Analyse théorique de l'apprentissage avec des fonctions de similarités pour l'adaptation de domaine

Sofien Dhouib, Ievgen Redko

Revisiting (ϵ, γ, τ)-similarity learning for domain adaptation

Sofiane Dhouib[1] et Ievgen Redko[1]

[1]Univ.Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69266, LYON, France

Abstract

Similarity learning is an active research area in machine learning that tackles the problem of finding a similarity function tailored to an observable data sample in order to achieve efficient classification. This learning scenario has been generally formalized by the means of a $(\epsilon, \gamma, \tau)-$good similarity learning framework in the context of supervised classification and has been shown to have important theoretical guarantees. In this paper, we propose to extend the theoretical analysis of similarity learning to the domain adaptation setting, a particular situation occurring when the similarity is learned and then deployed on samples following different probability distributions. We give a new definition of an $(\epsilon, \gamma)-$good similarity for domain adaptation and prove several results quantifying the performance of a similarity function on a target domain after it has been trained on a source domain. We particularly show that if the source domain support contains that of the target then a notable improvement of the adaptation is achievable.

Keywords: metric learning, similarity learning, domain adaptation

1 Introduction

Many popular supervised learning algorithms rely on pairwise metrics calculated based on the instances of a given data set in order to learn a classifier. For instance, a famous family of k-nearest neighbors algorithms [CH06] uses distance matrices in order to define the label of a given test point while support vector machines [BGV92] can be extended to handle the non-linear classification using kernel functions. Despite a widespread use of metrics in machine learning, existing distances often do not capture the intrinsic geometry of data with respect to the labels of the available data points. To tackle this problem, the emerging field of metric learning (also known as similarity learning) aims to provide solutions that allow to learn pairwise metrics explicitly from the data, thus making them tailored for the classification or regression problem at hand. As an example, one may consider the first approach of this kind presented in [XNJR02] that consisted in learning a positive semi-definite (PSD) matrix defining a Mahalanobis distance, and then plugging this distance to a k-means clustering algorithm with side information on different pairs. We refer to [BHS13] and [Kul13] for recent surveys on metric learning.

From the theoretical point of view, similarity learning was extensively analyzed in two seminal papers of [BBS08b, BBS08a] based on the $(\epsilon, \gamma, \tau)-$good similarity framework for binary classification task. This framework formalizes an intuitive definition of a good similarity function: given a set of landmarks or reasonable points of probability mass at least τ, most of data points (a $1 - \epsilon$ probability mass) should be on average more similar to reasonable points of their own class than to points of the opposite class. Based on the proposed formalization, the authors provided performance guarantees for a resulting linear classifier after mapping data into a new feature space defined via the good similarity function. We refer the interested reader to [BHS12] and [GY14] for other theoretical studies on $(\epsilon, \gamma)-$ framework in the supervised, and to [NGHS15] and [ISH$^+$15]) in the semi-supervised learning cases.

While most of the work based on the (ϵ, γ, τ) framework has been done in the classical context where training and testing data have the same distribution, in several practical scenarios, one may want to transfer the learned similarity function from one domain, usually called source domain, to another, related yet different domain, called target domain. This framework, known as transfer learning, is a notorious research topic in machine learning nowadays [PY10, Mar11, PGLC15, WKW16] often used in situations where the target domain contains few or no labeled instances in order to reduce the time and effort needed for manual labeling

or even collecting new data. As many domain adaptation algorithms proposed in the literature are based on metric learning [GTX11, KSD11, CNS+11], a question about the theoretical guarantees of the general similarity framework naturally arises.

In this paper, we present a theoretical study of the $(\epsilon, \gamma, \tau)-$ framework in the domain adaptation context where only the marginal distributions across the source and the target domains are assumed to change while the class labels set remains the same. To the best of our knowledge, the only other related work on the domain adaptation problem for (ϵ, γ, τ)-good similarities was presented in [MHA12] in which the authors proposed a theoretical analysis of an algorithm that selects landmarks defining a projection space in which the source and target distribution are close. In this paper, we aim to consider a more general setting without being attached to a particular algorithm in order to investigate to which extent a similarity that is good for a source domain remains good for the target domain.

The rest of the paper is organized as follows. Section 2 presents the learning setting that we consider with some necessary definitions and notations. Section 3 introduces a generalization of the (ϵ, γ)-goodness definition used to provide a theoretical result bounding a divergence term between the source and target goodnesses. The bound contains a term reflecting the distance between the distributions of two domains and a worst margin term measuring the worst error obtainable by the similarity function for some instance from the learning sample. Depending on the assumptions made about the source and target domains distributions, we further provide two variations of the obtained bound with two types of probability metrics. We analyze the obtained worst margin term in Section 4 and measure the confidence of its empirical estimation. Finally, Section 5 is dedicated to the comparison of our results with other papers. We conclude our paper in Section 6 and give several possible future perspectives of this work.

2 Preliminaries

In order to proceed, we now introduce the basic elements related to the $(\epsilon, \gamma, \tau)-$good similarity framework. In what follows, we denote by $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \{-1, 1\}$ the features and labels spaces, respectively. For any real t, t_+ denotes its positive part, i.e $\max(t, 0)$. As in [BBS08a], we define a similarity function as a pairwise function $K : \mathcal{X} \times \mathcal{X} \to [-1, 1]$. We now recall the definition of the (ϵ, γ, τ)-goodness with hinge loss.

Definition 1 (Balcan et. al. 2008)**.** *A similarity function K is (ϵ, γ, τ)-good in hinge loss for problem (distribution) $\mathcal{P}$ if there exists a (probabilistic) indicator function R of a set of "reasonable points" such that:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{P}} \left[\left(1 - \frac{y.g(x)}{\gamma} \right)_+ \right] \leq \epsilon, \tag{1}$$

$$\mathop{\mathbb{P}}_{x'\sim\mathcal{P}} [R(x')] \geq \tau, \tag{2}$$

where $g(x) = \mathop{\mathbb{E}}_{(x',y')\sim P} [y' K(x, x') | R(x')]$.

In this definition, ϵ is an upper bound for the expected hinge loss over all the margins $g(x)$, every margin being the average signed similarity of an instance to reasonable points defined by R. In order to control the loss sensitivity to the margin, a division by γ is applied. We assume that $0 < \gamma \leq 1$.

Following this definition, the authors of [BBS08a] prove a theorem that guarantees the existence of a linear separator in a new feature space defined via an $(\epsilon, \gamma, \tau)-$good similarity function, a result that is stated by the following theorem.

