A survey on domain adaptation theory: learning bounds and theoretical guarantees

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, Younès Bennani

A SURVEY ON DOMAIN ADAPTATION THEORY: LEARNING BOUNDS AND THEORETICAL GUARANTEES

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban
Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d'Optique Graduate School
Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Etienne, France
name.surname@univ-st-etienne.fr

Younès Bennani
Université Sorbonne Paris Nord, CNRS, Institut Galilée
Laboratoire d'Informatique de Paris Nord UMR 7030, F-93430, Villetaneuse, France
name.surname@sorbonne-paris-nord.fr

ABSTRACT

All famous machine learning algorithms that comprise both supervised and semi-supervised learning work well only under a common assumption: the training and test data follow the same distribution. When the distribution changes, most statistical models must be reconstructed from new collected data, which for some applications can be costly or impossible to obtain. Therefore, it has become necessary to develop approaches that reduce the need and the effort to obtain new labeled samples by exploiting data that are available in related areas, and using these further across similar fields. This has given rise to a new machine learning framework known as *transfer learning*: a learning setting inspired by the capability of a human being to extrapolate knowledge across tasks to learn more efficiently. Despite a large amount of different transfer learning scenarios, the main objective of this survey is to provide an overview of the state-of-the-art theoretical results in a specific, and arguably the most popular, sub-field of transfer learning, called *domain adaptation*. In this sub-field, the data distribution is assumed to change across the training and the test data, while the learning task remains the same. We provide a first up-to-date description of existing results related to domain adaptation problem that cover learning bounds based on different statistical learning frameworks.

Keywords Transfer learning · Domain adaptation · Learning theory

This survey is a shortened version of the recently published book **"Advances in Domain Adaptation Theory"** [Redko et al., 2019c] written by the authors of this survey. Its purpose is to provide a high-level overview of the book and to update it with some recent references. All of the proofs and most of the mathematical developments are omitted in this version, to keep the document to a reasonable length. For more details, we refer the interested reader to the original papers or to the full version of the book, available at https://www.elsevier.com/books/advances-in-domain-adaptation-theory/redko/978-1-78548-236-6.

1 Introduction

The idea behind *transfer learning* is inspired by the ability of human beings to learn with minimal or no supervision based on previously acquired knowledge. It is not surprising that this concept was not invented in the machine-learning community in the correct sense of the term, as the concept of "transfer of learning" had been used long before the construction of the first computer, and can be found in papers in the field of psychology from the early 20th century. From the statistical point of view, this learning scenario is different from supervised learning, as transfer learning does

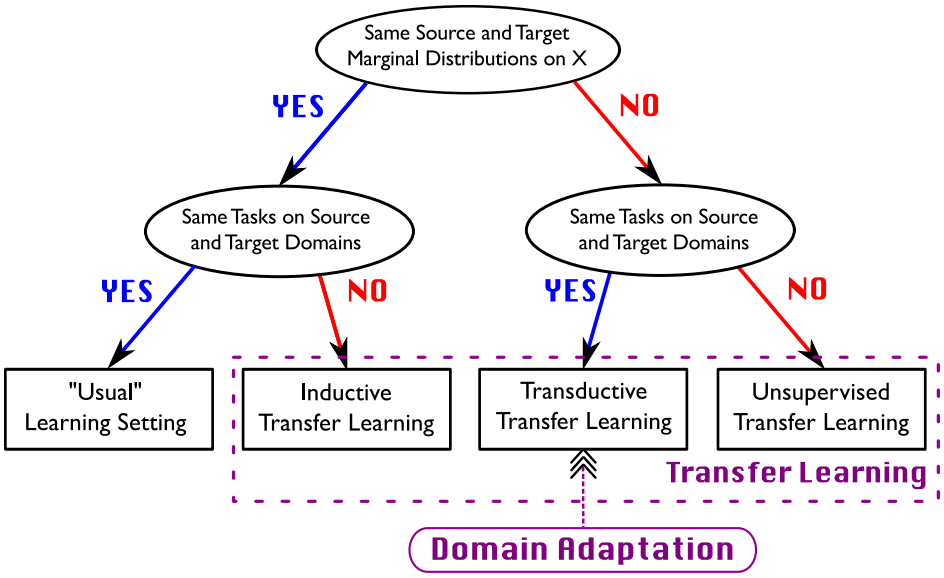

Figure 1: Comparison of standard supervised learning, transfer learning, and positioning of the domain adaptation.

not assume that the training and test data have to be drawn from the same probability distribution. It was argued that this assumption is often too restrictive to hold in practice, as in many real-world applications a hypothesis is learned and deployed in environments that differ and exhibit an important shift. A typical example often used in transfer learning is to consider a spam-filtering task where the spam filter is learned using an arbitrary classification algorithm for a corporate mailbox of a given user. In this case, the vast majority of the e-mails analyzed by the algorithm are likely to be of a professional character, with very few of them being related to the private life of the person considered. Imagine further a situation where this same user installs mailbox software on the personal computer and imports the settings of its corporate mailbox, with the hope that it will work equally well on this too. However, this is not likely to be the case, as many personal e-mails may appear to be spam to an algorithm that has learned purely on professional communications, due to the differences in their content and attached files, as well as the nonuniformity of e-mail addresses. Another illustrative example is that of species classification in oceanographic studies, where experts rely on video coverage of a certain sea area to recognize species of the marine habitat. For instance, in the Mediterranean Sea and in the Indian Ocean, the species of fish that can be found on the recorded videos are likely to belong to the same family, even though their actual appearance might be quite dissimilar due to the different climate and evolutionary backgrounds. In this case, the learning algorithm trained on the video coverage of the Mediterranean Sea will most likely fail to provide correct classification of species in the Indian Ocean without being specifically adapted by an expert.

For these kinds of applications, it might be desirable to find a learning paradigm that can remain robust to a changing environment and can adapt to a new problem at hand, by drawing parallels and exploiting the knowledge from the domain where it was learned initially. In response to this problem, the quest for new algorithms that can learn on a training sample and then provide good performance on a test sample from a different, but related, probability distribution gave rise to a new learning paradigm, known as *transfer learning*. Its definition is given as follows.

Definition 1. *(Transfer learning) We consider a source data distribution $\mathcal{S}$ called the source domain, and a target data distribution $\mathcal{T}$ called the target domain. Let $\mathbf{X}_\mathcal{S} \times Y_\mathcal{S}$ be the source input and output spaces associated to $\mathcal{S}$, and $\mathbf{X}_\mathcal{T} \times Y_\mathcal{T}$ be the target input and output spaces associated to $\mathcal{T}$. We use $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$ to denote the marginal distributions of $\mathbf{X}_\mathcal{S}$ and $\mathbf{X}_\mathcal{T}$, $t_\mathcal{S}$ and $t_\mathcal{T}$ to denote the source and target learning tasks depending on $Y_\mathcal{S}$ and $Y_\mathcal{T}$, respectively. Then, transfer learning aims to help to improve the learning of the target predictive function $f_\mathcal{T} : \mathbf{X}_\mathcal{T} \to Y_\mathcal{T}$ for $t_\mathcal{T}$ using knowledge gained from $\mathcal{S}$ and $t_\mathcal{S}$, where $\mathcal{S} \neq \mathcal{T}$.*

Note that the condition $\mathcal{S} \neq \mathcal{T}$ implies either $\mathcal{S}_\mathbf{X} \neq \mathcal{T}_\mathbf{X}$ (*i.e.*, $\mathbf{X}_\mathcal{S} \neq \mathbf{X}_\mathcal{T}$ or $\mathcal{S}_\mathbf{X}(\mathbf{X}) \neq \mathcal{T}_\mathbf{X}(\mathbf{X})$) or $t_\mathcal{S} \neq t_\mathcal{T}$ (*i.e.*, $Y_\mathcal{S} \neq Y_\mathcal{T}$ or $\mathcal{S}(Y|\mathbf{X}) \neq \mathcal{T}(Y|\mathbf{X})$). In transfer learning, three possible learning settings are often distinguished based on these different relationships (illustrated in Figure 1):

1. **Inductive transfer learning** where $\mathcal{S}_\mathbf{X} = \mathcal{T}_\mathbf{X}$ and $t_\mathcal{S} \neq t_\mathcal{T}$;
   For example, $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$ are the distributions of the data collected from the mailbox of one particular user, where $t_\mathcal{S}$ is the task of detecting spam, while $t_\mathcal{T}$ is the task of detecting a hoax;

2. **Transductive transfer learning** where $\mathcal{S}_{\mathbf{X}} \neq \mathcal{T}_{\mathbf{X}}$ but $t_{\mathcal{S}} = t_{\mathcal{T}}$;
For example, in the spam filtering problem, $\mathcal{S}_{\mathbf{X}}$ is the distribution of the data collected for one user, $\mathcal{T}_{\mathbf{X}}$ is the distribution of the data of another user, and $t_{\mathcal{S}}$ and $t_{\mathcal{T}}$ are both the task of detecting spam;
3. **Unsupervised transfer learning** where $t_{\mathcal{S}} \neq t_{\mathcal{T}}$ and $\mathcal{S}_{\mathbf{X}} \neq \mathcal{T}_{\mathbf{X}}$;
For example, $\mathcal{S}_{\mathbf{X}}$ generates the data collected from one user and $\mathcal{T}_{\mathbf{X}}$ generates the content of web-pages collected on the web, where $t_{\mathcal{S}}$ is to filter out spams, while $t_{\mathcal{T}}$ is to detect hoaxes.

