (middle) interpolation between the two previous extreme situations and its impact on $\hat{\lambda}$. Surprisingly, in both cases, the Wasserstein distances between the source and target marginal distributions do not change: 0.01 for the first source and 0.2 for the second one

hardness of adaptation. This observation agrees well with the arguments that motivated us to analyze the ideal joint error in DA: even when the discrepancy between the source and target marginal distributions is small, the existence of a good hypothesis for both domains is a crucial component for the success of DA. On the other hand, we studied the behavior of the $\hat{\lambda}$ term when the target labeling function [Fig. 2(right)] gradually changes towards that of Fig. 2(left). To this end, we performed interpolation between the two labeling functions by varying their weights $[\alpha; 1-\alpha]$ for all $\alpha \in [0, 1]$ with step 0.05 in order to move from the oppositely labeled domains to similarly labeled ones. This interpolation is portrayed by the 3d plot inside Fig. 2(middle) where the target labeling function changes from a bi-modal distribution in the first case to a uni-modal distribution in the second one while the source labeling functions remain uni-modal and fixed. In general, we note that one may tune the values of α following the asymptotic analysis of its optimal mixing values proposed in Blitzer et al. (2008) and included later in Ben-David et al. (2010a). In our study, however, we choose to cover the whole interval of possible α values to fully observe the behaviour of the $\hat{\lambda}$ term. For each labeling function obtained in this way, we calculated the empirical $\hat{\lambda}$ term as before. The results are given in Fig. 2(middle). We note that $\hat{\lambda}$ becomes smaller and smaller when

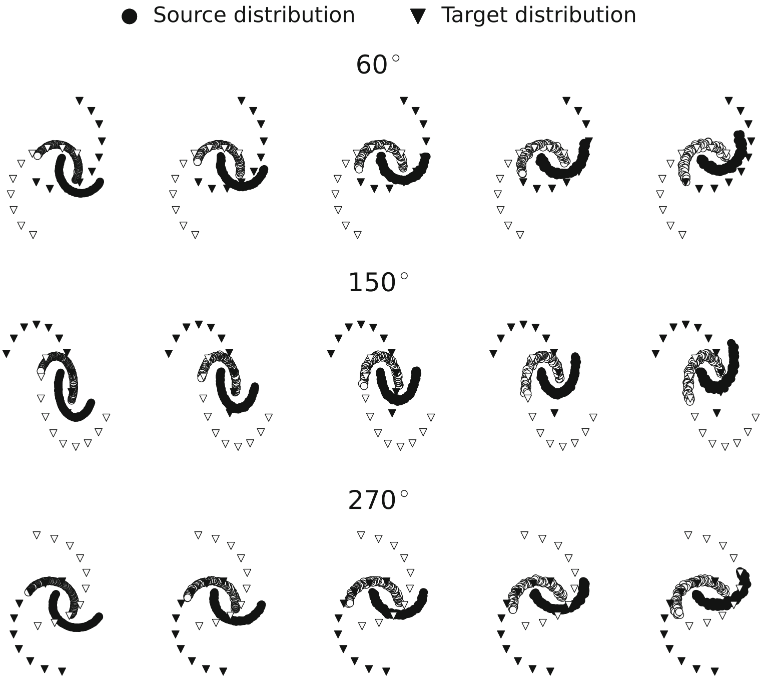

Fig. 3 Generated samples from the Moons dataset with different rotation degrees

the target labeling function approaches in shape the one considered in Fig. 2(left). This result agrees with the intuition behind the λ term explained in this study.

5.2 Moons data set

In order to go further in the empirical evaluation of our theoretical framework, we perform experiments on the well used Moons dataset described in Germain et al. (2013). Following this paper, we generate the first source domain as two entangled moons consisting of 300 data points such that each of the two moons corresponds to one of the classes. We further generate 4 more source domains by rotating the one generated previously at a random angle between $-15°$ and 15°. The target domain's data is obtained by generating 20 points using the same distribution as before and by rotating them at a given angle varying between 0° and 360°. Generated samples for several angles are given in Fig. 3.