Theorem 1 (Balcan et. al. 2008)**.** *Let K be an $(\epsilon, \gamma, \tau)-$good similarity function in hinge loss for a learning problem $\mathcal{P}$. For any $\epsilon_1 > 0$ and $0 < \delta < \frac{\gamma\epsilon_1}{4}$, let $S = \{x'_1, ..., x'_n\}$ be a (potentially unlabeled) sample of size*

$$n = \frac{2}{\tau} \log \left(\frac{2}{\delta} \right) \left(1 + \frac{16}{(\epsilon_1 \gamma)^2} \right)$$

of landmarks drawn from $\mathcal{P}$. Consider the mapping:

$$\begin{aligned} \phi^S : \mathcal{X} &\to \mathbb{R}^n \\ x &\mapsto (K(x, x'_1), ..., K(x, x'_n)). \end{aligned}$$

Then with a probability at least $1 - \delta$ over the draw of S, there exists $\beta \in \mathbb{R}^n$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{P}} \left[\left(1 - \frac{\langle \beta, \phi^S(x) \rangle}{\gamma} \right)_+ \right] \leq \epsilon + \epsilon_1. \tag{3}$$

In other words, the induced distribution $\phi^S(P)$ in $\mathbb{R}^n$ has a linear separator achieving hinge loss at most $\epsilon + \epsilon_1$ at margin γ.

One can see this theorem as a variation of the kernel trick used in the SVM algorithm. Indeed, if K is a kernel function and if $\tau = 1$, the expected loss in Equation (3) becomes the non-regularized loss of an SVM defined via kernel K. The authors furthermore derive an algorithm from this theorem that minimizes the empirical version of (3), which boils down to a linear programming problem that is solved efficiently.

3 $(\epsilon, \gamma)-$good similarity learning for domain adaptation

In this section, we introduce the main contributions of our paper. We start by giving a definition of (ϵ, γ)-goodness with an arbitrary distribution of landmarks, and then propose a generalization bound that relates the goodness of the same similarity function learned on the source and target domains.

3.1 Problem setup

As mentioned earlier, the main goal of this paper is to address the domain adaptation problem in the context of $(\epsilon, \gamma)-$ similarity learning. For this case, we assume to have access to samples S and T drawn from source and target probability distributions $\mathcal{S}$ and $\mathcal{T}$, respectively. In the context of domain adaptation, $S \subset (\mathcal{X} \times \mathcal{Y})^m$ is labeled whereas T can be partially or totally unlabeled. In the rest of the paper, we suppose that the labeling is deterministic, meaning that there exists a labeling function $f_{\mathcal{S}}$ (resp. $f_{\mathcal{T}}$) such that for every (x, y) in the source domain (resp. in the target domain), $y = f_{\mathcal{S}}(x)$ (resp. $y = f_{\mathcal{T}}(x)$). Hence, we replace every $(x, y) \sim \mathcal{P}$ by writing simply $x \sim \mathcal{P}$ for all probability distributions considered below.

As hinted in [BBS08a, Note 2, Theorem 14], the instances and landmarks can be potentially drawn from different distributions. Hence, we propose a slight modification of Definition 1 that we use from now on.

Definition 2. *A similarity function K is (ϵ, γ)-good in hinge loss for problem $(\mathcal{P}, \mathcal{R})$ (where $\mathcal{P}$ is the data distribution whereas $\mathcal{R}$ is the landmarks distribution) if:*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{P}}\left[\left(1 - \frac{y.g_{\mathcal{R}}(x)}{\gamma}\right)_+\right] \leq \epsilon,$$

where $g_{\mathcal{R}}(x) = \mathop{\mathbb{E}}_{x'\sim\mathcal{R}}[y'K(x, x')]$.

This is a generalization of Definition 1, and the two coincide when we consider the distribution $\mathcal{R}$ defined by $\mathop{\mathbb{P}}_{x\sim\mathcal{R}}[x \in A] = \mathop{\mathbb{P}}_{x\sim\mathcal{P}}[x \in A | R(x) = 1]$ for all measurable sets A. As for parameter τ, we do not explicitly mention it in the definition, but it is an upper bound for $\mathop{\mathbb{P}}_{x\sim\mathcal{P}}[x \in \operatorname{supp}\mathcal{R}]$ since in this case, we have $\operatorname{supp}\mathcal{R} \subset \{R(x) = 1\}$.

In the rest of the paper, we use the following notations for any data distribution $\mathcal{P}$ and landmark distribution $\mathcal{Q}$. We denote the goodness of K for problem $(\mathcal{P}, \mathcal{Q})$ by

$$\mathcal{E}_{\mathcal{P},\mathcal{Q}}(K) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{P}}\left[\left(1 - \frac{y.g_{\mathcal{Q}}(x)}{\gamma}\right)_+\right].$$

For simplicity, we further denote by l_γ the γ-scaled hinge loss function defined by:

$$l_\gamma : x \mapsto \left(1 - \frac{x}{\gamma}\right)_+.$$

We let μ be a dominating probability distribution, i.e $\operatorname{supp}\mu$ contains the support of all other probability measures used afterwards. In addition, $\mathcal{M}_{\mathcal{P},\mathcal{Q}}(K)$ stands for the worst margin achieved by an element of $x \in \operatorname{supp}\mathcal{P}$ associated with landmark distribution $\mathcal{Q}$, i.e:

$$\mathcal{M}_{\mu,\mathcal{Q}}(K) := \sup_{x\in\operatorname{supp}\mathcal{P}} l_\gamma(yg_{\mathcal{Q}}(x)).$$

Note that since K has values in $[-1, 1]$, $yg_{\mathcal{Q}}(x)$ is also bounded in the same interval and consequently $l_\gamma(yg_{\mathcal{Q}}(x))$ is bounded thanks to its continuity. This ensures that $\mathcal{M}_{\mathcal{P},\mathcal{Q}}(K)$ is finite. Finally, if B is a boolean expression, then $[B] := \mathbb{1}_B$ is an indicator of the set on which B holds (Iverson bracket notation).

3.2 Relating the source and target goodnesses

Given a similarity function that is (ϵ, γ)-good in hinge loss for problem $(\mathcal{S}, \mathcal{R}_1)$, our goal is to bound its goodness on the target set for problem $(\mathcal{T}, \mathcal{R}_2)$, where $\mathcal{R}_1$ and $\mathcal{R}_2$ are not supposed to be equal. Based on the last assumption, we further consider two cases: for the first case, we assume that the landmark distribution $\mathcal{R}_1 = \mathcal{R}_2 = \mathcal{R}$ is common for both domains; for the second, we derive bounds for the general case of different landmark distributions using the results obtained for the first case.

3.2.1 Shared landmarks distribution

The following lemma allows to bound the difference between the goodness of a given similarity function w.r.t. the source and target domains when the landmark distribution does not change across two domains.

