Margin-aware Adversarial Domain Adaptation with Optimal Transport

Sofien Dhouib [1] **Ievgen Redko** [2] **Carole Lartizien** [1]

Abstract

In this paper, we propose a new theoretical analysis of unsupervised domain adaptation (DA) that relates notions of large margin separation, adversarial learning and optimal transport. This analysis generalizes previous work on the subject by providing a bound on the target margin violation rate, thus reflecting a better control of the quality of separation between classes in the target domain than bounding the misclassification rate. The bound also highlights the benefit of a large margin separation on the source domain for adaptation and introduces an optimal transport (OT) based distance between domains that has the virtue of being task-dependent, contrary to other approaches. From the obtained theoretical results, we derive a novel algorithmic solution for domain adaptation that introduces a novel shallow OT-based adversarial approach and outperforms other OT-based DA baselines on several simulated and real-world classification tasks.

1. Introduction

Learning to classify elements of a data set is one of the most widespread tasks in machine learning. Classically, it is done after assuming that data used for testing and those used for training a classification model stem from the same probability distribution (Sen et al., 2020). However, such an assumption is usually violated in real-world applications: product reviews classification where the difference of product types change the words distributions (Blitzer et al., 2007) or image classification where the variation of acquisition methods and conditions between training and test data introduce non-negligible distribution shifts (Hutchison et al., 2010), to name a few. While manual labeling may be considered as a tractable solution, such an approach is time consuming, can necessitate an often costly intervention of experts for labeling (e.g in medical imaging applications) and totally discards the information available on a different, yet related, labeled training set. Such a setting has motivated the emergence of domain adaptation (Pan & Yang, 2010; Weiss et al., 2016), a branch of statistical learning that takes the distribution shift into account, and in which the test data is assumed to be partially or totally unlabeled. In the literature, the training set and test set distributions are respectively termed source and target domains. In this paper, we focus on the challenging setting of unsupervised domain adaptation (Margolis, 2011), were no labels are available for the target data.

Since the inception of the domain adaptation field, several theoretical contributions were proposed to analyze this problem in the statistical learning framework (Ben-David et al., 2007b; Mansour et al., 2009; Cortes & Mohri, 2014; Germain et al., 2016; Zhang et al., 2019). The general idea behind any such analysis usually consists in bounding the target domain error rate by a source error rate plus an estimable term reflecting a certain distance between domains, called the alignment or divergence term, plus a non estimable term that is assumed to be small for adaptation to be possible. To this end, the seminal work of (Ben-David et al., 2007b) considered the bounds for 0-1 loss in binary classification setting by introducing a divergence term that takes into account the complexity of the hypothesis space. Their results were further generalized for any loss function verifying the triangle inequality in (Mansour et al., 2009) and to a case when the hypothesis space is a RKHS in (Cortes & Mohri, 2014; Cortes et al., 2019). A somewhat different result was recently proposed (Zhang et al., 2019) where the authors provided generalization bounds for DA in the case of multi-class classification with source domain error defined by the margin violation rate. Finally, several DA bounds where proposed in (Redko et al., 2016; Courty et al., 2017; Shen et al., 2017) for the specific case when the considered alignment term is given by the Wasserstein distance (Santambrogio, 2016).

At the algorithmic level, there have been a plethora of algorithms that deal with the unsupervised domain adaptation problem, and they can be roughly divided to shallow (Kouw

[1]Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69100, LYON, France [2]Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d Optique Graduate School Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Etienne, France. Correspondence to: Sofien Dhouib <sofiane.dhouib@creatis.insa-lyon.fr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

& Loog, 2019; Zhang, 2019) and deep (Wilson & Cook, 2019) methods. Most of shallow methods try to solve the problem in a two-step fashion by first aligning the source and target domains to make them indistinguishable, which then allows to apply classical supervised algorithms on the transformed data. Such an alignment is usually performed via instance reweighing (Sugiyama et al., 2007) or feature space transformations (Gong et al., 2012; Fernando et al., 2014; Sun et al., 2016). A notable recent approach to perform alignment is the use of optimal transport (Courty et al., 2015; 2017), that provides a well-funded way of finding a mapping aligning the source and target domains that minimizes the cost of transforming the source distribution into the target one. Deep domain adaptation methods have also known an impressive surge in their number, with the basic idea being the exploitation of their feature extraction capacity to learn representations that align the two domains, while distinguishing between the different classes of the source domain (Tzeng et al., 2014). One of the main reasons of this surge is the adversarial training procedure (Goodfellow et al., 2014) for the first time used for domain adaptation in (Ganin & Lempitsky, 2014), where the main idea is built upon the theoretical contribution in (Ben-David et al., 2007b).

In this paper, we provide a novel theoretical study of the unsupervised domain adaptation problem that provides the following contributions to the field:

1. We bound the margin violation rate in the target domain by its counterpart from the source domain, a novel symmetric alignment term and a non estimable term that shows the benefit of large margin separation on source domain for the success of adaptation. This result includes the work of (Ben-David et al., 2007a) as a special case, does not require the loss function to satisfy the triangle inequality as in (Mansour et al., 2009) and strengthens the result of (Zhang et al., 2019) by replacing the misclassification target error with a stricter target margin violation rate.
2. We upper bound our alignment term by a distance defining the minimax variation of the classic Monge-Kantorovitch problem (Santambrogio, 2016). This latter is further shown to be upper-bounded by the original Wasserstein distance considered in (Redko et al., 2016; Courty et al., 2017; Shen et al., 2017) thus leading to tighter bounds.
3. We derive a first OT-driven adversarial DA algorithm that outputs a classifier minimizing the estimable part of the obtained bounds. This classifier is shown to outperform other OT-based DA methods on both synthetic and real-world data sets.

The rest of the paper is organized as follows. Section 2 introduces required preliminary knowledge and notations. Section 3 is dedicated to our theoretical contributions presenting a novel bound on the margin violation error on the target domain, its thorough analysis and relation to other existing bounds. Then, in Section 4, we use it to derive an algorithm that is further specialized to linear classifiers, resulting in a convex programming formulation. Finally, in the last section, we evaluate our algorithm on a toy data set and on a benchmark real-world problem.

2. Preliminary Knowledge

In this section, we present the problem setup of our study with the notations used throughout the paper. We also provide background knowledge on learning bounds in DA to allow a further comparison to them in the rest of the manuscript.