Arguably, the vast majority of situations where transfer learning is most needed fall into the second category. This second category has the name of *domain adaptation*, where we suppose that the source and the target tasks are the same, but where we have a source dataset with an abundant amount of labeled observations and a target dataset with no (or few) labeled instances. In this survey, we concentrate on theoretical advances related to the latter case, and we highlight their differences with respect to the traditional supervised learning paradigm. A brief overview of the contributions presented is given in Tables 1 and 2 for learning bounds and hardness results, respectively.

Table 1: Summary of the learning bounds presented in this survey for domain adaptation. **(Task)** refers to the considered learning problem; **(Framework)** specifies the statistical learning framework used in the analysis; **(Divergence)** is the metric used to compare the source and target distributions; **(Link)** represents the dependence between the source error and the divergence term; **(Non-estim.)** indicates the presence of a nonestimable term in the bounds.

Reference	Learning bounds				
	Task	Framework	Divergence	Link	Non-estim.
[Ben-David et al., 2007] [Blitzer et al., 2008] [Ben-David et al., 2010a]	Binary classification	VC	L^1, $\mathcal{H}\Delta\mathcal{H}$	Add.	+
[Mansour et al., 2009a]	Classification/ Regression	Rademacher	Discrepancy	Add.	+
[Kuroki et al., 2019]	Classification	Rademacher	(S-)Discrepancy	Add.	+
[Cortes et al., 2010] [Cortes and Mohri, 2014] [Cortes et al., 2015]	Regression	Rademacher	(Generalized) Discrepancy	Add.	+
[Mansour et al., 2008]	Classification/ Regression	–	–	–	–
[Mansour et al., 2009b] [Hoffman et al., 2018]	Classification/ Regression	–	Rényi	Mult.	–
[Dhouib and Redko, 2018]	Binary classification/ Similarity learning	–	L^1, χ^2	Mult.	+
[Redko et al., 2019a]	Binary classification	Rademacher	Discrepancy	Add.	+
[Zhang et al., 2012]	Regression/ Classification	Uniform entropy number	IPM	Add.	–
[Redko, 2015]	Regression	Rademacher	IPM/MMD	Add.	+
[Redko et al., 2017]	Regression	–	IPM/Wassertein	Add.	+
[Zhang et al., 2019]	Large-margin classification	Rademacher	IPM	Add.	+
[Dhouib et al., 2020b]	Large margin Binary classification	–	IPM/minimax Wasserstein	Add.	+
[Johansson et al., 2019]	Classification	–	IPM	Add.	+
[Shen et al., 2018]	Classification	–	Wasserstein	Add.	+
[Courty et al., 2017]	Classification	–	Wasserstein	Add.	+
[Germain et al., 2013]	Classification	PAC-Bayes	Domain disagreement	Add.	+
[Germain et al., 2016]	Classification	PAC-Bayes	β-divergence	Mult.	+

[Li and Bilmes, 2007]	Classification	PAC-Bayes	–	Add.	–
[McNamara and Balcan, 2017]	Binary classification	VC/PAC-Bayes	–	Add.	–
[Mansour and Schain, 2014]	Classification	Robustness	λ-shift	Add.	–
[Kuzborskij and Orabona, 2013] [Kuzborskij and Orabona, 2017] [Du et al., 2017]	Regression	Stability	–	–	–
[Perrot and Habrard, 2015]	Classification/ Similarity learning	Stability	–	–	–
[Morvant et al., 2012]	Classification/ Similarity learning	Robustness/VC	$\mathcal{H}\Delta\mathcal{H}$	Add.	+

Table 2: Summary of the contributions presented in this survey for hardness results in domain adaptation. **(Type)** is the type of result obtained; **(Setting)** indicates the presence or absence of target data (either labelled or unlabelled); **(Assumptions)** indicates the assumptions considered (individual or combined); **(Proper)** specifies whether the learned model is required to belong to a predefined class; **(Constr.)** indicates whether the result is of a constructive nature.

Reference	Hardness results				
	Type	Setting	Assumptions	Proper	Constr.
[Ben-David et al., 2010b]	Impossibility/ Sample compl.	Unlabelled target	Cov. shift, $\mathcal{H}\Delta\mathcal{H}$, $\lambda_{\mathcal{H}}$	–	+
[Ben-David et al., 2012]	Impossibility/ Sample compl.	No target/ Unlabelled target	Cov. shift, $C_{\mathcal{B}}$, Lipscht.	+	+/–
[Ben-David and Urner, 2012]	Impossibility/ Sample compl.	Unlabelled target	Cov. shift, $C_{\mathcal{B}}$, Realizab.	–	–
[Redko et al., 2019b]	Estimation/ Sample compl.	Labelled target	–	–	–
[Zhao et al., 2019]	Impossibility	Unlabelled target	Cov. shift, $\mathcal{H}\Delta\mathcal{H}$, $\lambda_{\mathcal{H}}$	–	+
[Johansson et al., 2019]	Impossibility	Unlabelled target	Cov. shift, $\mathcal{H}\Delta\mathcal{H}$, $\lambda_{\mathcal{H}}$	–	+
[Hanneke and Kpotufe, 2019]	Sample compl.	Labelled target	Relaxed cov. shift, Noise cond.	–	–

The rest of this survey is organized as follows. In Section 2, we briefly present the traditional statistical learning frameworks that are referred to throughout the survey. In Section 3, we present the first theoretical results of the domain adaptation theory from the seminal studies of [Ben-David et al., 2007, Mansour et al., 2009a, Cortes and Mohri, 2011] that rely on the famous $\mathcal{H}\Delta\mathcal{H}$ and discrepancy distances. We further turn our attention to hardness results for the domain adaptation problem in Section 4. Section 5 presents several studies that establish the generalization bounds for domain adaptation based on the popular integral probability metrics (IPMs). In Section 6, we highlight several learning bounds defined using the PAC-Bayesian framework. Finally, in Section 7, we give an overview of the contributions that take the actual learning algorithm into account when deriving the learning bounds, and we conclude the survey in Section 8.

2 Preliminary knowledge

Below we recall the usual supervised learning set-up and the different quantities used to derive generalization bounds in this context. This includes the concepts of Vapnik-Chervonenkis (VC) [Vapnik, 2006, Vapnik and Chervonenkis, 1971] and Rademacher complexities [Koltchinskii and Panchenko, 1999], the definitions related to the PAC-Bayesian theory [McAllester, 1999], and those from the more recent algorithmic stability [Bousquet and Elisseeff, 2002] and algorithmic robustness [Xu and Mannor, 2010] frameworks.

2.1 Definitions

Let a pair $(\mathbf{X}, Y)$ define the input and the output spaces where $\mathbf{X}$ is described by real-valued vectors of finite dimension d, *i.e.*, $\mathbf{X} \subseteq \mathbb{R}^d$, and for Y we distinguish between two possible scenarios: 1) when Y is continuous, *e.g.*, $Y = [-1, 1]$ or $Y = \mathbb{R}$, we talk about regression; 2) when Y is discrete and takes values from a finite set, we talk about classification. Two important cases of classification are binary classification and multi-class classification, where $Y = \{-1, 1\}$ (or $Y = \{0, 1\}$) and $Y = \{1, 2, \ldots, C\}$ with $C > 2$, respectively.

We assume that $\mathbf{X} \times Y$ is drawn from an unknown joint probability distribution $\mathcal{D}$ and that we observe them through a finite training sample (also called the *learning sample*) $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim (\mathcal{D})^m$ of m independent and identically distributed (*i.i.d.*) pairs (also called examples or data instances). We further use $\mathcal{H} = \{h | h : \mathbf{X} \to Y\}$ to denote a *hypothesis space* (also called the *hypothesis class*) that consists of functions that map each element of $\mathbf{X}$ to Y. These functions h are usually called hypotheses, or more specifically classifiers or regressors, depending on the nature of Y.

Let us now consider a loss function $\ell : Y \times Y \to [0, 1]$ that gives a cost of $h(\mathbf{x})$ deviating from the true output $y \in Y$. We can define the *true risk* and the empirical risk with respect to $\mathcal{D}$ and S, respectively, as follows.