Lemma 1. *Let K be an (ϵ, γ)-good similarity for problem $(\mathcal{S}, \mathcal{R})$. Then K is $(\epsilon + \epsilon', \gamma)$-good for problem $(\mathcal{T}, \mathcal{R})$, with:*

$$\epsilon' = \mathop{\mathbb{E}}_{x\sim\mu}\left[\left(\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathcal{S}}{\mathrm{d}\mu}\right)_+ l_\gamma(yg_{\mathcal{R}}(x))[yg_{\mathcal{R}}(x) < \gamma]\right].$$

Proof. We have

$$\begin{aligned}\mathcal{E}_{\mathcal{T},\mathcal{R}}(K) &= \mathcal{E}_{\mathcal{S},\mathcal{R}}(K) + \mathcal{E}_{\mathcal{T},\mathcal{R}}(K) - \mathcal{E}_{\mathcal{S},\mathcal{R}}(K) \\ &\leq \epsilon + \mathcal{E}_{\mathcal{T},\mathcal{R}}(K) - \mathcal{E}_{\mathcal{S},\mathcal{R}}(K) \end{aligned} \tag{4}$$

following from the $(\epsilon,\gamma)-$goodness of K for $(\mathcal{P},\mathcal{R})$. Now we focus on the difference between the two last terms in (4). We get the following:

$$\begin{aligned}&\mathcal{E}_{\mathcal{T},\mathcal{R}}(K) - \mathcal{E}_{\mathcal{S},\mathcal{R}}(K) \qquad (5)\\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{S}}[l_\gamma(y.g_\mathcal{R}(x))] - \mathop{\mathbb{E}}_{x\sim\mathcal{T}}[l_\gamma(y.g_\mathcal{R}(x))] \\ &= \mathop{\mathbb{E}}_{x\sim\mu}\left[\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mu} l_\gamma(y.g_\mathcal{R}(x))\right] - \mathop{\mathbb{E}}_{x\sim\mu}\left[\frac{\mathrm{d}\mathcal{S}}{\mathrm{d}\mu} l_\gamma(y.g_\mathcal{R}(x))\right] \\ &\leq \mathop{\mathbb{E}}_{x\sim\mu}\left[\left(\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathcal{S}}{\mathrm{d}\mu}\right)_+ l_\gamma(yg_\mathcal{R}(x))[yg_\mathcal{R}(x) < \gamma]\right], \qquad (6)\end{aligned}$$

where (6) is obtained by noticing that $t \leq t_+ \ \forall t \in \mathbb{R}$, and due to the positivity of l_γ and to its nullity when calculated at a point $t \geq \gamma$. □

In general, we note that the difference in (5) can be bounded by an integral probability metric ([Zol84, MÿŸ7, ZZY13]) by taking the supremum over all similarity functions K belonging to a certain hypothesis space $\mathfrak{K}$. Given a fixed landmark distribution $\mathcal{R}$, $\mathfrak{K}$ induces a space of hypotheses

$$\mathfrak{G}_\mathcal{R} = \{x \mapsto g_\mathcal{R}(x) = \mathop{\mathbb{E}}_{x'\sim\mathcal{R}}[y'K(x,x')]\,; K \in \mathfrak{K}\}$$

taking one argument (similar to the traditional supervised learning framework). The integral probability metric is then given by

$$\begin{aligned}d_{\mathfrak{K}}(\mathcal{S},\mathcal{T}) &= \sup_{K\in\mathfrak{K}} |\mathcal{E}_{\mathcal{T},\mathcal{R}}(K) - \mathcal{E}_{\mathcal{S},\mathcal{R}}(K)| \\ &= \sup_{g_\mathcal{R}\in\mathfrak{G}_\mathcal{R}} \left| \mathop{\mathbb{E}}_{x\sim\mathcal{S}}[l_\gamma(y.g_\mathcal{R}(x))] - \mathop{\mathbb{E}}_{x\sim\mathcal{T}}[l_\gamma(y.g_\mathcal{R}(x))]\right|.\end{aligned}$$

In Lemma 1, we chose to bound this difference in another manner by providing a first upper bound given in (6). In this bound, the expectation is taken only on the support of the hinge loss, i.e for instances having a signed margin smaller than γ, making it problem dependent. Its aim is to prepare for proving two upper bounds for ϵ', in terms of the L^1 distance and then χ^2 divergence in order to quantify the behavior of the target error as a function of the divergence between the two domains.

Lemma 2 (L^1 bound, shared landmarks)**.** *Let K be an (ϵ,γ)-good similarity for problem $(\mathcal{S},\mathcal{R})$. Then K is $(\epsilon+\epsilon',\gamma)$-good for problem $(\mathcal{T},\mathcal{R})$, where:*

$$\epsilon' = d_{1+,\gamma}(\mathcal{T},\mathcal{S})\mathcal{M}_{\mu,\mathcal{R}}(K)$$

with

$$d_{1+,\gamma}(\mathcal{T},\mathcal{S}) = \mathop{\mathbb{E}}_{x\sim\mu}\left[\left(\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathcal{S}}{\mathrm{d}\mu}\right)_+ [yg_\mathcal{R}(x) < \gamma]\right].$$

Proof.

$$\begin{aligned}&\mathop{\mathbb{E}}_{x\sim\mu}\left[\left(\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathcal{S}}{\mathrm{d}\mu}\right)_+ l_\gamma(yg_\mathcal{R}(x))[yg_\mathcal{R}(x) < \gamma]\right] \\ &\leq \mathop{\mathbb{E}}_{x\sim\mu}\left[\left(\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathcal{S}}{\mathrm{d}\mu}\right)_+ [yg_\mathcal{R}(x) < \gamma]\right]\mathcal{M}_{\mu,\mathcal{R}}(K) \qquad (7)\\ &= d_{1+,\gamma}(\mathcal{T},\mathcal{S})\mathcal{M}_{\mu,\mathcal{R}}(K)\end{aligned}$$

where we use Hölder's inequality with ℓ_1 and ℓ_∞ norms to obtain (7). □

We note the presence of the term $\mathcal{M}_{\mu,\mathcal{R}}(K)$ here which stands for the worst margin achieved by K on some instance of $\operatorname{supp}\mu$. In the case of the SVM, this term is analogous to the largest slack variable associated to an instance drawn from the dominating measure μ. However, depending on the choice of μ, it can be difficult to control, as we can estimate it only by observing data drawn from $\mathcal{S}$.

In order to tackle this limitation and to obtain a tighter bound, we further assume that $\mathcal{S}$ dominates $\mathcal{T}$ implying $\operatorname{supp}\mathcal{T} \subset \operatorname{supp}\mathcal{S}$. For this particular case, the following lemma can be proved.