2.1. Problem setup and notations

We consider a binary classification setting, in which source and target data are respectively drawn from $\mathcal{S}$ and $\mathcal{T}$, the joint distributions over the product space of instances and labels $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$. We denote their corresponding marginal distributions as $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$ and use bold upper-case letters for matrices (e.g., $\mathbf{D}$) and bold lower-case letters for vectors (e.g. $\mathbf{x}$). Although both domains are assumed to be labeled, only the labels of the source instances are observable during the learning stage. This setting is often referred to as unsupervised domain adaptation.

To proceed, let $\mathcal{H}$ and $\mathcal{H}'$ denote two compact classes of hypotheses acting on $\mathcal{X}$ and taking values in $[-1, 1]$. For further developments, we define several quantities to assess classifiers' performances on different domains. Let $l^{\rho,\beta}$ be the loss function defined by

$$l^{\rho,\beta}(t) := \begin{cases} 1 - \frac{(t-\rho)}{\beta}, & \text{if } \rho \leq t \leq \beta + \rho \\ [t < \rho], & \text{otherwise} \end{cases}$$

where $1 > \rho, \beta > 0$, and $[\cdot]$ denotes the Iverson bracket for indicator functions. From its definition, we note that $l^{\rho,0}(t) = [t < \rho]$ and that it verifies the following inequality for all $\rho, \beta > 0$ and $t \in \mathbb{R}$

$$l^{\rho,0}(t) = [t < \rho] < l^{\rho,\beta}(t) < l^{\rho+\beta,0}(t) = [t < \rho + \beta] \quad (1)$$

illustrated in Figure 1.

For any domain $\mathcal{P}$ with marginal feature distribution $\mathcal{D}_\mathcal{P}$ and any hypotheses h, f, we define their disagreement associated to the loss $l^{\rho,\beta}$ as

$$\epsilon_\mathcal{P}^{\rho,\beta}(h, f) := \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}_\mathcal{P}} \left[l^{\rho,\beta}(f(\mathbf{x})h(\mathbf{x})) \right].$$

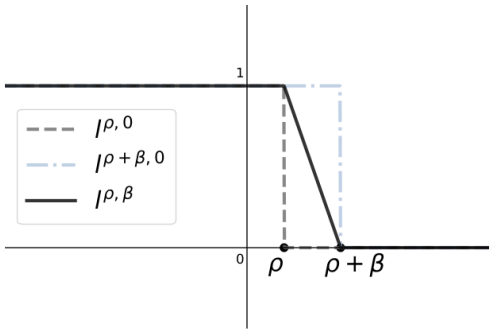

Figure 1. Loss function $l^{\rho,\beta}$ with its characteristic points and an illustration of the property from Equation (1).

This quantity can be further generalized to non deterministic hypotheses that define the labeling of domain $\mathcal{P}$, in which case the expectation is taken over $\mathcal{P}$:

$$\epsilon_{\mathcal{P}}^{\rho,\beta}(h) := \mathop{\mathbb{P}}_{\mathbf{x},y\sim\mathcal{P}}\left[l^{\rho,\beta}(yh(\mathbf{x}))\right].$$

This definition stands for the classification risk on $\mathcal{P}$: for $\beta = 0$, it is the $\rho-$margin violation rate measuring the probability of the event $\{yh(\mathbf{x}) < \rho\}$, while for $\rho = \beta = 0$, it is the 0-1 or misclassification rate.

2.2. Background on DA theory

All previous analyses of the DA problem aimed at bounding the target domain error by its source counterpart, an estimable divergence term and a non estimable term representing the a priori adaptability of the problem. Consequently, the major differences between the different available results lie in the considered definition of the error, the divergence measure they introduce and the form of the non estimable term.

To this end, the first rigorous theoretical analysis of domain adaptation, presented in (Ben-David et al., 2007b) (and later in (Ben-David et al., 2010)), introduces the alignment term given by the $\mathcal{H}\Delta\mathcal{H}-$divergence and the ideal joint error λ, respectively defined for a binary hypotheses class $\mathcal{H}$ as follows:

$$d_{\mathcal{H}\Delta\mathcal{H}} := 2\sup_{h,h'\in\mathcal{H}}\left|\epsilon_{\mathcal{S}}^{0,0}(h,h') - \epsilon_{\mathcal{T}}^{0,0}(h,h')\right|, \quad (2)$$

$$\lambda := \inf_{f\in\mathcal{H}} \epsilon_{\mathcal{S}}^{0,0}(f) + \epsilon_{\mathcal{T}}^{0,0}(f)$$

Note that the $\mathcal{H}\Delta\mathcal{H}$-divergence between the two domains does not require strict equality between distributions over all measurable sets, but only over supports of hypotheses of the $\mathcal{H}\Delta\mathcal{H}$ space defined as the space of symmetric differences between hypotheses in $\mathcal{H}$: if $h, h' \in \mathcal{H}$ take their values in $\{0,1\}$, it is simply equal to hypothesis $\{|h-h'|\}$, i.e., the hypothesis that takes value 1 if and only if h and h' disagree. The authors show that this quantity is estimable from finite samples and that the sample complexity of such estimation depends on the VC-dimension of the considered hypothesis space $\mathcal{H}$. As for λ, it cannot be estimated as it involves the target domain's labels, to which one has no access in the setting of unsupervised domain adaptation.

Remark 1. *If $h, h' \in \mathcal{H}$ take their values in $\{-1,1\}$, their disagreement at a point $\mathbf{x}$ is equal to $-h(\mathbf{x}).h'(\mathbf{x})$, as this latter equals 1 if and only if $h(\mathbf{x})$ and $h'(\mathbf{x})$ have opposite labels.*

Remark 2. *A similar bound for 0-1 loss with the Wasserstein distance and the same λ term was proved in (Shen et al., 2017).*

This result was further generalized in (Mansour et al., 2009) where the authors considered an arbitrary symmetric loss function l verifying the triangle inequality. This assumption makes their result more general than the one in (Ben-David et al., 2010) proved only for 0-1 loss. After considering the expected $l-$disagreement between two hypotheses on a marginal distribution $\mathcal{D}_{\mathcal{P}}$ defined as

$$\mathcal{L}_{\mathcal{D}_{\mathcal{P}}}(h,h') := \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{P}}}\left[l(h(\mathbf{x}),h(\mathbf{x}')\right],$$

the authors define their discrepancy distance as

$$\mathrm{disc}_l(\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{T}}) = \sup_{h,h'\in\mathcal{H}}|\mathcal{L}_{\mathcal{D}_{\mathcal{S}}}(h,h') - \mathcal{L}_{\mathcal{D}_{\mathcal{T}}}(h,h')|.$$

Such a generalization to arbitrary loss function was later also provided for the bounds with the Wasserstein distance between joint (Courty et al., 2017) and marginal domains distributions (Redko et al., 2016). While being more general than the bound of (Ben-David et al., 2010)[1], these bounds, however, do not cover the margin violation loss as this latter does not verify the triangle inequality.