Definition 2. *(True risk) Given a loss function $\ell : Y \times Y \to [0, 1]$, the true risk (also called the generalization error) $\mathrm{R}_{\mathcal{D}}^{\ell}(h)$ for a given hypothesis $h \in \mathcal{H}$ on a distribution $\mathcal{D}$ over $\mathbf{X} \times Y$ is defined as*

$$\mathrm{R}_{\mathcal{D}}^{\ell}(h) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \ell(h(\mathbf{x}), y).$$

By abuse of notations, for a given pair of hypotheses $(h, h') \in \mathcal{H}^2$, we can write

$$\mathrm{R}_{\mathcal{D}}^{\ell}(h, h') \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \ell(h(\mathbf{x}), h'(\mathbf{x})).$$

Definition 3. *(Empirical risk) Given a loss function $\ell : Y \times Y \to [0, 1]$ and a training sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, where each example is drawn i.i.d. from $\mathcal{D}$, the empirical risk $\mathrm{R}_{\hat{\mathcal{D}}}^{\ell}(h)$ for a given hypothesis $h \in \mathcal{H}$ is defined as*

$$\mathrm{R}_{\hat{\mathcal{D}}}^{\ell}(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i)\,,$$

where $\hat{\mathcal{D}}$ is the empirical distribution associated to the sample S.

The most natural loss function that can be used to count the number of errors committed by hypothesis $h \in \mathcal{H}$ on the distribution $\mathcal{D}$ is the $0-1$ loss function $\ell_{0-1} : Y \times Y \to \{0, 1\}$, which is defined for a training example $(\mathbf{x}, y)$ as

$$\ell_{01}(h(\mathbf{x}), y) = \mathbf{I}\left[h(\mathbf{x}) \neq y\right] = \begin{cases} 1, & \text{if } h(\mathbf{x}) \neq y\,, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

A popular proxy to this nonconvex function is the hinge loss defined for a given pair $(\mathbf{x}, y)$ by

$$\ell_{\text{hinge}}(h(\mathbf{x}), y) = \left[1 - yh(\mathbf{x})\right]_+ = \max\left(0, 1 - yh(\mathbf{x})\right).$$

Another loss function often used in practice that extends the $0-1$ loss to the case of real values is the linear loss $\ell_{\text{lin}} : \mathbb{R} \times \mathbb{R} \to [0, 1]$, defined by:

$$\ell_{\text{lin}}(h(\mathbf{x}), y) = \frac{1}{2}\left(1 - yh(\mathbf{x})\right).$$

The three above-mentioned loss functions are illustrated in Figure 2. Note that in Figure 2, the X-axis are $yh(\mathbf{x})$ values, as $h(\mathbf{x}) = y$ is equivalent to $yh(\mathbf{x}) \geq 0$ when $Y = \{-1, 1\}$.

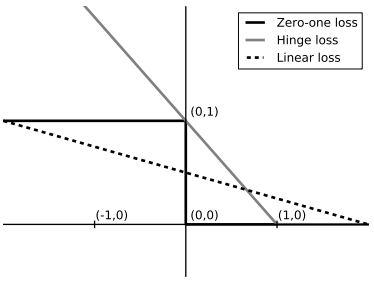

Figure 2: Illustration of different loss functions.

Notations Below, we present the notations that are used throughout the survey.

$\mathbf{X}$	Input space
Y	Output space
$\mathcal{D}$	A domain: a yet unknown distribution over $\mathbf{X} \times Y$
$\mathcal{D}_{\mathbf{X}}$	Marginal distribution of $\mathcal{D}$ on $\mathbf{X}$

$\hat{\mathcal{D}}_{\mathbf{X}}$	Empirical distribution associated with a sample drawn from $\mathcal{D}_{\mathbf{X}}$
$\text{SUPP}(\mathcal{D})$	Support of distribution $\mathcal{D}$
$\mathbf{Pr}(\cdot)$	Probability of an event
$\mathbb{E}(\cdot)$	Expectation of a random variable
$\mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$	A d-dimensional real-valued vector
$(\mathbf{x}, y) \sim \mathcal{D}$	$(\mathbf{x}, y)$ is drawn *i.i.d.* from $\mathcal{D}$
$S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim (\mathcal{D})^m$	Labeled learning sample constituted by m examples drawn *i.i.d.* from $\mathcal{D}$
$S_u = \{(\mathbf{x}_i)\}_{i=1}^m \sim (\mathcal{D}_{\mathbf{X}})^m$	Unlabeled learning sample constituted by m examples drawn *i.i.d.* from $\mathcal{D}_{\mathbf{X}}$
$\lvert S \rvert$	Size of the set S
$\mathcal{H}$	Hypothesis space
$\mathbf{I}[a]$	Indicator function: returns 1 if a is true, 0 otherwise
$\text{sign}[a]$	Return the sign of a: 1 if $a \geq 0$, -1 otherwise
$\mathbf{M}$	An arbitrary matrix
$\mathbf{M}^\top$	Transpose of the matrix $\mathbf{M}$
$\mathbf{0}$	Null vector (matrix)
$\lVert \cdot \rVert_1$	L_1-norm
$\lVert \cdot \rVert_\infty$	L_∞-norm

2.2 Probably approximately correct setting

Statistical learning theory [Vapnik, 1995] provides us with results regarding the conditions that ensure the convergence of the empirical risk to the true risk for a given hypothesis class. These results are known as the *generalization bounds*, and they are usually expressed in the form of probably approximately correct (PAC) inequalities [Valiant, 1984] that have the following form:

$$\underset{S \sim (\mathcal{D})^m}{\mathbf{Pr}} \left\{ \lvert \mathrm{R}_S^\ell(h) - \mathrm{R}_{\mathcal{D}}^\ell(h) \rvert \leq \varepsilon \right\} \geq 1 - \delta,$$

where $\varepsilon > 0$ and $\delta \in (0, 1]$. This expression essentially tells us that we want to upper-bound the gap between the true risk and its estimated value by the smallest possible value of ε and with a high probability over the random choice of the training sample S. The major question now is to understand whether $\mathrm{R}_S^\ell(h)$ converges to $\mathrm{R}_{\mathcal{D}}^\ell(h)$ with an increasing size of the learning sample, and what is the speed of this convergence. We now proceed to a presentation of several theoretical paradigms that were proposed in the literature to show the different characteristics of a learning model or a data sample that this speed can depend on.

2.3 Vapnik-Chervonenkis complexity

Vapnik-Charvonenkis (VC) bounds [Vapnik and Chervonenkis, 1971, Vapnik, 2006] are based on the original definition that allows quantification of the complexity of a given hypothesis class. This concept of complexity is captured by the famous VC dimension that is defined as follows.

Definition 4. *(VC dimension) The VC dimension VC($\mathcal{H}$) of a given hypothesis class $\mathcal{H}$ for the problem of binary classification is defined as the largest possible cardinality of some subset $\mathbf{X}' \subset \mathbf{X}$ for which there exists a hypothesis $h \in \mathcal{H}$ that perfectly classifies elements from $\mathbf{X}'$ whatever their labels are. More formally, we have*

$$VC(\mathcal{H}) = \max\{\lvert \mathbf{X}' \rvert : \forall y_i \in \{-1, +1\}^{\lvert \mathbf{X}' \rvert}, \exists h \in \mathcal{H} \text{ so that } \forall \mathbf{x}_i \in \mathbf{X}', h(\mathbf{x}_i) = y_i\}.$$

As follows from the definition, the VC dimension is the cardinality of the biggest subset of a given sample that can be subject to perfect classification provided by a hypothesis from $\mathcal{H}$ for all possible labelings of its observations. To illustrate this, we can consider the classical example given in Figure 3, where the hypothesis class $\mathcal{H}$ consists of half-planes in $\mathbb{R}^d$. In this particular case with $d = 2$, we can perfectly classify only $d + 1$ elements, regardless their labeling, as for the case with $d + 2$ points this will no longer be possible. This means that the VC dimension of the class of half-planes in $\mathbb{R}^d$ is $d + 1$. Note that the result obtained reveals that in this particular scenario, the VC dimension is equal to the number of parameters needed to define the function of the hypothesis plane. This, however, is not true in general, as some classes might have an infinite VC dimension despite the finite number of parameters needed to define the hypothesis class. A common

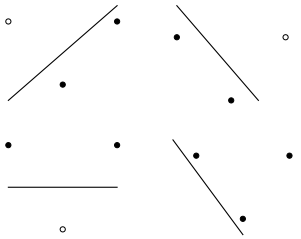

Figure 3: Illustration of the Vapnik-Charvonenkis (VC) dimension. Here, half-planes in $\mathbb{R}^d$ with $d = 2$ can correctly classify at most three points for all possible 2^3 labelings. The VC dimension here is $2 + 1$.

example used in the literature to show this is given by

$$\mathcal{H} = \{h_\theta(\mathbf{x}) : \mathbf{X} \to \{0,1\} : h_\theta(\mathbf{x}) = \frac{1}{2}\sin(\theta\mathbf{x}), \theta \in \mathbb{R}\}.$$

It can be proven that the VC dimension of this class is infinite.

The following theorem uses the VC dimension of a hypothesis class to upper-bound the gap between the true and the empirical error for a given loss function and a finite sample of size m.