Lemma 3 (χ^2-bound, same landmarks)**.** *Let K be an (ϵ,γ)-good similarity for problem $(\mathcal{S},\mathcal{R})$. Assume that $\operatorname{supp}\mathcal{T} \subset \operatorname{supp}\mathcal{S}$. Then K is $(\epsilon+\epsilon',\gamma)$-good for problem $(\mathcal{T},\mathcal{R})$, where:*

$$\epsilon' = \sqrt{d_{\chi^2_+,\gamma}(\mathcal{T},\mathcal{S})\mathcal{M}_{\mathcal{S},\mathcal{R}}(K)}\sqrt{\epsilon}$$

with

$$d_{\chi^2_+,\gamma}(\mathcal{T},\mathcal{S}) = \mathop{\mathbb{E}}_{x\sim\mathcal{S}}\left[\left(\left(\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mathcal{S}} - 1\right)_+\right)^2 [yg_\mathcal{R}(x) < \gamma]\right].$$

Proof. We bound the same quantity as in the proof of lemma 2. Since $\mathcal{S}$ dominates $\mathcal{T}$, we take $\mu = \mathcal{S}$ and we have:

$$\begin{aligned}&\mathop{\mathbb{E}}_{x\sim\mu}\left[\left(\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathcal{S}}{\mathrm{d}\mu}\right)_+ l_\gamma(yg_\mathcal{R}(x))[yg_\mathcal{R}(x) < \gamma]\right]^2 \\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{S}}\left[\left(\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mathcal{S}} - 1\right)_+ l_\gamma(yg_\mathcal{R}(x))[yg_\mathcal{R}(x) < \gamma]\right]^2 \\ &\leq \mathop{\mathbb{E}}_{x\sim\mathcal{S}}\left[\left(\left(\frac{\mathrm{d}\mathcal{T}}{\mathrm{d}\mathcal{S}} - 1\right)_+\right)^2 [yg_\mathcal{R}(x) < \gamma]\right] \mathop{\mathbb{E}}_{x\sim\mathcal{S}}\left[l_\gamma(yg_\mathcal{R}(x))^2\right] \qquad (8)\end{aligned}$$

$$
\begin{aligned}
&= d_{\chi^2_+,\gamma}(\mathcal{T},\mathcal{S}) \mathop{\mathbb{E}}_{x\sim\mathcal{S}} \left[l_\gamma(y g_{\mathcal{R}}(x))^2 \right] \\
&\leq d_{\chi^2_+,\gamma}(\mathcal{T},\mathcal{S}) \mathcal{M}_{\mathcal{S},\mathcal{R}}(K) \mathop{\mathbb{E}}_{x\sim\mathcal{S}} [l_\gamma(y g_{\mathcal{R}}(x))] \qquad (9) \\
&\leq d_{\chi^2_+,\gamma}(\mathcal{T},\mathcal{S}) \mathcal{M}_{\mathcal{S},\mathcal{R}}(K) \epsilon.
\end{aligned}
$$

To obtain (8), we applied the Cauchy-Schwartz inequality. Inequality 9 is obtained thanks to the boundedness and positivity of l_γ via a Hölder inequality for norms ℓ_1 and ℓ_∞. The last line follows from the $(\epsilon,\gamma)-$goodness of K for problem $(\mathcal{S},\mathcal{R})$. □

This last result clearly shows the benefit of assuming $\operatorname{supp}\mathcal{T} \subset \operatorname{supp}\mathcal{S}$: the distance term in the bound is multiplied by $\sqrt{\epsilon}$ meaning that having a similarity function achieving a low error on the source domain can leverage the difference between the domains' distributions. Note that this assumption is quite common in the domain adaptation literature and has already been used in [ZSMW13]. As mentioned in this latter paper, it roughly means that the source domain is richer than the target one, an assumption that is quite reasonable to make in practice.

3.2.2 Different landmarks case

We now turn our attention to a more general case where the landmarks distributions vary across two domains. To this end, we assume that a similarity function K is (ϵ,γ)-good for $(\mathcal{S},\mathcal{R}_1)$. Given these assumptions, our goal now is to provide a learning guaranty for the goodness of K for the $(\mathcal{T},\mathcal{R}_2)$ learning problem using the results established for the previous case.

To proceed, we first note that the difference between $\mathcal{E}_{\mathcal{T},\mathcal{R}_2}(K)$ and $\mathcal{E}_{\mathcal{S},\mathcal{R}_1}(K)$ can be equivalently written as:

$$
\begin{aligned}
&\mathcal{E}_{\mathcal{T},\mathcal{R}_2}(K) - \mathcal{E}_{\mathcal{S},\mathcal{R}_1}(K) \\
&= \mathcal{E}_{\mathcal{T},\mathcal{R}_1}(K) - \mathcal{E}_{\mathcal{S},\mathcal{R}_1}(K) + \mathcal{E}_{\mathcal{T},\mathcal{R}_2}(K) - \mathcal{E}_{\mathcal{T},\mathcal{R}_1}(K).
\end{aligned}
$$

By analyzing the obtained expression, we note that the difference between the first two terms can be bounded directly using Lemma 2 or 3 after taking into account the necessary hypotheses. Consequently, in what follows we focus solely on the last two terms and, similar to the previous case, provide a result based on both the L_1 and χ^2 distances.

For this case, we obtain the following proposition.

Proposition 1. *Let K be an (ϵ,γ)-good similarity for problem $(\mathcal{S},\mathcal{R}_1)$. Then K is $(\epsilon+\epsilon'+\epsilon'',\gamma)$-good for problem $(\mathcal{T},\mathcal{R}_2)$, with:*

$$
\epsilon'' = \frac{1}{\gamma} d_1(\mathcal{R}_1,\mathcal{R}_2)
$$

and

$$
\epsilon' = d_{1+,\gamma}(\mathcal{T},\mathcal{S}) \mathcal{M}_{\mu,\mathcal{R}_1}(K),
$$

where $d_1(\mathcal{R}_1,\mathcal{R}_2) = \mathop{\mathbb{E}}_{x'\sim\mu} \left[\left| \frac{\mathrm{d}\mathcal{R}_1}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathcal{R}_2}{\mathrm{d}\mu} \right| \right]$. Moreover, if $\operatorname{supp}\mathcal{T} \subset \operatorname{supp}\mathcal{S}$, then the obtained result holds with

$$
\epsilon' = \sqrt{d_{\chi^2_+,\gamma}(\mathcal{T},\mathcal{S}) \mathcal{M}_{\mathcal{S},\mathcal{R}_1}(K)} \sqrt{\epsilon}.
$$

Proof.