Finally, in a more recent work the authors of (Zhang et al., 2019) generalized the bounds mentioned above to the multi-class setting and introduced a classification margin $\beta > 0$ into their results (Koltchinskii & Panchenko, 2004). Below, we present their alignment term and their ideal joint hypothesis term, restricted to the case of binary classification with labels encoded in $\{-1,1\}$:

$$d_{h,\mathcal{H}}^{(\beta)} := \sup_{h'\in\mathcal{H}}\left(\epsilon_{\mathcal{T}}^{0,\beta}(\mathrm{sgn}\,(h),h') - \epsilon_{\mathcal{S}}^{0,\beta}(\mathrm{sgn}\,(h),h')\right), \quad (3)$$

$$\lambda^{(\beta)} := \inf_{f\in\mathcal{H}} \epsilon_{\mathcal{S}}^{0,\beta}(f) + \epsilon_{\mathcal{T}}^{0,\beta}(f).$$

The alignment term in (3) involves a supremum over one hypothesis instead of two, making it lower than $\mathcal{H}\Delta\mathcal{H}$-divergence defined in Equation (2) for $\beta = 0$. It also offers

[1] As mentioned in (Mansour et al., 2009), the bounds based on the discrepancy distance are in general incomparable to those of (Ben-David et al., 2007b).

new insights on domain adaptation problem by introducing the margin violation rate and scoring functions that give the confidence level of belonging to a class of interest rather than functions with binary output. However, as they bound the 0-1 loss on the target domain, i.e., $\epsilon_{\mathcal{T}}^{0,0}(h, f)$, their bound does not indicate the behaviour of the margin violation rate on this latter. As for $\lambda^{(\beta)}$, it remains conceptually similar to the λ term of the other bounds with the only difference consisting in the definition of the error terms.

We now proceed to the presentation of our main theoretical contributions.

3. Bounding the Target Margin Violation Risk

This section is dedicated to our theoretical contributions. We begin by bounding the margin violation rate on the target domain for a classifier h picked from a given hypothesis class $\mathcal{H}$. This bound introduces a new divergence term for which we provide a convex proxy afterwards. All of the proofs can be found in the supplementary material.

3.1. Bound with non convex divergence between distributions

The theorem below aims at providing a first theoretical result for DA that includes only interconnected terms depending on the margin of the considered hypothesis. Such interdependence allows to better highlight the possible trade-offs between the different terms in the bound and to gain new insights into the conditions leading to a successful adaptation.

Theorem 1. *Assume that for any $h' \in \mathcal{H}'$, we have $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}}[h'(\mathbf{x}) = 0] = \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{T}}[h'(\mathbf{x}) = 0] = 0$. Let $\rho, \beta, \alpha > 0$ be such that $\rho + \beta < \alpha < 1$. Then, for any $h \in \mathcal{H}$, the following bound holds:*

$$\epsilon_{\mathcal{T}}^{\rho,0}(h) \leq \epsilon_{\mathcal{S}}^{\frac{\rho+\beta}{\alpha},0}(h) + d_{h,\mathcal{H}'}^{\rho,\beta}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T}) + \lambda_\alpha$$

where

$$d_{h,\mathcal{H}'}^{\rho,\beta}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T}) := \sup_{h'\in\mathcal{H}'} \left|\epsilon_{\mathcal{S}}^{\rho,\beta}(h, h') - \epsilon_{\mathcal{T}}^{\rho,\beta}(h, h')\right|$$

and

$$\lambda_\alpha := \inf_{f\in\mathcal{H}'} \epsilon_{\mathcal{T}}^{0,0}(f) + \epsilon_{\mathcal{S}}^{0,0}(f) + \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}}[|f| < \alpha].$$

Proof idea. The proof uses the fact that for any $s, t, u, v \geq 0$, $st \leq uv \Rightarrow s \leq u$ or $t \leq v$ in order to bound probabilities of margin violation. The rest follows from the techniques used to establish the bound in (Ben-David et al., 2010) and on inequality (1). □

While being similar in shape to the previous DA bounds (Ben-David et al., 2007b; Mansour et al., 2009; Zhang et al., 2019), the obtained result has several fundamental distinctions. First, our bound concerns the margin violation rate on the target domain, rather than the misclassification rate used in Equation (3) (Zhang et al., 2019) and thus, it offers a better estimation of the quality of separation between the classes in the target domain. Second, its estimable part given by the alignment term $d_{h,\mathcal{H}'}^{\rho,\beta}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T})$ does not introduce the decision function associated to a hypothesis (its sign in case of binary classification with labels encoded as -1 and 1) as in Equation (3) (Zhang et al., 2019), and thus avoids discontinuities in this term making it more suitable for optimization algorithms. One can further show that it is an integral probability metric (Müller, 1997) as for a fixed $h \in \mathcal{H}$, the set $\{\mathbf{x} \mapsto l^{\rho,\beta}(h(\mathbf{x})h'(\mathbf{x}); h' \in \mathcal{H}'\}$ is a class of bounded measurable functions implying that $d_{h,\mathcal{H}'}^{\rho,\beta}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T})$ is a pseudometric on the space of probability distributions over $\mathcal{X}$. Finally, the non estimable term λ_α is non symmetric with respect to the the source and target domains' roles as, in addition to having low errors in both domains, it requires *only* a large absolute margin on the source domain, reflected by $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}}[|f_\alpha(\mathbf{x})| < \alpha]$ where f_α is a function achieving the minimum of λ_α. This latter thus reflects an intuitively understandable behaviour: if one has a large margin of separation on $\mathcal{S}$, i.e., there exists α large enough with $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}}[|f_\alpha(\mathbf{x})| < \alpha]$ small enough, the space of classifiers that are good for the source becomes bigger. Consequently, it becomes more likely to find among them a classifier that is not only good on the source, but on the target domain too. This claim is also supported by the fact that for fixed $\rho, \beta \geq 0$, the source error term is the violation rate of margin $\frac{\beta+\rho}{\alpha}$: as α increases, that rate decreases, implying that less concentration on the performance in the target domain is needed. Of course, no gain can be obtained from augmenting α if it considerably increases λ_α.