Theorem 1. *Let $\mathbf{X}$ be an input space, $Y = \{-1, +1\}$ the output space, and $\mathcal{D}$ their joint distribution. Let S be a finite sample of size m drawn i.i.d. from $\mathcal{D}$, and $\mathcal{H} = \{h : X \to Y\}$ be a hypothesis class of VC dimension $VC(\mathcal{H})$. Then for any $\delta \in (0, 1]$ with probability of at least $1 - \delta$ over the random choice of the training sample $S \sim (\mathcal{D})^m$, the following holds*

$$\forall h \in \mathcal{H}, \quad \mathrm{R}^\ell_{\mathcal{D}}(h) \leq \mathrm{R}^\ell_S(h) + \sqrt{\frac{4}{m}\left(VC(\mathcal{H})\ln\frac{2em}{VC(\mathcal{H})} + \ln\frac{4}{\delta}\right)}.$$

2.4 Rademacher complexity

Intuitively, the Rademacher complexity measures the capacity of a given hypothesis class to resist against noise that might be present in the data. This, in turn, was shown to lead to more accurate bounds than those based on the VC dimension [Koltchinskii and Panchenko, 1999]. To present the Rademacher bounds, we first provide a definition of a Rademacher variable.

Definition 5. *(Rademacher variable) A random variable κ defined as*

$$\kappa = \begin{cases} 1, \text{ with probability } \frac{1}{2} \\ -1, \text{ otherwise}, \end{cases}$$

is called the Rademacher variable.

From this definition, a Rademacher variable defines a random binary labeling as it takes values -1 and 1 with equal probability and allows the introduction of the Rademacher complexity for an unlabeled sample of size m, as follows.

Definition 6. *(Rademacher complexity) For a given unlabeled sample $S = \{(\mathbf{x}_i)\}_{i=1}^m$ and a given hypothesis class $\mathcal{H}$, the Rademacher complexity is defined as follows:*

$$\mathcal{R}_S(\mathcal{H}) = \mathop{\mathbf{E}}_{\boldsymbol{\kappa}}\left[\sup_{h\in\mathcal{H}}\frac{2}{m}\sum_{i=1}^{m}\kappa_i h(\mathbf{x}_i)\right],$$

where $\boldsymbol{\kappa}$ is a vector of m independent Rademacher variables. The Rademacher complexity for the whole hypothesis class is thus defined as the expected value of $\mathcal{R}_S(\mathcal{H})$ by

$$\mathcal{R}_m(\mathcal{H}) = \mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m}\mathcal{R}_S(\mathcal{H}).$$

In this definition, $\mathcal{R}_S(\mathcal{H})$ encodes the complexity of a given hypothesis class $\mathcal{H}$ based on the observed sample S, while $\mathcal{R}_m(\mathcal{H})$ is the expected value of this complexity over all possible samples that were drawn from some joint probability distribution. Contrary to the VC dimension, this complexity measure is defined in terms of the expected value over all labelings, and not only the worst one. The following theorem presents the Rademacher-based generalization bound [Koltchinskii and Panchenko, 1999, Bartlett and Mendelson, 2002].

Theorem 2. *Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ be a finite sample of m examples drawn i.i.d. from $\mathcal{D}$, and $\mathcal{H} = \{h : \mathbf{X} \to Y\}$ be a hypothesis class. Then, for any $\delta \in (0, 1]$ with probability of at least $1 - \delta$ over the choice of the sample $S \sim (\mathcal{D})^m$, the following holds*

$$\forall h \in \mathcal{H}, \quad \mathrm{R}^\ell_{\mathcal{D}}(h) \;\leq\; \mathrm{R}^\ell_S(h) + \mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

2.5 PAC-Bayesian bounds

The PAC-Bayesian approach [Shawe-Taylor and Williamson, 1997, McAllester, 1999] provides generalization bounds for a hypothesis expressed as a weighted majority vote over the hypothesis space $\mathcal{H}$, as, for instance, in ensemble methods [Dietterich, 2000, Re and Valentini, 2012]. In this section, we recall the general PAC-Bayesian generalization bound as presented in [Germain et al., 2015] in the setting of binary classification, where $Y = \{-1, 1\}$ with the $0-1$ loss or the linear loss. To derive such a generalization bound, a prior distribution π over $\mathcal{H}$ is assumed, which models an *a-priori* belief on the hypotheses from $\mathcal{H}$ before the observation of the training sample $S \sim (\mathcal{D})^m$. Given S, the learner aims to find a posterior distribution ρ over $\mathcal{H}$ that leads to a well-performing ρ-weighted majority vote $B_\rho(\mathbf{x})$ (also called the Bayes classifier), defined as

$$B_\rho(\mathbf{x}) = \operatorname{sign}\left[\mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x})\right].$$

In other words, rather than finding the best hypothesis from $\mathcal{H}$, we want to learn ρ over $\mathcal{H}$, such that this minimizes the true risk $\mathrm{R}_\mathcal{D}(B_\rho)$ of the ρ-weighted majority vote. However, PAC-Bayesian generalization bounds do not directly focus on the risk of the deterministic ρ-weighted majority vote B_ρ, but on giving an upper bound over the expectation over ρ of all of the individual hypothesis true risks, called the *Gibbs risk*: $\mathbf{E}_{h\sim\rho}\,\mathrm{R}^\ell_\mathcal{D}(h)$. The Gibbs risk is associated to a stochastic classifier, called the Gibbs classifier, which draws a hypothesis h from $\mathcal{H}$ according to the posterior distribution ρ, and predicts the label of $\mathbf{x}$ given by $h(\mathbf{x})$. An important behavior of the Gibbs risk is that it is closely related to the deterministic ρ-weighted majority vote. Indeed, if B_ρ miss-classifies $\mathbf{x} \in \mathbf{X}$, then at least half of the classifiers (under measure ρ) make a prediction error on $\mathbf{x}$. Therefore, we have

$$\mathrm{R}^\ell_\mathcal{D}(B_\rho) \leq 2 \mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^\ell_\mathcal{D}(h). \tag{2}$$

Thus, an upper bound on $\mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^\ell_\mathcal{D}(h)$ provides an upper bound on $\mathrm{R}^\ell_\mathcal{D}(B_\rho)$ as well.

Note that PAC-Bayesian generalization bounds do not directly take into account the complexity of the hypothesis class $\mathcal{H}$, contrary to the Rademacher complexity or the VC dimension, but they measure the deviation between the prior distribution π and the posterior distribution ρ on $\mathcal{H}$ through the Kullback-Leibler divergence:

$$\mathrm{KL}(\rho|\pi) \;=\; \mathop{\mathbf{E}}_{h\sim\rho} \ln \frac{\rho(h)}{\pi(h)}.$$

The result that follows is a general PAC-Bayesian theorem that takes the form of an upper bound on the deviation between the true and empirical Gibbs risks when measured by a convex function $D : [0,1] \times [0,1] \to \mathbb{R}$.

Theorem 3 ([Germain et al., 2009, Germain et al., 2015])**.** *For any distribution $\mathcal{D}$ on $\mathbf{X} \times Y$, for any hypothesis class $\mathcal{H}$, for any prior distribution π on $\mathcal{H}$, for any $\delta \in (0,1]$, for any convex function $D : [0,1] \times [0,1] \to \mathbb{R}$, with a probability of at least $1-\delta$ over the random choice of $S \sim (\mathcal{D})^m$, we have, for all posterior distribution ρ on $\mathcal{H}$,*

$$D\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^\ell_S(h), \mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^\ell_\mathcal{D}(h)\right) \leq \frac{1}{m}\left[\mathrm{KL}(\rho|\pi) + \ln\left(\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m} \mathop{\mathbf{E}}_{h\sim\pi} e^{m\, D\left(\mathrm{R}^\ell_S(h), \mathrm{R}^\ell_\mathcal{D}(h)\right)}\right)\right].$$

By upper-bounding $\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m} \mathop{\mathbf{E}}_{h\sim\pi} e^{m\, D(\mathrm{R}^\ell_S(h))}$ and by selecting a well-suited deviation function D, we can retrieve the classical versions of the PAC-Bayesian theorem (*i.e.*, [McAllester, 1999, Seeger, 2002, Catoni, 2007]).

2.6 Uniform stability

As the complexity of the hypothesis class intuitively depends directly on the properties of a learning algorithm, it might be desirable to have the generalization bounds that manifest this relationship explicitly. [Bousquet and Elisseeff, 2002] introduced generalization bounds that provide a solution to this problem based on the concept of uniform stability of a learning algorithm. We now give its definition.

Definition 7. *(Uniform stability) An algorithm $\mathcal{A}$ has uniform stability β with respect to the loss function ℓ if the following holds*

$$\forall S \in \{\mathbf{X} \times Y\}^m, \forall i \in \{1, \ldots, m\}, \sup_{(\mathbf{x},y)\in S} |\ell(h_S(\mathbf{x}), y) - \ell(h_{S^{\setminus i}}(\mathbf{x}), y)| \leq \beta\,,$$

where the hypothesis h_S is learned on the sample S while $h_{S^{\setminus i}}$ is obtained on S with its i^{th} observation being deleted.

The intuition behind this definition is to say that an algorithm that is expected to generalize well should be robust to small perturbations in the training sample. Consequently, stable algorithms should have an empirical error that remains close to their generalization error. This idea is confirmed by the following theorem.