$$
\begin{aligned}
&\mathcal{E}_{\mathcal{T},\mathcal{R}_2}(K) - \mathcal{E}_{\mathcal{T},\mathcal{R}_1}(K) \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{T}} [l_\gamma(y g_{\mathcal{R}_2}(x)) - l_\gamma(y g_{\mathcal{R}_1}(x))] \\
&\leq \frac{1}{\gamma} \mathop{\mathbb{E}}_{x\sim\mathcal{T}} [| y g_{\mathcal{R}_1}(x) - y g_{\mathcal{R}_2}(x) |] \qquad (10) \\
&= \frac{1}{\gamma} \mathop{\mathbb{E}}_{x\sim\mathcal{T}} \left[\left| \mathop{\mathbb{E}}_{x'\sim\mu} \left[\left(\frac{\mathrm{d}\mathcal{R}_1}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathcal{R}_2}{\mathrm{d}\mu} \right) y y' K(x,x') \right] \right| \right] \\
&\leq \frac{1}{\gamma} \mathop{\mathbb{E}}_{x\sim\mathcal{T}} \left[\mathop{\mathbb{E}}_{x'\sim\mu} \left[\left| \left(\frac{\mathrm{d}\mathcal{R}_1}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathcal{R}_2}{\mathrm{d}\mu} \right) y y' K(x,x') \right| \right] \right] \qquad (11) \\
&\leq \frac{1}{\gamma} \mathop{\mathbb{E}}_{x'\sim\mu} \left[\left| \frac{\mathrm{d}\mathcal{R}_1}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathcal{R}_2}{\mathrm{d}\mu} \right| \right]. \qquad (12)
\end{aligned}
$$

where (10) holds because l_γ is $\frac{1}{\gamma}-$lipschitz. (11) is obtained by the Jensen inequality with the convexity of the $|.|$ function. Line (12) comes from the fact that $y y' K(x,x') \leq 1$. As for ϵ', it is directly obtained by Lemma 2 or 3 depending on the assumption made about the support of the target distribution. □

This result shows the effect of different landmark distributions on the adaptation capacity of a given similarity function. It proves to which extent different landmark distributions can be penalizing as the L^1 distance cannot be estimated from finite samples ([KBDG04, BFR$^+$00]) making the bound potentially vacuous for an arbitrary pair of distributions $\mathcal{R}_1$ and $\mathcal{R}_2$. For this reason, we focus on the case of a shared landmark distribution in the rest of the paper.

4 Analysis of the worst margin term

As the worst margin term $\mathcal{M}_{\mu,\mathcal{R}}(K)$ is present in both bounds obtained in the previous section (Lemmas 2 and 3), we now proceed to its analysis below. It tells us that if there is at least one instance from the source distribution (or from a distribution dominating it) that has a high loss, then the deviation between the target error and the source error is expected to be large. In what follows, we provide an analysis of this term showing first that it can be bounded in terms of γ and then presenting a guarantee for its deviation from its empirical counterpart.

4.1 A simple bound for the worst margin

The first bound for the worst margin term can be obtained as follows:

$$\begin{aligned}\mathcal{M}_{\mu,\mathcal{R}}(K) &= \sup_{x\in\operatorname{supp}\mu} l_\gamma(yg_{\mathcal{R}}(x)) \\ &= \left(1-\frac{1}{\gamma}\inf_{x\in\operatorname{supp}\mu} y.g_{\mathcal{R}}(x)\right)_+ \\ &= \left(1-\frac{1}{\gamma}\inf_{x\in\operatorname{supp}\mu}\mathbb{E}_{x'\sim\mathcal{R}}\left[yy'K(x,x')\right]\right)_+ \\ &\leq 1+\frac{1}{\gamma}\leq\frac{2}{\gamma}.\end{aligned}$$

The last inequality comes from the fact that $K:\mathcal{X}\times\mathcal{X}\to[-1,1]$ and that $0<\gamma\leq 1$.

Based on the obtained expression, we note that the bounds given in Lemma 2 and Lemma 3 become proportional to $\frac{1}{\gamma}$ and $\sqrt{\frac{\epsilon}{\gamma}}$, respectively. The second bound especially suggests that if K is good on the first domain (small ϵ), and if the source support contains the target's, then K performs moderately on the target domain, as the $\sqrt{\epsilon}$ term reduces the effect of the divergence between the two domains.

The worst margin term multiplies a divergence term between $\mathcal{S}$ and $\mathcal{T}$ in both Lemmas 2 and 3. If it has a high value then focusing on minimizing the divergence between the two domains becomes crucial for the potential success of adaptation. Thus, it can be useful to estimate this term empirically from the observed data sample by taking the empirical maximum for the source instances and the empirical mean for the landmarks. For the sake of simplicity, we only consider the case where $\mathcal{S}$ dominates $\mathcal{T}$.

4.2 An empirical estimation of the worst margin

We intend to measure our confidence in the empirical estimation of the worst margin term by bounding the deviation between the real worst margin term and its empirical counterpart. To this end, we suppose having access to a labeled data sample $S=\{(x_1,y_1),...,(x_m,y_m)\}\subset(\mathcal{X}\times\mathcal{Y})^m$ drawn from $\mathcal{S}$, inducing an empirical distribution $\hat{\mathcal{S}}$. Similarly, we define a sample $S_{\mathcal{R}}=\{(x'_1,y'_1),\ldots,(x'_r,y'_r)\}$ and the corresponding empirical distribution $\hat{\mathcal{R}}$. As the main result of this section relies on the notion of the Rademacher complexity, we give its definition below.

Definition 3. *Let $\mathcal{G}$ be a family of mappings from $\mathcal{X}$ to $\mathbb{R}$ and $\mathcal{P}$ be a probability distribution on $\mathcal{X}$. The Rademacher complexity of $\mathcal{G}$ w.r.t. $\mathcal{P}$ and to a sample size n is defined as*

$$\operatorname{Rad}_n(\mathcal{G})=\mathop{\mathbb{E}}_{S\sim\mathcal{P}^n}\left[\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n\sigma_i g(s_i)\right]\right]$$

where σ_i are independent uniform random variables in $\{-1,+1\}$ called Rademacher random variables and $S=\{s_1,...,s_n\}$.

Intuitively, the Rademacher complexity is large if we can find a function $g\in\mathcal{G}$ that looks like random noise, i.e. highly correlated with the Rademacher random variables.

Under these notations, the following result can be proved.

Proposition 2. *Let K be a similarity function defined on a feature space $\mathcal{X}$. Let $\mathcal{M}_{\mathcal{S},\mathcal{R}}(K)$ denote its worst performance associated to loss function l_γ and achieved by an example drawn from $\mathcal{S}$, where $\mathcal{R}$ is the landmarks distribution. Assume the cumulative distribution function F_{l_γ} of the loss function associated with $\mathcal{S}$ and $\hat{\mathcal{R}}$ is k times differentiable at $\mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K)$, and that $k>0$ is the minimum integer such that $F^{(k)}_{l_\gamma}\neq 0$. Then for all $\alpha>1, r\geq 1$, there exists $m_0\geq 1$ such that for all $m\geq m_0$, we have with probability at least $1-\delta$:*

$$\begin{aligned}\mathcal{M}_{\mathcal{S},\mathcal{R}}(K)\leq\mathcal{M}_{\hat{\mathcal{S}},\hat{\mathcal{R}}}(K)&+\frac{2}{\gamma}\operatorname{Rad}_r\left(\mathfrak{H}_1(K)\right)\\&+\frac{1}{\gamma^2}\sqrt{2\frac{\log\left(\frac{2}{\delta}\right)}{r}}+\left(\frac{(-1)^{k+1}\log\left(\frac{2\alpha}{\delta}\right)k!}{F^{(k)}_{l_\gamma}(\mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K))m}\right)^{\frac{1}{k}},\end{aligned}$$