Additionally, compared to non estimable terms from previous works, λ_α is at least equal to its counterpart λ from (Ben-David et al., 2010), and is not directly comparable to $\lambda^{(\rho)}$ from (Zhang et al., 2019). However, it verifies the following inequality:

$$\lambda_\alpha \leq \min_{f\in\mathcal{H}'} \epsilon_{\mathcal{S}}^{\alpha,0}(f) + \epsilon_{\mathcal{T}}^{\alpha,0}(f)$$

This is not very different from $\lambda^{(\rho)}$, as this latter minimizes the joint $l^{0,\rho}$ risk, while the right hand side in the previous bound minimizes the joint rate of violating margin α. Finally β, for the moment, appears as the cost of the Lipschitz property of the loss function used in defining the discrepancy term. Its role will become clear when we use it to bound our alignment term by a convex one.

In the following corollary, we formerly link our bound to that of (Ben-David et al., 2007b).

Corollary 1. *If $\mathcal{H} = \mathcal{H}'$ is a class of binary hypotheses*

taking values in $\{-1, 1\}$, the bound from Theorem 1 is equivalent to the one in (Ben-David et al., 2010).

This corollary shows that our bound generalizes that of (Ben-David et al., 2010) to more informative scoring functions, and to the margin violation rate criterion, without requiring the loss function to verify the triangle inequality as in (Mansour et al., 2009).

3.2. A convex domain divergence based on optimal transport

Although the previous bound offers several novel insights on the behaviour of the target error with respect to different components of the bound, its estimable part is non convex as a function of hypothesis $h \in \mathcal{H}$. In order to convexify its two components, i.e., the error on the source domain and the divergence term, it is sufficient to use a convex loss proxy for the first, while for the second we propose to leverage optimal transport (OT) theory (Santambrogio, 2016). To this end, let us introduce two projection operators $\pi_1 : (\mathbf{x}_1, \mathbf{x}_2) \mapsto \mathbf{x}_1$ and $\pi_2 : (\mathbf{x}_1, \mathbf{x}_2) \mapsto \mathbf{x}_2$ defined over $\mathcal{X} \times \mathcal{X}$. The set Π of transport plans between $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$ is the set of probability distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{X}$ that verify the following two properties:

$$\pi_1 \# \mathcal{D} = \mathcal{D}_\mathcal{S}, \qquad \pi_2 \# \mathcal{D} = \mathcal{D}_\mathcal{T},$$

where $\#$ denotes the pushforward measure. With the previous notations, the convex bound for $d^{\rho,\beta}_{h,\mathcal{H}'}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T})$ is given in the following proposition.

Proposition 1 (Convex bound for alignment term)**.** *For any $\rho, \beta > 0$, we have*

$$d^{\rho,\beta}_{h,\mathcal{H}'}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T}) \leq \frac{1}{\beta} \inf_{\mathcal{D} \in \Pi} \Delta_{\mathcal{H}'}(h, \mathcal{D})$$

where

$$\Delta_{\mathcal{H}'}(h, \mathcal{D}) := \sup_{h' \in \mathcal{H}'} \mathop{\mathbb{E}}_{\mathbf{x}_s, \mathbf{x}_t \sim \mathcal{D}} \left[|hh'(\mathbf{x}_s) - hh'(\mathbf{x}_t)| \right].$$

Proof idea. From the $\frac{1}{\beta}-$Lipschitz property of $l^{\rho,\beta}$, we use the dual form of the Wasserstein distance between distributions $hh'\#\mathcal{D}_\mathcal{S}$ and $hh'\#\mathcal{D}_\mathcal{T}$, then its primal form (with an infimum), and finally the inf-sup inequality. □

To see the convexity of the introduced alignment term, we note that $d^{\rho,\beta}_{h,\mathcal{H}'}$ can be bounded by $\Delta_{\mathcal{H}'}(h, \mathcal{D})$ for any transport plan $\mathcal{D}$. Furthermore,

$$\{h \mapsto \mathop{\mathbb{E}}_{\mathbf{x}_s, \mathbf{x}_t \sim \mathcal{D}} \left[|hh'(\mathbf{x}_s) - hh'(\mathbf{x}_t)| \right], h' \in \mathcal{H}'\}$$
$$\{\mathcal{D} \mapsto \mathop{\mathbb{E}}_{\mathbf{x}_s, \mathbf{x}_t \sim \mathcal{D}} \left[|hh'(\mathbf{x}_s) - hh'(\mathbf{x}_t)| \right], h' \in \mathcal{H}'\}$$

defined for a fixed $\mathcal{D} \in \Pi$ and a fixed $h \in \mathcal{H}$, respectively are two families of convex functions in h and in $\mathcal{D}$. Thus, taking the supremum over $h' \in \mathcal{H}'$ is as well convex in h (resp. in $\mathcal{D}$). We note that the function $h \mapsto \inf_{\mathcal{D} \in \Pi} \Delta_{\mathcal{H}'}(h, \mathcal{D})$ is not necessarily convex, but when we derive our algorithm, we show that only convexity of $\Delta_{\mathcal{H}'}(\cdot, \cdot)$ is needed.

The bound in Proposition 1 also has the form of a robust version of the Wasserstein distance between 1D distributions $hh'\#\mathcal{D}_\mathcal{S}$ and $hh'\#\mathcal{D}_\mathcal{T}$ and admits the following adversarial interpretation: for a fixed joint distribution $\mathcal{D} \in \Pi$, taking the supremum over $h' \in \mathcal{H}'$ is trying to separate the two domains, while taking $\mathcal{D}$ that achieves the infimum resists to this separation.

Combined with Theorem 1, Proposition 1 allows to immediately deduce a domain adaptation bound involving the introduced OT-based divergence.

Proposition 2 (Optimal transport bound on the target risk)**.** *With the assumptions and notations of Theorem 1 and Proposition 1, we have for any $h \in \mathcal{H}$:*

$$\epsilon^{\rho,0}_\mathcal{T}(h) \leq \epsilon^{\frac{\rho+\beta}{\alpha},0}_\mathcal{S}(h) + \frac{1}{\beta} \inf_{\mathcal{D} \in \Pi} \Delta_{\mathcal{H}'}(h, \mathcal{D}) + \lambda_\alpha. \quad (4)$$

Compared to other DA bounds involving the Wasserstein distance (Redko et al., 2016; Courty et al., 2017; Shen et al., 2017), our divergence term takes into account the considered hypothesis classes, making it a pseudo-metric that is less strict than the Wasserstein distance between marginal distributions of the domains. To support this claim, we bound our optimal transport based alignment term by the Wasserstein distance between the two domains.