Note that for this particular data generating procedure, it is commonly assumed that increasing the rotation angle leads to a more difficult adaptation problem. Indeed, one can observe from Fig. 3 that the source samples become more and more shifted with respect to the target samples when the rotation angle varies from 0° to 90°. For the rotation angle between 90° and 180°, the unlabeled samples start to become geometrically closer even though the source and target points that can be found in the same regions have opposite labels. Consequently, the classifier learned on source sample and applied directly on the target sample is expected to have a degrading performance for angles between 0° and 180° while the distance between the source and target samples should be bell-shaped with a peak around 90°. The process is then reversed on the range of rotation angles between 180° and

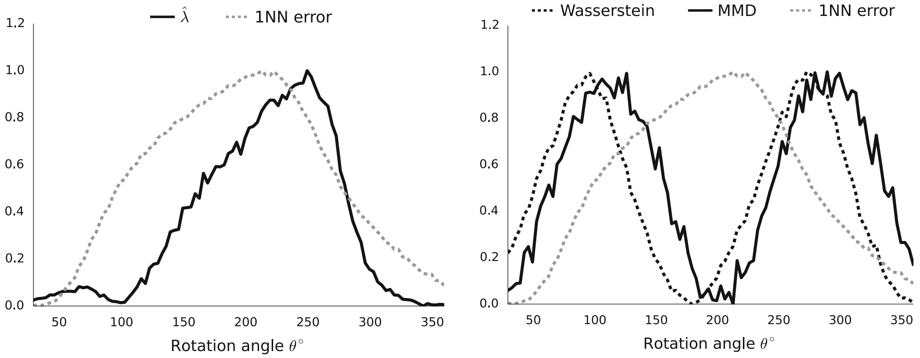

Fig. 4 (Left) $\hat{\lambda}$ and the true target error of a 1NN classifier as a function of the rotation angle; (right) Wasserstein distance, MMD distance and the true target error of a 1NN classifier as a function of the rotation angle

360°. To verify this, we resort to a standard 1NN classifier on the source domain's data and measure its error in classifying the target samples averaged over 30 randomly drawn samples. We further calculate the Wasserstein and Maximum mean discrepancy (MMD) (Smola et al. 2007) distances between the source and target samples for each rotation angle considered and the $\hat{\lambda}$ term as before. The results are presented in Fig. 4(left) and (right) for our method and the calculated divergence measures, respectively. From this figure, we can see that the proposed estimation procedure for $\hat{\lambda}$ allows to predict and captures correctly the behavior of the true target error while only the distance between the unlabeled samples fails to do so. As before, this can be explained by the fact that the considered divergence measures do not take into account the information about the labels of multiple source and target domains and thus reflects only the geometric proximity of points across two domains.

6 Conclusions and future perspectives

In this paper we proposed a new theoretical analysis for the λ term describing the a priori success of DA. The main idea of our work was to express the joint ideal error term by using the Wasserstein distance-based error function. We proved the uniqueness and the existence of the λ term for each DA task and derived new finite-sample inequalities for it. These generalization bounds cover two cases: in the first one, the considered source and target measures are supported on the real line; the second focus on the d-dimensional case with an entropic barycenter regularization. These theoretical results are quite important as, to the best of our knowledge, the term depicting the existence of the best hypothesis for source and target domains has never been thoroughly analyzed in the DA field and was the only remaining element that has never been estimated in the bounds.

Our work can be extended in different directions. We plan to investigate a possible application of the proposed analysis to multi-view and multi-task learning. In multi-view learning, the analogue of the adaptability term may be defined as the error achieved by the best hypothesis function over all views. This scenario is pretty similar to multi-source DA while the peculiarity here lies in the fact that the available views may be interdependent and that learning a good hypothesis should benefit from their interaction. In this case, the proposed analysis should be extended in order to correctly model the learning problem. In multi-task learning, we may also consider the problem of characterizing the possible benefit of learning a set of tasks simultaneously by a term defined as an agreement between the predictors that are learned for each task. In this case, similar to multi-source DA, the success of learning will naturally depend on the existence of a shared representation, where all the task are learned.