where $\mathfrak{H}_1(K)$ is the hypothesis class defined by

$$\mathfrak{H}_1(K)=\{x'\mapsto K(x,x'), x\in\operatorname{supp}\mathcal{S}\}.$$

Proof. To proceed, we first rewrite the quantity of interest as

$$\mathcal{M}_{\mathcal{S},\mathcal{R}}(K)=\mathcal{M}_{\mathcal{S},\mathcal{R}}(K)-\mathcal{M}_{\hat{\mathcal{S}},\hat{\mathcal{R}}}(K)+\mathcal{M}_{\hat{\mathcal{S}},\hat{\mathcal{R}}}(K)$$

and further focus on bounding the difference between the first two terms which can be separated into two quantities as follows:

$$\begin{aligned}M_1&=\mathcal{M}_{\mathcal{S},\mathcal{R}}(K)-\mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K),\\M_2&=\mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K)-\mathcal{M}_{\hat{\mathcal{S}},\hat{\mathcal{R}}}(K).\end{aligned}$$

We begin by bounding M_1:

$$M_1=\sup_{x\in\operatorname{supp}\mathcal{S}} l_\gamma(yg_{\mathcal{R}}(x))-\sup_{x\in\operatorname{supp}\mathcal{S}} l_\gamma(yg_{\hat{\mathcal{R}}}(x)) \tag{13}$$

$$\leq \sup_{x\in\operatorname{supp}\mathcal{S}} \{l_\gamma(yg_{\mathcal{R}}(x)) - l_\gamma(yg_{\hat{\mathcal{R}}}(x))\} \tag{14}$$

$$\leq \frac{1}{\gamma} \sup_{x\in\operatorname{supp}\mathcal{S}} |g_{\mathcal{R}}(x) - g_{\hat{\mathcal{R}}}(x)| \tag{15}$$

$$= \frac{1}{\gamma} \sup_{x\in\operatorname{supp}\mathcal{S}} \left| \mathop{\mathbb{E}}_{x'\sim\mathcal{R}} \left[y'K(x,x')\right] - \frac{1}{r}\sum_{i=1}^{r} y'_i K(x,x'_i) \right|. \tag{16}$$

where (15) holds by the $\frac{1}{\gamma}$-lipschitzness of l_γ. The quantity in (16) is known as the representativeness (see, for example, [SSBD14]) of sample $S_{\mathcal{R}}$ drawn from $\mathcal{R}$ associated with the hypothesis set $\mathcal{Y}.\mathfrak{H}_1(K)$. In what follows, we denote it by $\operatorname{Rep}_{\mathcal{R}}(\mathcal{Y}.\mathfrak{H}_1(K), S_{\mathcal{R}})$ and notice that its value changes at most by $\dfrac{2}{\gamma r}$ if an instance of $S_{\mathcal{R}}$ is replaced since K takes values in $[-1,1]$. By applying Mc-Diarmid's inequality, we have with a probability at least $1-\frac{\delta}{2}$ for $0<\delta\leq 1$

$$\operatorname{Rep}_{\mathcal{R}}(\mathcal{Y}.\mathfrak{H}_1(K), S_{\mathcal{R}}) \leq \mathop{\mathbb{E}}_{S_{\mathcal{R}}\sim\mathcal{R}^m} \left[\operatorname{Rep}_{\mathcal{R}}(\mathcal{Y}.\mathfrak{H}_1(K), S_{\mathcal{R}})\right] + \frac{1}{\gamma}\sqrt{2\frac{\log\left(\frac{2}{\delta}\right)}{r}}. \tag{17}$$

The expectation term in (17) can be bounded by twice the Rademacher complexity of hypotheses class $\mathcal{Y}.\mathfrak{H}_1(K)$ (see, for example, [SSBD14, Lemma 26.2]), denoted by $\operatorname{Rad}_r(\mathcal{Y}.\mathfrak{H}_1(K))$, which also equals $\operatorname{Rad}_r(\mathfrak{H}_1(K))$. Hence, with a probability at least $1-\frac{\delta}{2}$, we have:

$$M_1 \leq \frac{2}{\gamma}\operatorname{Rad}_r(\mathfrak{H}_1(K)) + \frac{1}{\gamma^2}\sqrt{2\frac{\log\left(\frac{2}{\delta}\right)}{r}}. \tag{18}$$

Now, we focus on M_2 and examine the probability over the draw of S that it exceeds a certain threshold. For a given $t>0$, we have:

$$\begin{aligned}
&\mathop{\mathbb{P}}_{S\sim\mathcal{S}^m}[M_2 \geq t] \\
&= \mathop{\mathbb{P}}_{S\sim\mathcal{S}^m}\left[\mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K) - \mathcal{M}_{\hat{\mathcal{S}},\hat{\mathcal{R}}}(K) \geq t\right] \\
&= \mathop{\mathbb{P}}_{S\sim\mathcal{S}^m}\left[\mathcal{M}_{\hat{\mathcal{S}},\hat{\mathcal{R}}}(K) \leq \mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K) - t\right] \\
&= \mathop{\mathbb{P}}_{S\sim\mathcal{S}^m}\left[\max_{1\leq i\leq m} l_\gamma(y_i g_{\hat{\mathcal{R}}}(x_i)) \leq \mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K) - t\right] \\
&= \mathop{\mathbb{P}}_{x\sim\mathcal{S}}\left[l_\gamma(yg_{\hat{\mathcal{R}}}(x)) \leq \mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K) - t\right]^m \\
&= F_{l_\gamma}\left(\mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K) - t\right)^m.
\end{aligned}$$

By the assumptions made on the regularity of F_{l_γ}, setting t to $\dfrac{t}{m^{\frac{1}{k}}}$ yields:

$$\begin{aligned}
&\mathop{\mathbb{P}}_{S\sim\mathcal{S}^m}\left[M_2 \geq \frac{t}{m^{\frac{1}{k}}}\right] \\
&= \left(1 + F^{(k)}_{l_\gamma}(\mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K))\frac{(-t)^k}{mk!} + o\left(\frac{t^k}{m}\right)\right)^m \qquad (19)\\
&\underset{m\to\infty}{\longrightarrow} \exp\left(F^{(k)}_{l_\gamma}(\mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K))\frac{(-t)^k}{k!}\right). \qquad (20)
\end{aligned}$$

where (19) is obtained from a Taylor expansion. This implies for any $\alpha>1$ that there exists $m_0\in\mathbb{N}^*$ such that for all $m\geq m_0$,

$$\mathop{\mathbb{P}}_{S\sim\mathcal{S}^m}\left[M_2 \geq \frac{t}{m^{\frac{1}{k}}}\right] \leq \alpha\exp\left(F^{(k)}_{l_\gamma}(\mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K))\frac{(-t)^k}{k!}\right).$$