Theorem 4. *Let $\mathcal{A}$ be an algorithm with uniform stability β with respect to a loss function ℓ, such that $0 \leq \ell(h_S(\mathbf{x}, y) \leq M$, for all $(\mathbf{x}, y) \in (\mathbf{X} \times Y)$ and all sets S. Then, for any $m \geq 1$, and any $\delta \in (0, 1]$, the following bound holds with probability of at least $1 - \delta$ over the random choice of the sample S,*

$$\mathrm{R}^{\ell}_{\mathcal{D}}(h_S) \leq \mathrm{R}^{\ell}_{S}(h_S) + 2\beta + (4m\beta + M)\sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

2.7 Algorithmic robustness

The main underlying idea of algorithmic robustness [Xu and Mannor, 2010, Xu and Mannor, 2012] is to say that a robust algorithm should have similar performance in terms of the classification error for testing and training samples that are close. The measure of similarity used to define whether two samples are close or not relies on partitioning the joint space $\mathbf{X} \times Y$ in a way that puts two similar points of the same class in the same partition. This partition is further defined using the concept of covering numbers [Kolmogorov and Tikhomirov, 1959], as introduced below.

Definition 8. *(Covering number) Let (Z, ϱ) denote a metric space with metric $\varrho(\cdot)$ defined on Z. For $Z' \subset Z$, we say that $\hat{Z}'$ is a γ covering of Z', if for any element $t \in Z'$ there is an element $\hat{t} \in \hat{Z}'$ such that $\varrho(t, \hat{t}) \leq \gamma$. Then the number of γ covering of Z' is expressed as*

$$N(\gamma, Z', \varrho) = \min\left\{\left|\hat{Z}'\right| : \hat{Z}' \text{ is a } \gamma\text{-covering of } Z'\right\}.$$

In the case where $\mathbf{X}$ is a compact space, its covering number $N(\gamma, \mathbf{X}, \varrho)$ is finite. Furthermore, for the product space $\mathbf{X} \times Y$, the number of γ-covering is also finite and is equal to $|Y|\, N(\gamma, \mathbf{X}, \varrho)$. As previously explained, the above partitioning ensures that two points from the same subset are from the same class and are close to each other with respect to metric ϱ. Bearing this in mind, the algorithmic robustness is defined as follows.

Definition 9. *(Algorithmic robustness) Let S be a training sample of size m where each example is drawn from the joint distribution $\mathcal{D}$ on $\mathbf{X} \times Y$. An algorithm $\mathcal{A}$ is $(M, \epsilon(\cdot))$-robust on $\mathcal{D}$ with respect to a loss function ℓ for $M \in \mathbb{N}$ and $\epsilon(\cdot) : (\mathbf{X} \times Y)^m \to \mathbb{R}$ if $\mathbf{X} \times Y$ can be partitioned into M disjoint subsets denoted by $\{Z_j\}_{j=1}^{M}$, so that for all $(\mathbf{x}, y) \in \mathbf{X} \times Y$, $(\mathbf{x}', y')$ drawn from $\mathcal{D}$ and $j \in \{1, \ldots, M\}$ we have*

$$\big((\mathbf{x}, y), (\mathbf{x}', y')\big) \in Z_j^2 \quad \longrightarrow \quad |\ell(h_S(\mathbf{x}), y) - \ell(h_S(\mathbf{x}'), y')| \leq \epsilon(S)\,,$$

where h_S is a hypothesis learned by $\mathcal{A}$ on S.

We are now ready to present the generalization guarantees that characterize robust algorithms that verify the definition presented above.

Theorem 5. *Let S be a finite sample of size m drawn i.i.d. from $\mathcal{D}$, $\mathcal{A}$ be $(M, \epsilon(\cdot))$-robust on $\mathcal{D}$ with respect to a loss function $\ell(\cdot, \cdot)$, such that $0 \leq \ell(h_S(\mathbf{x}), y) \leq M_\ell$, for all $(\mathbf{x}, y) \in (\mathbf{X} \times Y)$. Then, for any $\delta \in (0, 1]$, the following bound holds with probability of at least $1 - \delta$ over the random draw of the sample $S \sim (\mathcal{D})^m$,*

$$\mathrm{R}^{\ell}_{\mathcal{D}}(h_S) \leq \mathrm{R}^{\ell}_{S}(h_S) + \epsilon(S) + M_\ell\sqrt{\frac{2M \ln 2 + 2 \ln \frac{1}{\delta}}{m}}\,,$$

where h_S is a hypothesis learned by $\mathcal{A}$ on S.

Note that the algorithmic robustness focuses on measuring the divergence between the costs associated to two similar points, assuming that the learned hypothesis function should be locally consistent. Uniform stability, in turn, explores the variation in the cost due to perturbations of the training sample, and thus assumes that the learned hypothesis does not change much.

3 Seminal divergence-based learning bounds

In this section, we provide the description of domain adaptation generalization bounds that laid the foundation of this field. These seminal bounds mainly relied on traditional divergence measures between the probability distributions, to relate the source and target domains.

3.1 Learning bound based on the L^1-distance

From a theoretical point of view, the domain adaptation problem was rigorously investigated for the first time by [Ben-David et al., 2007] and [Ben-David et al., 2010a][1]. The authors of these papers focused on the domain adaptation problem following VC theory (recalled in Section 2.3) and considered the $0-1$ loss (Equation 1) function in the setting of binary classification with $Y = \{-1, +1\}$. They further proposed to make use of the L^1-*distance*, the definition of which is given below.

Definition 10. *(L^1-distance) Let $\mathcal{B}$ denote the set of measurable subsets under two probability distributions $\mathcal{D}_1$ and $\mathcal{D}_2$. The L^1-distance or the total variation distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ is defined as*

$$d_1(\mathcal{D}_1, \mathcal{D}_2) = 2 \sup_{B \in \mathcal{B}} \left| \underset{\mathcal{D}_1}{\mathbf{Pr}} (B) - \underset{\mathcal{D}_2}{\mathbf{Pr}} (B) \right| .$$

The L^1-distance is a proper metric on the space of probability distributions that informally quantifies the largest possible difference between the probabilities that the two probability distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ can assign to the same event B. This distance is relatively popular in many real-world applications, such as image denoising or numerical approximations of partial derivative equations.

Starting from Definition 10, the first important result from their work was formulated as follows.

Theorem 6 ([Ben-David et al., 2007])**.** *Given two domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X} \times Y$ and a hypothesis class $\mathcal{H}$, the following holds*

$$\forall h \in \mathcal{H}, \quad \mathrm{R}_{\mathcal{T}}^{\ell_{01}}(h) \leq \mathrm{R}_{\mathcal{S}}^{\ell_{01}}(h) + d_1(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \min \left\{ \underset{\mathbf{x} \sim \mathcal{S}_{\mathbf{X}}}{\mathbf{E}} \left[|f_{\mathcal{S}}(\mathbf{x}) - f_{\mathcal{T}}(\mathbf{x})| \right], \underset{\mathbf{x} \sim \mathcal{T}_{\mathbf{X}}}{\mathbf{E}} \left[|f_{\mathcal{T}}(\mathbf{x}) - f_{\mathcal{S}}(\mathbf{x})| \right] \right\} ,$$

where $f_{\mathcal{S}}(\mathbf{x})$ and $f_{\mathcal{T}}(\mathbf{x})$ are the source and target true labeling functions associated to $\mathcal{S}$ and $\mathcal{T}$, respectively.

This theorem presents the first result that relates the performance of a given hypothesis function with respect to two different domains. It implies that the error achieved by a hypothesis in the source domain upper-bounds the true error on the target domain where the tightness of the bound depends on the distance between their distributions and that of the labeling functions.

3.2 Learning bound based on $\mathcal{H}\Delta\mathcal{H}$-divergence

Despite being the first result of this kind proposed in the literature, the idea of bounding the error in terms of the L^1-distance between the marginal distributions of the two domains includes two important restrictions: 1) the L^1-distance cannot be estimated from finite samples for arbitrary probability distributions; and 2) it does not allow the divergence measure to be linked to the considered hypothesis class, and thus leads to very loose inequality.

To address these issues, the authors further defined the $\mathcal{H}\Delta\mathcal{H}$-divergence based on the $\mathcal{A}$-divergence introduced in [Kifer et al., 2004] for detection of changes in data streams. We give its definition below.