Acknowledgements Funding was provided by Agence nationale de la recherche (Grant No. ANR-15-CE23-0026).

References

Agueh, M., & Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, *43*(2), 904–924.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. *ICML*, *70*, 214–223.

Álvarez Esteban, P., del Barrio, E., Cuesta-Albertos, J. A., & Matrán, C. (2011). Uniqueness and approximate computation of optimal incomplete transportation plans. *Annales de l'Institut Henri Poincaré B: Probability and Statistics*, *47*(2), 358–375.

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2007). Analysis of representations for domain adaptation. In: NIPS (pp. 137–144).

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. (2010a). A theory of learning from different domains. *Machine Learning*, *79*, 151–175.

Ben-David, S., Lu, T., Luu, T., & Pàl, D. (2010b). Impossibility theorems for domain adaptation. *AISTATS*, *9*, 129–136.

Bigot, J., Gouet, R., Klein, T., & Lopez, A. (2018). Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line. *Electronic Journal of Statistics, 12*(02), 2253–2289.

Bigot, J., Cazelles, E., & Papadakis, N. (2018a). Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration. ArXiv e-prints arXiv:1804.08962v2.

Bigot, J., Cazelles, E., & Papadakis, N. (2018b). Penalization of barycenters in the Wasserstein space. ArXiv e-prints.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Wortman, J. (2008). Learning bounds for domain adaptation. In: NIPS.

Boissard, E., Le Gouic, T., & Loubes, J. M. (2015). Distribution's template estimate with Wasserstein metrics. *Bernoulli*, *21*(2), 740–759.

Chizat, L., Peyré, G., Schmitzer, B., & Vialard, F. X. (2015). Unbalanced optimal transport: Geometry and Kantorovich formulation. ArXiv e-prints arXiv:1508.05216v2.

Cortes, C., & Mohri, M. (2014). Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, *519*, 103–126.

Crammer, K., Kearns, M., & Wortman, J. (2008). Learning from multiple sources. *Journal of Machine Learning Research*, *9*, 1757–1774.

Cuturi , M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In: NIPS (pp. 2292–2300).

Cuturi, M., & Doucet, A. (2014). Fast computation of Wasserstein barycenters. In: ICML (pp. 685–693).

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

Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., & Poggio, T. A. (2015). Learning with a Wasserstein loss. In: NIPS (pp. 2053–2061).

Genevay, A., Peyré, G., & Cuturi, M. (2018). Learning generative models with Sinkhorn divergences. In: AISTATS (pp. 1608–1617).

Germain, P., Habrard, A., Laviolette, F., & Morvant, E. (2013). A Pac-Bayesian approach for domain adaptation with specialization to linear classifiers. In: ICML (pp. 738–746).

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In: NIPS (pp. 2672–2680).

Le Gouic, T., & Loubes, J. M. (2017). Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, *168*, 901–917.

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009a). Domain adaptation: Learning bounds and algorithms. In: COLT.

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009b). Multiple source adaptation and the Rényi divergence. In: UAI (pp. 367–374).

Margolis, A. (2011). A literature review on domain adaptation with unlabeled data. Technical report, University of Washington.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. Histoire de l'Académie Royale des Sciences (pp. 666–704).

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Redko, I., Habrard, A., & Sebban, M. (2017). Theoretical analysis of domain adaptation with optimal transport. In: ECML/PKDD (pp. 737–753).

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99–121.

Sinkhorn, R., & Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, *21*, 343–348.

Smola, A., Gretton, A., Song, L., & Schölkopf, B. (2007). A hilbert space embedding for distributions. In: ALT (pp. 13–31).

Zhang, C., Zhang, L., & Ye, J. (2012). Generalization bounds for domain adaptation. In: NIPS (pp. 3320–3328).

Zhang, Z., & Müller, H. G. (2011). Functional density synchronization. *Computational Statistics and Data Analysis*, *55*(7), 2234–2249.

Zolotarev, V. M. (1984). Probability metrics. *Theory of Probability and Its Applications*, *28*(2), 278–302.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.