Setting this bound to $\frac{\delta}{2}$ and solving for t yields that with a probability at least $1-\frac{1}{\delta}$

$$M_2 \leq \left(\frac{(-1)^{k+1}\log\left(\frac{2\alpha}{\delta}\right)k!}{F^{(k)}_{l_\gamma}(\mathcal{M}_{\mathcal{S},\hat{\mathcal{R}}}(K))m}\right)^{\frac{1}{k}}. \tag{21}$$

Finally we use a union bound to bound the probability that the two inequalities (18) and (21) occur simultaneously in order to obtain the theorem's result. □

This theorem shows that under certain conditions, the empirical maximum is guaranteed to converge in probability to the real supremum of the distribution's support. The convergence rate depends heavily on the complexity of the similarity function search space represented by the Rademacher complexity term and on the regularity of the loss distribution function reflected by the $m^{-\frac{1}{k}}$ term. This last term dominates the convergence rate when $k>2$, and we have in general a convergence rate that is $\mathcal{O}(m^{-\frac{1}{\max\{2,k\}}})$.

5 Discussion

In this section, we briefly compare the obtained results with some previous works that prove generalization bounds for domain adaptation. We choose these particular works based on the similarity of their results with ours in order to highlight the main differences between them. In [BDBC$^+$10], the authors provide first learning guarantees for the general adaptation problem in the following form:

$$\epsilon_{\mathcal{T}}(h, f_{\mathcal{T}}) \leq \epsilon_{\mathcal{S}}(h, f_{\mathcal{S}}) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S},\mathcal{T}) + \lambda, \tag{22}$$

where $\epsilon_{\mathcal{D}}(h, f_{\mathcal{D}}) := \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[|h(x) - f_{\mathcal{D}}(x)|]$ is the error function defined over some probability distribution $\mathcal{D}$ for a hypothesis and a labeling function $h, f_{\mathcal{D}} : \mathcal{X} \to \mathcal{Y}$ with zero-one loss and λ is the combined error of the ideal hypothesis h^* that minimizes $\epsilon_{\mathcal{S}}(h) + \epsilon_{\mathcal{T}}(h)$. In the proposed framework, the main quantity of interest is the introduced $\mathcal{H}\Delta\mathcal{H}$ divergence defined as follows:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) = 2 \sup_{h,h'\in\mathcal{H}} \Big| \mathop{\mathbb{P}}_{x\sim\mathcal{S}}[h(x) \neq h'(x)] - \mathop{\mathbb{P}}_{x\sim\mathcal{T}}[h(x) \neq h'(x)] \Big|.$$

This divergence measure is a slight modification of the $\mathcal{A}$-divergence introduced in [KBDG04] in order to deal with drawbacks and limitations of the L_1 distance as it can be estimated from finite samples. The obtained generalization result provided above was further generalized in [MMR09] for an arbitrary loss function $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ using the discrepancy distance:

$$\mathrm{disc}_L(\mathcal{S}, \mathcal{T}) = \max_{h,h'\in\mathcal{H}} \Big| \mathop{\mathbb{E}}_{x\sim\mathcal{S}}[l(h'(x), h(x))] - \mathop{\mathbb{E}}_{x\sim\mathcal{T}}[l(h'(x), h(x))] \Big|$$

that coincides with $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ when l is the zero-one loss.

The established bound in Lemma 1, and consequently those presented in Lemmas 2 and 3, involve distances restricted to the $[y.g_R(x) < \gamma]$ set which is the support of the scaled hinge loss. In this sense, this distance shares some similarity with the $d_{\mathcal{H}\Delta\mathcal{H}}$ distance or more generally with the discrepancy distance as both are related to the considered hypothesis class. The main difference is that we do not take a supremum over a class of hypotheses (similarity functions in our case), but rather concentrate on one learned hypothesis that is $(\epsilon, \gamma)-$good for the source and the associated landmark domains.

These bounds also enclose a worst margin term, analyzed in Section 4, that is comparable to the η term appearing in [CM11, Theorem 2] which represents the greatest deviation between the source and target labelling functions on the source's support (η is defined below). More precisely, they define $h, h' \in \mathcal{H}$ as two minimizers of a certain objective function on the source and target domains respectively, where $\mathcal{H}$ is a hypothesis space and prove the following inequality

$$|l(h'(x), y) - l(h(x), y)| \leq \mu R \sqrt{\frac{\mathrm{disc}(\mathcal{S}, \mathcal{T}) + \mu\eta}{\lambda}}$$

for all $x \in \mathcal{X}, y \in \mathcal{Y}$, where μ is a Lipschitz constant of the loss function l w.r.t. its first argument, λ is the regularization coefficient in the considered objective function, R is a bound on the reproducing kernel Hilbert space of hypotheses and $\eta = \max\{l(f_{\mathcal{S}}(x), f_{\mathcal{T}}(x); x \in \mathrm{supp}\,\mathcal{S}\}$. Besides the different definitions of η and our worst margin term, the latter multiplies the divergence term while the former is added to it. Furthermore, their result is proven for pointwise deviation of losses between the best hypothesis on the source domain and that on the target one while in our case it is defined between the expected losses.

Another counterpart of our worst margin term is found in [GHLM16, Theorem 3] and is given by an $e_{\mathcal{S}}(\rho)$ term. This term is the expected joint error on the source domain of a pair of classifiers drawn from a set of voters $\mathcal{H}$ according to an arbitrary distribution ρ. More precisely, they prove that for any distribution ρ over $\mathcal{H}$

$$\mathbf{R}_{\mathcal{T}}(G_\rho) \leq \frac{1}{2}\mathbf{d}_{\mathcal{T}_{\mathbf{x}}}(\rho) + \beta_{\mathrm{q}}(\mathcal{T}\|\mathcal{S}) \times \mathrm{e}_{\mathcal{S}}(\rho)^{1-\frac{1}{\mathrm{q}}}, \quad (23)$$

where $\beta_q(\mathcal{T}\|\mathcal{S})$ is a divergence term between the two domains and

$$\mathbf{R}_{\mathcal{T}}(G_\rho) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{T}}\left[\mathop{\mathbb{E}}_{h\sim\rho}[[h(x) \neq y]]\right],$$
$$\mathbf{d}_{\mathcal{T}_{\mathbf{x}}}(\rho) = \mathop{\mathbb{E}}_{x\sim\mathcal{T}_{\mathbf{x}}}\left[\mathop{\mathbb{E}}_{h,h'\sim\rho}[[h(x) \neq h'(x)]]\right],$$
$$e_{\mathcal{S}}(\rho) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{T}}\left[\mathop{\mathbb{E}}_{h,h'\sim\rho}[[h(x) \neq y].[h'(x) \neq y]]\right].$$

From (23), we observe a certain similtude with our $\mathcal{M}_{\mathcal{S},\mathcal{R}}(K)$ term as $e_{\mathcal{S}}(\rho)$ also multiplies the divergence term $\beta_q(\mathcal{T}\|\mathcal{S})$ even though the bounds are not directly comparable in general.