Proposition 3 (Bounding by the Wasserstein distance)**.** *Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a metric, and assume that all of the hypotheses from $\mathcal{H}$ and $\mathcal{H}'$ verify the $L-$Lipschitz continuity with respect to metric c for some $L > 0$. Then, the following holds*

$$\sup_{h \in \mathcal{H}} \inf_{\mathcal{D} \in \Pi} \Delta_{\mathcal{H}'}(h, \mathcal{D}) \leq 2LW_1(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T})$$

where

$$W_1(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T}) := \inf_{\mathcal{D} \in \Pi} \mathop{\mathbb{E}}_{\mathbf{x}_s, \mathbf{x}_t \sim \mathcal{D}} \left[c(\mathbf{x}_s, \mathbf{x}_t) \right] \quad (5)$$

is the Wasserstein distance associated to metric c.

Proof idea. The proof relies on the triangle inequality verified by the absolute value, the $L-$Lipschitzness and the boundedness of the classifiers, as well as the monotonicity of the inf and the sup operators. □

This inequality comes essentially from the fact that the space of $L-$Lipchitz functions is bigger than the considered hypothesis spaces $\mathcal{H}$ and $\mathcal{H}'$, and the supremum over $h \in \mathcal{H}$ is due to the independence of the Wasserstein distance

between distributions of classifier h. It formally shows that the attachment of our alignment term to the task at hand, via $h \in \mathcal{H}$ and the supremum over $h' \in \mathcal{H}'$, makes it far less strict than the Wasserstein distance between data marginal distributions of the domains.

4. Domain Adaptation Algorithm

With our considered setting and notations, the goal of our domain adaptation task is to find $h \in \mathcal{H}$ such that $\epsilon_{\mathcal{T}}^{\rho,0}(h)$, the margin violation rate on the target domain, is as small as possible. As we assume that no access to the labels of $\mathcal{T}$ is given, we look for a hypothesis that minimizes the estimable part of our bound of Proposition 2.

4.1. Minimizing the estimable part of the bound

To derive the objective function for the estimable part of the bound, we have to consider two terms. First is the margin violation error on the source domain $\epsilon_{\mathcal{S}}^{\rho'}(h)$ (where $\rho' = \frac{\rho+\beta}{\alpha}$) whose minimization is known to be a NP-hard problem (Arora et al., 1997). Hence, we replace it by its commonly used proxy, the hinge risk $\underset{\mathbf{x},y\sim\mathcal{S}}{\mathbb{E}}[(\rho' - y.h(\mathbf{x}))_+]$. The second term is the domain alignment term, which is the infimum over $\mathcal{D} \in \Pi$ of convex functions of h, as we mentioned earlier. The resulting optimization problem is then given as follows:

$$\min_{\substack{h\in\mathcal{H}\\ \mathcal{D}\in\Pi}} \underset{\mathbf{x},y\sim\mathcal{S}}{\mathbb{E}}[(\rho' - y.h(\mathbf{x}))_+] + \frac{1}{\beta}\Delta_{\mathcal{H}'}(h,\mathcal{D}). \quad (6)$$

We would like to stress out that the cost function in this case contains a supremum over the potentially infinite hypothesis space $\mathcal{H}'$, hence, it might be difficult to solve problem (6), even though convexity is verified. However, we show in the next section that a particular choice of $\mathcal{H}$ and $\mathcal{H}'$ allows one to deal efficiently with this term.

Below, we specify the proposed method to a particular case of linear classifiers thus introducing a shallow adversarial DA approach.

4.2. Application to linear classification

We consider our algorithm's formulation in the linear classification case, where $\mathcal{H}$ is the space of ℓ^2 bounded linear classifiers, and $\mathcal{H}'$ is the space of ℓ^1 bounded classifiers.

Proposition 4. *Let $\mathcal{H}$ be the space of linear classifier with bounded ℓ^2 norm, and $\mathcal{H}'$ be the space of linear classifiers with bounded ℓ^1 norm. Let l denote the hinge loss, $\mathbf{D}_{st} := \mathbf{x}_s\mathbf{x}_s^T - \mathbf{x}_t\mathbf{x}_t^T$ and $(|\mathbf{D}_{st}\mathbf{w}|)_i := |(\mathbf{D}_{st}\mathbf{w})_i|$ for $1 \leq i \leq d$, where $\mathbf{w} \in \mathbb{R}^d$. Then, Problem (6) can be equivalently expressed as the following convex program:*

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^n\\ \mathcal{D}\in\Pi}} \underset{\mathbf{x},y\sim\mathcal{S}}{\mathbb{E}}\left[l(y.\mathbf{w}^T\mathbf{x})\right] + \delta\left\|\underset{\mathbf{x}_s,\mathbf{x}_t\sim\mathcal{D}}{\mathbb{E}}\left[|\mathbf{D}_{st}\mathbf{w}|\right]\right\|_\infty + \zeta\|\mathbf{w}\|_2^2 \quad (7)$$

where $\delta, \zeta > 0$ are two hyper-parameters related to the bounds on $\mathcal{H}$ and $\mathcal{H}'$.

Proof idea. The supremum over $\mathcal{H}'$ is a supremum of a convex function on the ℓ_1 unit ball, hence over its vertices which appears via the $\infty-$norm. The regularization is to take into account the boundedness of classifiers from $\mathcal{H}$. □

The previous proposition introduces two hyper-parameters linked to regularization of the classifier and to the alignment term. The further the domains are from each other, the more concentration we need on the alignment term, which is achieved by increasing δ. Also, we note that this is a strongly convex optimization problem, due to the strong convexity of the regularization $\|\mathbf{w}\|^2$ which is an important feature for numerical optimization with gradient descent, thus justifying our choice of the space $\mathcal{H}$.