Definition 11 (Based on [Kifer et al., 2004])**.** *Given two domains' marginal distributions $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$ over the input space $\mathbf{X}$, let $\mathcal{H}$ be a hypothesis class, and let $\mathcal{H}\Delta\mathcal{H}$ denote the symmetric difference hypothesis space defined as*

$$h \in \mathcal{H}\Delta\mathcal{H} \iff h(\mathbf{x}) = g(\mathbf{x}) \oplus g'(\mathbf{x}) ,$$

for some $(g, g')^2 \in \mathcal{H}^2$, where $\oplus$ stands for the XOR operation. Let $I(h)$ denote the set for which $h \in \mathcal{H}\Delta\mathcal{H}$ is the characteristic function, i.e., $\mathbf{x} \in I(h) \Leftrightarrow g(\mathbf{x}) = 1$. The $\mathcal{H}\Delta\mathcal{H}$-divergence between $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$ is defined as:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) = 2 \sup_{h \in \mathcal{H}\Delta\mathcal{H}} \left| \underset{\mathcal{S}_{\mathbf{X}}}{\mathbf{Pr}} (I(h)) - \underset{\mathcal{T}_{\mathbf{X}}}{\mathbf{Pr}} (I(h)) \right| .$$

The $\mathcal{H}\Delta\mathcal{H}$-divergence solves both problems associated with the L^1-distance. First, from its definition, we can see that $\mathcal{H}\Delta\mathcal{H}$-divergence explicitly takes into account the considered hypothesis class. This ensures that the bound remains meaningful and directly related to the learning problem at hand. On the other hand, the $\mathcal{H}\Delta\mathcal{H}$-divergence for any class $\mathcal{H}$ is never larger than the L^1-distance, and thus can lead to a tighter bound. Finally, for a given hypothesis class $\mathcal{H}$ of finite VC dimension, the $\mathcal{H}\Delta\mathcal{H}$-divergence can be estimated from finite samples using the following lemma.

[1] Note that in [Ben-David et al., 2010a], the authors presented an extended version of the results previously published in [Ben-David et al., 2007] and [Blitzer et al., 2008].

Lemma 7. *Let $\mathcal{H}$ be a hypothesis space of VC dimension VC($\mathcal{H}$). If S_u, T_u are unlabeled samples of size m each, drawn independently from $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$, respectively, then for any $\delta \in (0,1)$ with probability of at least $1-\delta$ over the random choice of the samples, we have*

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \leq \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(S_u, T_u) + 4\sqrt{\frac{2\,VC(\mathcal{H})\log(2m) + \log(\frac{2}{\delta})}{m}}\,, \tag{3}$$

where $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(S_u, T_u)$ is the empirical $\mathcal{H}\Delta\mathcal{H}$-divergence estimated on S_u and T_u.

Inequality (3) shows that with an increasing number of instances and for a hypothesis class of finite VC dimension, the empirical $\mathcal{H}\Delta\mathcal{H}$-divergence can be a good proxy for its true counterpart. The former can be further calculated thanks to the following result.

Lemma 8 ([Ben-David et al., 2010a])**.** *Let $\mathcal{H}$ be a hypothesis space. Then, for two unlabeled samples S_u, T_u of size m, we have*

$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(S_u, T_u) = 2\left(1 - \min_{h \in \mathcal{H}\Delta\mathcal{H}} \left[\frac{1}{m}\sum_{\mathbf{x}:h(\mathbf{x})=0} \mathbf{I}\left[\mathbf{x} \in S_u\right] + \frac{1}{m}\sum_{\mathbf{x}:h(\mathbf{x})=1} \mathbf{I}\left[\mathbf{x} \in T_u\right]\right]\right).$$

It can be noted that the expression of the empirical $\mathcal{H}\Delta\mathcal{H}$-divergence given above is essentially the error of the best classifier for the binary classification problem of distinguishing between the source and target instances pseudo-labeled with 0's and 1's. In practice, this means that the value of the $\mathcal{H}\Delta\mathcal{H}$-divergence depends explicitly on the hypothesis class used to produce such a classifier. This dependence and the intuition behind the $\mathcal{H}\Delta\mathcal{H}$-divergence are illustrated in Figure 4. In Figure 4, we consider two different domain adaptation problems, where for one of them the source and target samples are well separated, while for the other, the source and target data are mixed together. To calculate the value of the $\mathcal{H}\Delta\mathcal{H}$-divergence, we need to choose a hypothesis class used to produce a classifier that distinguishes between them. Here, we consider two different families of classifiers: a linear support vector machine classifier, and its nonlinear version with radial basis function kernels. For each solution, we also plot the decision boundaries to see how the source and target instances are classified in both cases. From the visualization of the decision boundaries, we note that the linear classifier fails to distinguish between the mixed source and target instances, while the nonlinear classifier manages to do this relatively well. This is reflected by the value of the $\mathcal{H}\Delta\mathcal{H}$-divergence, which is zero in the first case for both classifiers, and is drastically different for the second adaptation problem. Having two different divergence values for the same adaptation problem might appear surprising at first sight, but this has a simple explanation. By choosing a richer hypothesis class composed of nonlinear functions, we have increased the VC complexity of the considered hypothesis space, and have thus increased the complexity term in Lemma 7. This shows the trade-off that has to be borne in mind when the $\mathcal{H}\Delta\mathcal{H}$-divergence is calculated in the same way as is suggested by general VC theory.

At this point, we already have a "reasonable" version of the L^1-distance used to derive the first seminal result. We have also presented its finite sample approximation, but we have not yet applied this to relate the source and target error functions. The next lemma gives the final key needed to obtain a learning bound for domain adaptation that is linked to a specific hypothesis class and is derived for the available source and target finite size samples. This reads as follows.

Lemma 9 ([Ben-David et al., 2010a])**.** *Let $\mathcal{S}$ and $\mathcal{T}$ be two domains on $\mathbf{X} \times Y$. For any pair of hypotheses $(h, h') \in \mathcal{H}\Delta\mathcal{H}^2$, we have*

$$\left|\mathrm{R}_{\mathcal{T}}^{\ell_{01}}(h, h') - \mathrm{R}_{\mathcal{S}}^{\ell_{01}}(h, h')\right| \;\leq\; \frac{1}{2}\, d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}).$$

Note that in this lemma, the source and target risk functions are defined for the same pairs of hypotheses, while the true risk should be calculated based on a given hypothesis and the corresponding labeling function. This result presents the complete learning bound for domain adaptation with $\mathcal{H}\Delta\mathcal{H}$-divergence, and it is established by means of the following theorem.

Theorem 10 ([Ben-David et al., 2010a])**.** *Let $\mathcal{H}$ be a hypothesis space of VC dimension VC($\mathcal{H}$). If S_u, T_u are unlabeled samples of size m' each, which are drawn independently from $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$, respectively, then for any $\delta \in (0,1)$ with probability of at least $1-\delta$ over the random choice of the samples, then for all $h \in \mathcal{H}$*

$$\mathrm{R}_{\mathcal{T}}^{\ell_{01}}(h) \;\leq \mathrm{R}_{\mathcal{S}}^{\ell_{01}}(h) + \tfrac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(S_u, T_u) + 4\sqrt{\frac{2\,VC(\mathcal{H})\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda\,,$$

where λ is the combined error of the ideal hypothesis h^* that minimizes $\mathrm{R}_{\mathcal{S}}(h) + \mathrm{R}_{\mathcal{T}}(h)$.

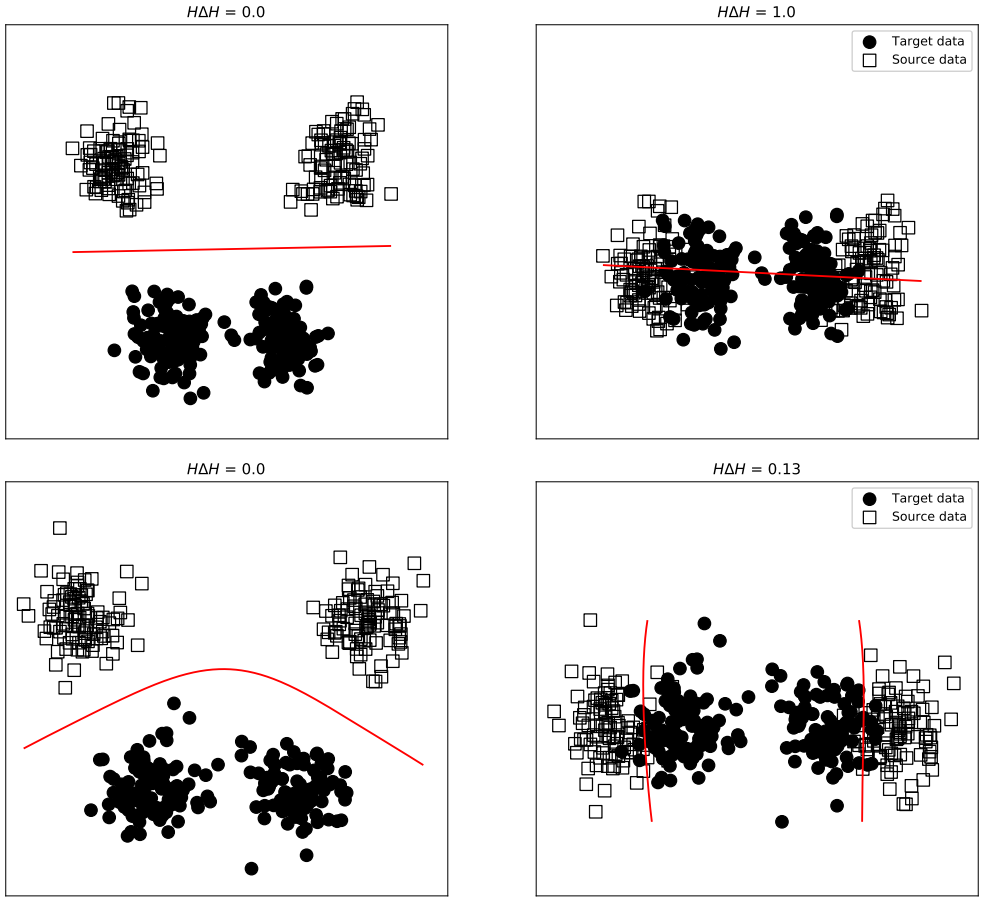

Figure 4: Illustration of the $\mathcal{H}\Delta\mathcal{H}$-divergence when the hypothesis class consists of linear (**top row**) and nonlinear (**bottom row**) classifiers. Note that the indicated value of $\mathcal{H}\Delta\mathcal{H}$ is the error of the obtained classifier without subtracting 1 and multiplying the result by two, as in Lemma 8.