Concerning the $(\epsilon, \gamma, \tau)-$good similarity learning framework in particular, we mentioned in the introduction that [MHA12] is the only paper dealing with it in a domain adaptation context. The generalization guarantee presented in their work first proves that their algorithm is robust on the source domain in the sense of the algorithmic robustness presented in [XM10]. The obtained bound for domain adaptation scenario then follows directly from this result and the one presented in Equation (22) leading to:

$$\epsilon_{\mathcal{T}}(h, f_{\mathcal{T}}) \leq \hat{\epsilon}_S(h, f_{\mathcal{S}}) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda + \frac{N_\eta}{\beta B_R + \nu} + \sqrt{\frac{4M_\eta \log 2 + 2\log\frac{1}{\delta}}{r}},$$

where M_η is the η-covering number of $\mathcal{X}$, β and ν are algorithm hyperparameters, $B_R = \min_{x'_j\in S_{\mathcal{R}}}\{\max_{(x_s,x_t)\in S\times T}\big|K(x'_j, x_s) - K(x'_j, x_t)\big|\}$ and $N_\eta = \max_{x_1,x_2\sim\mathcal{S}}\|{}^t\phi^{S_{\mathcal{R}}}(x_1) -^t \phi^{S_{\mathcal{R}}}(x_2)\|_\infty$ where the maximum

is taken over points for which the distance between x_1 and x_2 does not exceed η and ${}^t x$ is a transpose of x.

The main difference between our paper and theirs is that they focus on the performance of the resulting classifier, while we address the similarity performance problem itself. In addition, their paper focuses on an algorithm selecting landmarks making the two distributions close in the projection space, while in our paper, we do not focus on a particular algorithm, and we examine the divergence between the two domains without applying any transformation.

6 Conclusions and future perspectives

Through this paper, we proved guarantees of the $(\epsilon, \gamma)-$goodness of a similarity function on a target domain if assumed to be good on a source domain. A divergence term between the two domains naturally appears when bounding the deviation between the same similarity's performance on both of them. When the source domain's support contains that of the target, we showed in Lemma 3 that the bound is improved via a $\sqrt{\epsilon}$ factor, but a worst margin term remains to be dealt with, thus leading to a section about its estimation. We showed that its convergence to its true value depends on the complexity of the search space of the similarity function, as well as on the regularity of the hinge loss's cumulative distribution function at a neighborhood of its maximum (worst) value. Since this term multiplies the divergence term between the two domains, it gives us a first idea to which extent that divergence must be minimized. Hence, a generalization guarantee involving the divergence and its empirical counterpart, as well as an algorithm that tries to reduce it are crucial future perspectives to be explored. This reduction can be hopefully achieved via a re-weighting procedure applied to the instances of the source domain in a similar approach to that used in [MMR09] to reduce the discrepancy distance, or potentially via other transformations of the data.

Moreover, our new definition of an $(\epsilon, \gamma)-$good similarity uses a landmark domain that is not necessarily included in the source domain. It can be thought of as a "universal landmarks domain" which is independent of the source or target domains. In the case of sentiment classification for example, it might correspond to negative or positive vocabulary used to express one's opinion independently of the type of the concerned product. One problem to be handled is to detect the data points that are close or even included in this landmark domain, and if it is possible to formalize the intuition of using it as a medium of knowledge transfer between the two domains.

References

[BBS08a] Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. Improved guarantees for learning via similarity functions. In *COLT*, pages 287–298, 2008.

[BBS08b] Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008.

[BDBC+10] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

[BFR+00] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *FOCS*, pages 259–269, 2000.

[BGV92] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT*, pages 144–152, 1992.

[BHS12] Aurélien Bellet, Amaury Habrard, and Marc Sebban. Similarity learning for provably accurate sparse linear classification. In *ICML*, 2012.

[BHS13] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

[CH06] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions Information Theory*, 13(1):21–27, 2006.

[CM11] Corinna Cortes and Mehryar Mohri. Domain adaptation in regression. In *ALT*, 2011.

[CNS+11] Bin Cao, Xiaochuan Ni, Jian-Tao Sun, Gang Wang, and Qiang Yang. Distance metric learning under covariate shift. In *IJCAI*, pages 1204–1210, 2011.

[GHLM16] Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A new pac-bayesian perspective on domain adaptation. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 859–868. JMLR.org, 2016.

[GTX11] B. Geng, D. Tao, and C. Xu. Daml: Domain adaptation metric learning. *IEEE Transactions on Image Processing*, 20(10):2980–2989, 2011.

[GY14] Zheng-Chu Guo and Yiming Ying. Guaranteed classification via regularized similarity learning. *Neural Computation*, 26(3):497–522, 2014.

[ISH+15] Nicolae Irina, Marc Sebban, Amaury Habrard, Eric Gaussier, and Massih-Reza Amini. Algorithmic Robustness for Semi-Supervised (ϵ, γ, τ)-Good Metric Learning. In *ICONIP*, page 10, 2015.

[KBDG04] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *VLDB*, pages 180–191, 2004.

[KSD11] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, pages 1785–1792, 2011.

[Kul13] Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013.

[MÖ97] Alfred Müller. Integral Probability Metrics and Their Generating Classes of Functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[Mar11] Anna Margolis. A literature review on domain adaptation with unlabeled data, 2011.

[MHA12] Emilie Morvant, Amaury Habrard, and Stéphane Ayache. Parsimonious unsupervised and semi-supervised domain adaptation with good similarity functions. *Knowledge and Information Systems*, 33(2):309–349, 2012.

[MMR09] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

[NGHS15] Maria-Irina Nicolae, Éric Gaussier, Amaury Habrard, and Marc Sebban. Joint semi-supervised similarity learning for linear classification. In *ECML/PKDD*, pages 594–609, 2015.

[PGLC15] Vishal M. Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015.

[PY10] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

[WKW16] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1), 2016.

[XM10] Huan Xu and Shie Mannor. Robustness and generalization. In *COLT*, pages 503–515, 2010.

[XNJR02] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *NIPS*, pages 521–528, 2002.

[Zol84] V. M. Zolotarev. Probability metrics. *Theory of Probability & Its Applications*, 28(2):278–302, 1984.

[ZSMW13] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *ICML*, pages 819–827, 2013.

[ZZY13] Chao Zhang, Lei Zhang, and Jieping Ye. Generalization bounds for domain adaptation. *CoRR*, abs/1304.1574, 2013.