4.3. Learning in similarity induced spaces

In the previous section, we chose $\mathcal{H}'$ as the space of linear classifiers with bounded ℓ^1 norm in order to tackle the difficulty of computing the supremum over $h' \in \mathcal{H}'$, transforming it into a maximum over the vertices of the unit ℓ^1 ball. However, the main idea behind this choice is not only to simplify the computation of the supremum, but also to have a theoretical justification for our algorithm when the data lies in a space induced by an $(\epsilon,\gamma,\tau)-$good similarity function, introduced in the seminal paper (Balcan et al., 2008). Indeed, this latter theoretically guarantees the existence of a low-error linear classifier with bounded ℓ^1 norm in the similarity space defined by transformation

$$\Psi(\mathbf{x}) = (K(\mathbf{x},\tilde{\mathbf{x}}_1), ..., K(\mathbf{x},\tilde{\mathbf{x}}_L)) \quad (8)$$

where $\{\tilde{\mathbf{x}}_1, ..., \tilde{\mathbf{x}}_L\}$ is a finite set of landmarks usually defined as a subset of the original data set.

Hence, if we consider all of the points from both the source and target domains to be landmarks, then the existence of a good similarity function able of separating the classes implies the existence of an ℓ_1-bounded classifier that performs well on both domains, i.e., an ideal joint hypothesis with a low error in the similarity induced space. Thus, in the next experimental section, we solve Problem (7) after applying the mapping Ψ, but with a regularization term $\|\mathbf{w}\|_2^2$. Note that this latter that does not depend on the similarity matrix $(\mathbf{K})_{ij} := K(\mathbf{x}_i,\mathbf{x}_j)$ for $\mathbf{x}_i,\mathbf{x}_j \in \mathcal{X}$, unlike in the case of kernelized approaches where it is equal to $\mathbf{w}^T\mathbf{K}\mathbf{w}$, with $\mathbf{K}$ being the kernel matrix. We note further that in

(Balcan et al., 2008), the authors recommend bounding ℓ_1 norm of $\mathbf{w}$ as a constraint. However, this is equivalent to bounding its ℓ_2 norm due to norm equivalence in finite dimension, which in turn is equivalent to adding a quadratic regularization. This latter is more suitable for optimization as we indicate later.

4.4. Optimization procedure of the discrete problem

In the empirical case, one has access to finite data sets $\mathbf{S}_m = \{((\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\} \sim \mathcal{S}^m$, and $\mathbf{T}_n = \{(\mathbf{x}'_1, y'_1), ..., (\mathbf{x}'_n, y'_n)\} \sim \mathcal{T}^n$, where the labels of $\mathbf{T}_n$ are not used for learning classifier $\mathbf{w}$. We define $\hat{\Pi}$, the empirical counterpart of Π, as the set:

$$\hat{\Pi} = \{\mathbf{\Gamma} \in \mathbb{R}_+^{m \times n}; \mathbf{\Gamma}\mathbf{1}_n = \frac{1}{m}\mathbf{1}_m; \mathbf{\Gamma}^T\mathbf{1}_m = \frac{1}{n}\mathbf{1}_n\}$$

where $\mathbf{1}_p = (1, 1, ..., 1) \in \mathbb{R}^p$. Denoting $\mathbf{D}_{ij} = \mathbf{x}_i\mathbf{x}_i^T - \mathbf{x}'_j{\mathbf{x}'_j}^T \in \mathbb{R}^{d \times d}$, the empirical cost function of Problem (7) becomes:

$$\frac{1}{m}\sum_{1 \le i \le m} l(y_{s,i}.\mathbf{w}^T\mathbf{x}_{s,i}) + \delta \left\| \sum_{\substack{1 \le i \le m \\ 1 \le j \le n}} \gamma_{ij}|\mathbf{D}_{ij}\mathbf{w}| \right\|_\infty + \zeta\|\mathbf{w}\|_2^2 \tag{9}$$

which is a function of $\mathbf{w} \in \mathbb{R}^d$ and $\mathbf{\Gamma} \in \hat{\Pi}$ having elements γ_{ij}.

Similar to (Courty et al., 2017), the objective function of our minimization problem is convex in two sets of variables: the classifier $\mathbf{w}$ and a transport matrix $\mathbf{\Gamma}$. Following their procedure, we use block coordinate descent (Grippo & Sciandrone, 2000) which alternates between the two steps:

1. For a fixed transport matrix $\mathbf{\Gamma}$, minimize over $\mathbf{w}$. To this end, we use the L-BFGS quasi-Newton method.

2. For a fixed linear classifier $\mathbf{w}$, the minimization over $\mathbf{\Gamma}$ only involves the term multiplied by δ in (9), and due to the positivity of all coordinates of vector $\sum_{ij} \gamma_{ij}|\mathbf{D}_{ij}\mathbf{w}|$, this minimization is equivalent to:

$$\min_{\mathbf{q} \in \Delta_d} \max_{\mathbf{\Gamma} \in \hat{\Pi}} \left(-\sum_{ij} \gamma_{ij}\mathbf{q}^T|\mathbf{D}_{ij}\mathbf{w}| \right)$$

   where Δ_d is the probability simplex in dimension d. In this case, we use the min-max algorithm from Blankenship & Falk (1976, Algorithm 2.2) to find the optimal transport matrix $\mathbf{\Gamma}$.

We use smooth proxies of the positive part $(\cdot)_+$, the absolute value $|\cdot|$ and the infinite norm $\|\cdot\|_\infty$ (for more details, we refer the interested reader to the supplementary material).

5. Empirical Evaluation

In this section, we evaluate our method on two domain adaptation problems: a toy set with controllable adaptation difficulty and a real-world sentiment analysis problem. For all experiments, we use the version of our algorithm specialized to linear classifiers as described in problem (7). We further use a similarity function K to be specified for each data set considered as in Equation (8) to calculate the features from the raw data. Finally, we denote our method by **MADAOT** following the abbreviation of our paper's title. The code for the different experiments is available on this link[2].

5.1. Hyper-parameter tuning

Hyper-parameter tuning is a longstanding problem in unsupervised domain adaptation that was mainly addressed by the introduction of the reversed validation procedure proposed in (Zhong et al., 2010; Bruzzone & Marconcini, 2010). Although this latter may seem as the most suitable cross-validation procedure for the unlabeled scenario, it was shown to fail at selecting the best hyper-parameters for several methods (Wilson & Cook, 2019, Section 8.2), (Bousmalis et al., 2016). One possible reason for this failure is its dependence on accurate estimation of the ratio between the marginal distributions that was proved to require a very large number of samples to be approximated correctly (Ben-David & Urner, 2012).