As indicated at the beginning of this section, a meaningful domain adaptation generalization bound should include two terms that reflect both the divergence between the marginal distribution of the source and target domains, and the divergence between their labeling functions. The first term here is obviously reflected by the $\mathcal{H}\Delta\mathcal{H}$-divergence between the observable samples, while the second term is given by the λ term, as it depends on the true labels (and can be seen as a measure of capacity to adapt). The presence of the trade-off between source risk, divergence, and capability to adapt is a very important phenomenon in domain adaptation. Indeed, it shows that the reduction in the divergence between the samples can be insufficient when there is no hypothesis that can achieve a low error on both the source and target samples.

The semi-supervised case In the unsupervised case that we have considered previously, it is assumed that there is no access to labeled instances in the target domain that can help to guide adaptation. For this case, the main strategy that leads to an efficient adaptation is to have a classifier learned on a target-aligned labeled sample from the source domain, and to apply it directly in the target domain afterwards. While this situation occurs relatively often in practice, many applications can be found where several labeled target instances are available during the learning stage. In what follows, we consider this situation and give a generalization bound for it, which shows that the error obtained by a classifier that has been learned on a mixture of source and target labeled data can be upper-bounded by the error of the best classifier learned using the target domain data only.

To proceed, let us now assume that we have βm instances drawn independently from $\mathcal{T}$ and $(1-\beta)m$ instances drawn independently from $\mathcal{S}$ and labeled by $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$, respectively. A natural goal for this setting is to use the available labeled instances from the target domain to find a trade-off between minimizing the source and the target errors depending on the number of instances available in each domain and the distance between them. In this case, we can consider the empirical combined error [Blitzer et al., 2008] defined as a convex combination of errors on the source

and target training data for $\alpha \in [0,1]$:

$$\hat{\mathrm{R}}^{\alpha}(h) = \alpha \mathrm{R}^{\ell_{01}}_{\hat{\mathcal{T}}}(h) + (1-\alpha)\mathrm{R}^{\ell_{01}}_{\hat{\mathcal{S}}}(h).$$

The use of the combined error is motivated by the fact that if the number of instances in the target sample is small compared to the number of instances in the source domain (which is usually the case in domain adaptation), minimizing only the target error might not be appropriate. Instead, there might be the need to find a suitable value of α that ensures the minimum of $\mathrm{R}^{\alpha}(h)$ with respect to a given hypothesis h. Note that in this case, the shape of the generalization bound that we are interested in becomes different. Indeed, in all previous theorems the goal was to upper-bound the target error by the source error, while in this case we would like to know whether learning a classifier minimizing the combined error is better than minimizing the target error using the available labeled instances alone. The answer to this question is given by the following theorem.

Theorem 11 ([Blitzer et al., 2008, Ben-David et al., 2010a])**.** *Let $\mathcal{H}$ be a hypothesis space of VC dimension VC($\mathcal{H}$). Let $\mathcal{S}$ and $\mathcal{T}$ be the source and target domains, respectively, defined on $\mathbf{X} \times Y$. Let S_u, T_u be unlabeled samples of size m' each, drawn independently from $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$, respectively. Let S be a labeled sample of size m generated by drawing $\beta\, m$ points from $\mathcal{T}$ ($\beta \in [0,1]$) and $(1-\beta)\, m$ points from $\mathcal{S}$ and labeling them according to $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$, respectively. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\hat{\mathrm{R}}^{\alpha}(h)$ on S and $h^*_T = \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{R}^{\ell_{01}}_{\mathcal{T}}(h)$ then for any $\delta \in (0,1)$, with probability of at least $1-\delta$ over the random choice of the samples, we have*

$$\mathrm{R}^{\ell_{01}}_{\mathcal{T}}(\hat{h}) \;\leq\; \mathrm{R}^{\ell_{01}}_{\mathcal{T}}(h^*_T) + c_1 + c_2\,,$$

where

$$c_1 \;=\; 4\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2\,VC(\mathcal{H})\log(2(m+1)) + 2\log(\frac{8}{\delta})}{m}}\,,$$

$$\text{and } c_2 \;=\; 2(1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(S_u, T_u) + 4\sqrt{\frac{2\,VC(\mathcal{H})\log(2m') + \log(\frac{8}{\delta})}{m'}} + \lambda\right). \tag{4}$$

This theorem presents an important result that reflects the usefulness of the combined minimization of the source and target errors based on the available labeled samples in both domains compared to the minimization of the target error only. This essentially shows that the error achieved by the best hypothesis of the combined error in the target domain is always upper-bounded by the error achieved by the hypothesis of the best target domain. Furthermore, this indicates two important consequences:

1. if $\alpha = 1$, the term related to the $\mathcal{H}\Delta\mathcal{H}$-divergence between the domains disappears, as in this case we have enough labeled data in the target domain and a low-error hypothesis can be produced solely from the target data;
2. if $\alpha = 0$, the only way to produce a low-error classifier on the target domain is to find a good hypothesis in the source domain while minimizing the $\mathcal{H}\Delta\mathcal{H}$-divergence between the domains. In this case, it has also to be assumed that λ is low, so that the adaptation is possible.

Additionally, Theorem 11 can provide some insights into the optimal mixing value of α depending on the quantity of labeled instances in the source and target domains. To illustrate this, the right-hand side of Equation (4) can be rewritten as a function of α, to understand when this function is minimized. This gives

$$f(\alpha) = 2B\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}} + 2(1-\alpha)A,$$

where

$$B = \sqrt{\frac{2\,\mathrm{VC}(\mathcal{H})\log(2(m+1)) + 2\log(\frac{8}{\delta})}{m}}$$

is a complexity term that is approximately equal to $\sqrt{\mathrm{VC}(\mathcal{H})/m}$ and

$$A = \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(S_u, T_u) + 4\sqrt{\frac{2\,\mathrm{VC}(\mathcal{H})\log(2m') + \log(\frac{8}{\delta})}{m'}} + \lambda$$

is the total divergence between the two domains.

It then follows that the optimal value α^* is a function of the number of target examples $m_T = \beta m$, the number of source examples $m_S = (1-\beta)m$, and the ratio $D = \sqrt{\text{VC}(\mathcal{H})}/A$:

$$\alpha^*(m_S, m_T, D) = \begin{cases} 1, & m_T \geq D^2 \\ \min(1, \nu), & m_T \leq D^2 \end{cases}$$

where

$$\nu = \frac{m_T}{m_T + m_S}\left(1 + \frac{m_S}{\sqrt{D^2(m_S + m_T) - m_S m_T}}\right).$$

As mentioned in [Ben-David et al., 2010a], this reformulation offers two interesting insights. First, if $m_T = 0$ ($\beta = 0$) then $\alpha^* = 0$, and if $m_S = 0$ (*i.e.*, $\beta = 1$) then $\alpha^* = 1$. As mentioned above, this implies that if we have only source or only target labeled data, the most appropriate choice is to use them for learning directly. Secondly, if the divergence between two domains is zero, then the optimal combination is to use the training data with uniform weighting of the examples. On the other hand, if there are enough target data, *i.e.*, $m_T \geq D^2 = \text{VC}(\mathcal{H})/A^2$, then no source data are required for efficient learning, and using it will be detrimental to the overall performance. This is because the possible error decrease as a result of using additional source data is always subject to its increase due to the increasing divergence between the source and target data. Secondly, for a few target examples, we might not have enough source data to justify its use. In this case, the sample of the source domain can be simply ignored. Finally, once we have enough source instances combined with a few target instances, α^* takes on intermediate values. This analysis is illustrated in Figure 5.