Hence, we choose to present our algorithm's performance for two cases. In the first one, we do not use target labels during training phase, but we use them as a validation set to select the best hyper-parameters (defined in Proposition 4) via a 5-fold cross-validation procedure for 10 values of δ ranging from 10^{-2} to 10^2, and 10 values for ζ from $\{10^{-6}$ to 10^{-2}, both on a logarithm scale. This is a rather standard procedure in unsupervised domain adaptation used in several other papers on the subject (Courty et al., 2015; Bousmalis et al., 2016). We use this procedure for the first dataset. As for the real-world data set, we run all experiments by setting $\delta = 1$ and $\zeta = 10^{-5}$.

5.2. Inter-twinning moons data set

We carry on our experiments on the moons data set used in (Courty et al., 2015). For this data set, the source domain's data sample is represented by two inter-twinning moons centered at the origin $(0, 0)$, and composed of 300 instances. The source domain's data are then rotated around their center by a certain angle to get the target domain data. Obviously, the greater is the angle, the further from each other the two domains are and the harder is the adaptation. Similar to (Courty et al., 2015), we cope with the non-linearity of this

[2] https://github.com/sofiendhouib/MADAOT.

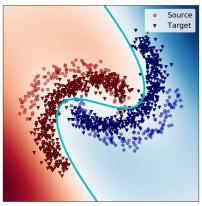

Table 1. Average accuracy (percentage) over 10 realizations for the moons toy set.

Angle (°)	10	20	30	40	50	70	90
SVM (Courty et al., 2015)	**100**	89.6	76	68.8	60	26.6	17.2
OT-GL (Courty et al., 2015)	**100**	**100**	**100**	**98.7**	80.4	62.2	49.2
JDOT (Courty et al., 2017)	98.9	95.5	90.6	86.5	81.5	**70.5**	**60**
MADAOT	**99.5**	**99.3**	**99.6**	**99.6**	**98.9**	**77**	**64.1**

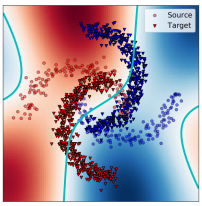

Table 2. Accuracy on the Amazon Reviews data set.

Task	B → D	B → E	B → K	D → B	D → E	D → K	E → B	E → D	E → K	K → B	K → D	K → E
SVM (CV)	79.5	69.2	71.4	**78.5**	71.7	76.6	71	73.8	85.3	72.3	73.4	84.4
DANN	**80.6**	74.7	76.7	74.7	73.8	76.5	71.8	72.6	85	71.8	73	**84.7**
OT-GL	75.7	75.4	77.9	75.9	70	78	71.2	69.6	81.4	**73.1**	74.1	81.7
JDOT_{SVM}	75	**77.1**	77.9	70.9	**78.3**	78	65.9	71.7	79.4	66.8	66.1	77.5
JDOT_{NN}	79.5	**78.1**	**79.4**	76.3	**78.8**	**82.1**	**74.9**	**73.7**	**87.2**	72.8	**76.5**	84.5
MADAOT	**82.4**	75	**80.4**	**80.9**	73.5	**81.5**	**77.2**	**78.1**	**88.1**	**75.6**	**75.9**	**87.1**

data set by using a Gaussian kernel as similarity function K, where the width parameter is chosen as the mean Euclidean distance between the source instances, as suggested in (Kar & Jain, 2011). As for the baselines, our algorithm is compared to SVM classifier with Gaussian kernel trained on the source domain (without adaptation) and two optimal transport based domain adaptation algorithms OT-GL (Courty et al., 2015) and JDOT (Courty et al., 2017). Note that we report only the variation of the method proposed in (Courty et al., 2015) with the group-Lasso regularization as this latter was showed to be the most efficient for this data set. Finally, as the results for JDOT on moons were not presented in the original paper, we run it with the hyper-parameter ranges suggested by the authors. The final results averaged over 10 tests on independent data sets of 1000 instances are presented in Table 1. From it, we can make several conclusions. First, all considered DA baselines manage to achieve an almost perfect score on the angles from $10°$ to $40°$, while SVM without adaptation has a 30% drop in accuracy for these angles. This shows that moons data set presents a challenging adaptation task that goes beyond the generalization capacities of a standard supervised learning algorithm. As for the DA baselines, their performance is rather not surprising as for these angles the adaptation problem remains fairly easy. Starting from $50°$ (Table 1(left)) and up to $90°$ (Table 1(left)), out method provides a better performance that those obtained with both JDOT and OT-GL with the most significant improvement obtained for the angle of $50°$. One should note that OT-GL method relies on the information about the source labels encoded in the group-lasso term but even this does not help to maintain its performance for larger angles. We conclude by saying that the theoretical features of the introduced OT-based distance used by our algorithm are highlighted by its efficiency in this experiment compared to strong OT baselines.

5.3. Sentiment analysis data set

Below, we consider the famous Amazon product reviews dataset (Blitzer et al., 2007) related to the sentiment analysis task. For this data set, we choose 4 of its subsets corresponding to different product categories, namely: books, dvd, electronics and kitchen (denoted by B, D, E, K, respectively). This leads to 12 domain adaptation tasks of varying difficulty as the proximity and the number of semantic relationship between the different domains vary a lot. As the original data is represented by over 100 000 features given by uni- and bigrams, we follow the pre-processing of (Chen et al., 2011) (resulting in between 20000 and 40000 features) and consider a linear kernel as a similarity function K. For each task, we use predefined sets of 2000 instances of source and target data samples for training, and keep 4000 instances of the target domain for testing. We compare our method to SVM with cross-validated hyper-parameters as a baseline, to a state-of-the-art adversarial DA approach DANN (Ganin et al., 2016) and to JDOT (NN) with a neural network used as a classifer, as done in (Courty et al., 2017). As our method uses only linear classifiers, we also run two baseline shallow algorithms for comparison, JDOT with a linear SVM and OT-GL (Courty et al., 2015) with 1-Nearest Neighbor classifier. The results of our experiments are reported in Table 2. From this table, we see that MADAOT outperforms other methods on 8 out of 12 tasks, and has the second best performance on 2 others. This is rather surprising considering that both DANN and JDOT (NN) rely on neural networks to learn the final classifier and this latter

is expected to have a higher discriminative power than the class of linear classifiers. Consequently, we attribute this performance gain to the efficiency of our task-dependent OT-based alignment term that manages to better align the two distributions compared to the minimization of the original 2-Wasserstein distance considered in OT-GL and JDOT. Furthermore, the minimax formulation of our alignment term addresses the curse of dimensionality problem related to OT as the sample complexity of this latter is known to scale exponentially in dimension. Several approaches were proposed to address this problem recently in order for the calculation of the Wasserstein distance to make sense for high-dimensional data and our findings show that applying it in the DA context can lead to an improved performance.

6. Conclusion and Future Perspectives

In this paper, we presented a novel theoretical analysis of unsupervised domain adaptation problem for binary classification that considers the margin violation loss on the target domain as the error measure. We proved a new bound on this latter that involves source margin violation error, a novel convex alignment term given by a task-dependent variant of the Wasserstein distance between the source and target domains and a non estimable term that offers new insights on domain adaptation problem and on the importance of the notion of margin violation for its a priori success. Our analysis generalizes several prior works on this subject and includes them as special cases. Our algorithm, derived from the established learning bounds, has proved to be efficient on both simulated and real-world problems compared to several state-of-the-art methods.

The future research directions of this paper are many. First, we would like to study the direct maximization of our non convex alignment term introduced in Theorem 1 using a deep adversarial approach. Even though our method offers a remarkable performance compared to several deep learning baselines, its efficiency can be further improved by this latter extension. Second, we plan to investigate in more detail the theoretical properties of our data dependent optimal transport term by establishing a concentration inequality for this latter. This would allow to theoretically highlight the success of our algorithm for high-dimensional data.

References

Arora, S., Babai, L., Stern, J., and Sweedyk, Z. The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, 54(2):317 – 331, 1997.

Balcan, M.-F., Blum, A., and Srebro, N. Improved guarantees for learning via similarity functions. *Computer Science Department*, pp. 126, 2008.

Ben-David, S. and Urner, R. On the Hardness of Domain Adaptation and the Utility of Unlabeled Target Samples. In *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pp. 139–153. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34106-9.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of Representations for Domain Adaptation. In *Advances in Neural Information Processing Systems 19*, pp. 137–144. MIT Press, 2007a.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 19*, pp. 137–144. MIT Press, 2007b.

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

Blankenship, J. W. and Falk, J. E. Infinitely constrained optimization problems. *Journal of Optimization Theory and Applications*, 19(2):261–281, 1976.

Blitzer, J., Dredze, M., and Pereira, F. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *In ACL*, pp. 187–205, 2007.

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks. *arXiv:1612.05424 [cs]*, 2016.

Bruzzone, L. and Marconcini, M. Domain Adaptation Problems: A DASVM Classification Technique and a Circular Validation Strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):770–787, 2010.

Chen, M., Weinberger, K. Q., and Blitzer, J. Co-Training for Domain Adaptation. In *Advances in Neural Information Processing Systems 24*, pp. 2456–2464. Curran Associates, Inc., 2011.

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

Cortes, C., Mohri, M., and Medina, A. M. Adaptation Based on Generalized Discrepancy. *Journal of Machine Learning Research*, 20(1):1–30, 2019.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal Transport for Domain Adaptation. *arXiv:1507.00504 [cs]*, 2015.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. Joint Distribution Optimal Transportation for Domain Adaptation. *arXiv:1705.08848 [cs, stat]*, 2017.

Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T. Subspace Alignment For Domain Adaptation. *arXiv:1409.5241 [cs]*, 2014.

Ganin, Y. and Lempitsky, V. Unsupervised Domain Adaptation by Backpropagation. *arXiv:1409.7495 [cs, stat]*, 2014.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. A New PAC-Bayesian Perspective on Domain Adaptation. In *International Conference on Machine Learning*, pp. 859–868, 2016.

Gong, B., Shi, Y., Sha, F., and Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073, 2012. doi: 10.1109/CVPR.2012.6247911.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

Grippo, L. and Sciandrone, M. On the convergence of the block nonlinear gauss–seidel method under convex constraints. *Operations Research Letters*, 26(3):127 – 136, 2000.

Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *Computer Vision – ECCV 2010*, volume 6314, pp. 213–226. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15560-4 978-3-642-15561-1.

Kar, P. and Jain, P. Similarity-based learning via data driven embeddings. In *Advances in Neural Information Processing Systems 24*, pp. 1998–2006. Curran Associates, Inc., 2011.

Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers, 2004.

Kouw, W. M. and Loog, M. A review of single-source unsupervised domain adaptation. *arXiv:1901.05335 [cs, stat]*, 2019.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain Adaptation: Learning Bounds and Algorithms. *arXiv:0902.3430 [cs]*, 2009.

Margolis, A. A literature review of domain adaptation with unlabeled data. *Rapport Technique, University of Washington*, 01 2011.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29 (2):429–443, 1997.

Pan, S. J. and Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Redko, I., Habrard, A., and Sebban, M. Theoretical Analysis of Domain Adaptation with Optimal Transport. *arXiv:1610.04420 [cs, stat]*, 2016.

Santambrogio, F. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2016. ISBN 9783319365817.

Sen, P. C., Hajra, M., and Ghosh, M. Supervised classification algorithms in machine learning: A survey and review. In *Emerging Technology in Modelling and Graphics*, pp. 99–111. Springer Singapore, 2020. ISBN 978-981-13-7403-6.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. Wasserstein Distance Guided Representation Learning for Domain Adaptation. *arXiv:1707.01217 [cs, stat]*, 2017.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate Shift Adaptation by Importance Weighted Cross Validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Sun, B., Feng, J., and Saenko, K. Return of Frustratingly Easy Domain Adaptation, 2016.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. Deep Domain Confusion: Maximizing for Domain Invariance. *arXiv:1412.3474 [cs]*, 2014.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

Wilson, G. and Cook, D. J. A Survey of Unsupervised Deep Domain Adaptation. *arXiv:1812.02849 [cs, stat]*, 2019.

Zhang, L. Transfer Adaptation Learning: A Decade Survey. *arXiv:1903.04687 [cs]*, 2019.

Zhang, Y., Liu, T., Long, M., and Jordan, M. Bridging Theory and Algorithm for Domain Adaptation. In *International Conference on Machine Learning*, pp. 7404–7413, 2019.

Zhong, E., Fan, W., Yang, Q., Verscheure, O., and Ren, J. Cross Validation Framework to Choose amongst Models and Datasets for Transfer Learning. In *Machine Learning and Knowledge Discovery in Databases*, volume 6323, pp. 547–562. Springer Berlin Heidelberg, 2010.