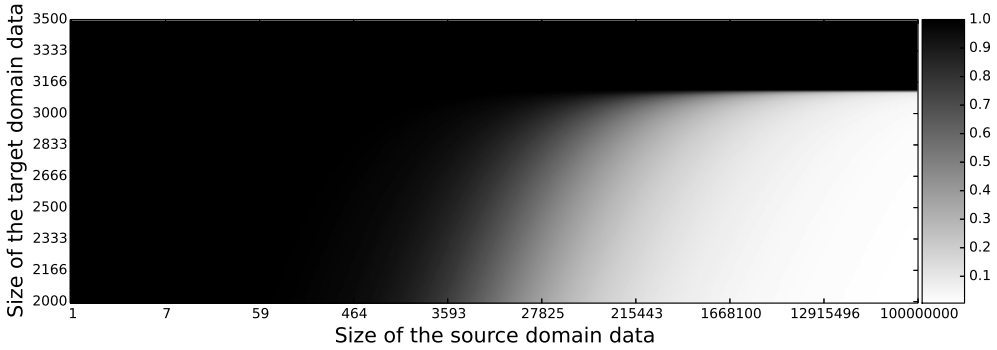

Figure 5: Illustration of the optimal value for α as a function of the number of source and target labeled instances.

3.3 Generalization bounds based on a discrepancy distance

One important limitation of the $\mathcal{H}\Delta\mathcal{H}$-divergence is its explicit dependence on a particular choice of a loss function, which is taken to be a $0-1$ loss. In general, however, it would be preferred to have generalization results for a more general domain adaptation setting, where any arbitrary loss function ℓ with some reasonable properties can be considered. In this section, we present a series of results that allow the first theoretical analysis of domain adaptation presented in the previous section to be extended to any arbitrary loss function. As we will show, the new divergence measure considered in this section is not restricted to be used exclusively for the task of binary classification, but can also be used for large families of regularized classifiers and regression. Moreover, the results in this section use the concept of the Rademacher complexity, as recalled in Section 2. This particular improvement will lead to data-dependent bounds that are usually tighter that the bounds obtained using the VC theory.

Discrepancy distance We start with the definition of the new divergence measure that was first introduced in [Mansour et al., 2009a]. As they mentioned, its name, the *discrepancy distance*, is due to the relationship between this concept and the discrepancy problems that arise in combinatorial contexts.

Definition 12 ([Mansour et al., 2009a])**.** *Given two domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X} \times Y$, let $\mathcal{H}$ be a hypothesis class, and let $\ell : Y \times Y \to \mathbb{R}_+$ define a loss function. The discrepancy distance $disc_\ell$ between the two marginals $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$ over $\mathbf{X}$ is defined by*

$$disc_\ell(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}) = \sup_{(h,h') \in \mathcal{H}^2} \left| \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{S}_\mathbf{X}} \left[\ell\left(h'(\mathbf{x}), h(\mathbf{x})\right)\right] - \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{T}_\mathbf{X}} \left[\ell\left(h'(\mathbf{x}), h(\mathbf{x})\right)\right] \right|.$$

We note that the $\mathcal{H}\Delta\mathcal{H}$-divergence and the discrepancy distance are related. First, for the $0-1$ loss, we have

$$disc_{\ell_{01}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \;=\; \frac{1}{2}\, d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})\,,$$

which shows that in this case the discrepancy distance coincides with the $\mathcal{H}\Delta\mathcal{H}$-divergence that appears in Theorems 10 and 11, and it suffers from the same computational restrictions as the latter. Furthermore, their tight connection is illustrated by the following proposition.

Proposition 12 ([Mansour et al., 2009a]). *Given two domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X} \times Y$, let $\mathcal{H}$ be a hypothesis class, and let $\ell : Y \times Y \to \mathbb{R}_+$ define a loss function that is bounded, $\forall (y, y') \in Y^2,\ \ell(y, y') \leq M$ for some $M > 0$. Then, for any hypothesis $h \in \mathcal{H}$, we have*

$$disc_{\ell}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \;\leq\; M\, d_1(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}).$$

This proposition establishes a link between the seminal results [Ben-David et al., 2010a] presented in the previous section, and shows that for a loss function bounded by M, the discrepancy distance can be upper-bounded in terms of the L^1-distance.

Learning bounds To present a generalization bound, we first need to understand how the discrepancy distance can be estimated from finite samples. To this end, [Mansour et al., 2009a] proposed the following lemma that bounds the discrepancy distance using the Rademacher complexity (see Section 2.4) of the hypothesis class.

Lemma 13 ([Mansour et al., 2009a]). *Let $\mathcal{H}$ be a hypothesis class, and let $\ell : Y \times Y \to \mathbb{R}_+$ define a loss function that is bounded, $\forall (y, y') \in Y^2,\ \ell(y, y') \leq M$ for some $M > 0$ and let $L_{\mathcal{H}} = \{\mathbf{x} \to \ell(h'(\mathbf{x}), h(\mathbf{x})) : h, h' \in \mathcal{H}\}$. Let $\mathcal{D}_{\mathbf{X}}$ be a distribution over $\mathbf{X}$, and let $\hat{\mathcal{D}}_{\mathbf{X}}$ denote the corresponding empirical distribution for a sample $S = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$. Then, for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over the choice of sample S, we have*

$$disc_{\ell}(\mathcal{D}_{\mathbf{X}}, \hat{\mathcal{D}}_{\mathbf{X}}) \leq \mathcal{R}_S(L_{\mathcal{H}}) + 3M\sqrt{\frac{\log \frac{2}{\delta}}{2m}},$$

where $\mathcal{R}_S(L_{\mathcal{H}})$ is the empirical Rademacher complexity of $L_{\mathcal{H}}$ based on the observations from S.

It can be noted that this lemma looks very much like the usual generalization inequalities obtained using the Rademacher complexities presented in Section 2.4. Using this result, we can further prove the following corollary for the case of more general loss functions defined as $\forall (y, y') \in Y^2,\ \ell_q(y, y') = |y - y'|^q$ for some q. This parametric family of functions is a common choice of a loss function for a regression task.

Corollary 14 ([Mansour et al., 2009a]). *Let $\mathcal{S}$ and $\mathcal{T}$ be the source and target domains over $\mathbf{X} \times Y$, respectively. Let $\mathcal{H}$ be a hypothesis class, and let $\ell_q : Y \times Y \to \mathbb{R}_+$ be a loss function that is bounded, $\forall (y, y') \in Y^2,\ \ell_q(y, y') \leq M$ for some $M > 0$, and defined as $\forall (y, y') \in Y^2,\ \ell_q(y, y') = |y - y'|^q$ for some q. Let S_u and T_u be samples of size m_s and m_t drawn independently from $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$ and let $\hat{\mathcal{S}}_{\mathbf{X}}, \hat{\mathcal{T}}_{\mathbf{X}}$ denote the empirical distributions corresponding to $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$. Then, for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over the random choice of the samples, we have*

$$disc_{\ell_q}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \;\leq\; disc_{\ell_q}(\hat{\mathcal{S}}_{\mathbf{X}}, \hat{\mathcal{T}}_{\mathbf{X}}) + 4q\left(\mathcal{R}_{S_u}(\mathcal{H}) + \mathcal{R}_{T_u}(\mathcal{H})\right) + 3M\left(\sqrt{\frac{\log(\frac{4}{\delta})}{2m_s}} + \sqrt{\frac{\log(\frac{4}{\delta})}{2m_t}}\right).$$

This result highlights one of the major differences between the approach of [Ben-David et al., 2010a] and that of [Mansour et al., 2009a], which arises from the way that they estimate the introduced distance. While Theorem 10 relies on the VC dimension to bound the true $\mathcal{H}\Delta\mathcal{H}$-divergence by its empirical counterpart, $disc_{\ell}$ is estimated using the quantities based on the Rademacher complexity. To illustrate what this implies for the generalization guarantees, we now present the analog of Theorem 10, which relates the source and target error functions using the discrepancy distance, and compare this to the original result.

Theorem 15 ([Mansour et al., 2009a]). *Let $\mathcal{S}$ and $\mathcal{T}$ be the source and target domains over $\mathbf{X} \times Y$, respectively. Let $\mathcal{H}$ be a hypothesis class, and let $\ell : Y \times Y \to \mathbb{R}_+$ be a loss function that is symmetric, obeys the triangle inequality, and is bounded, $\forall (y, y') \in Y^2,\ \ell(y, y') \leq M$ for some $M > 0$. Then, for $h^*_{\mathcal{S}} = \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{R}^{\ell}_{\mathcal{S}}(h)$ and $h^*_{\mathcal{T}} = \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{R}^{\ell}_{\mathcal{T}}(h)$ denoting the ideal hypotheses for the source and target domains, we have*

$$\forall h \in \mathcal{H},\ \mathrm{R}^{\ell}_{\mathcal{T}}(h) \;\leq\; \mathrm{R}^{\ell}_{\mathcal{S}}(h, h^*_S) + disc_{\ell}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \epsilon\,,$$

where $\mathrm{R}^{\ell}_{\mathcal{S}}(h, h^*_{\mathcal{S}}) = \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{S}_{\mathbf{X}}} \ell\left(h(\mathbf{x}), h^*_{\mathcal{S}}(\mathbf{x})\right)$ and $\epsilon \;=\; \mathrm{R}^{\ell}_{\mathcal{T}}(h^*_T) + \mathrm{R}^{\ell}_{\mathcal{S}}(h^*_T, h^*_{\mathcal{S}})$.

Comparison with the $\mathcal{H}\Delta\mathcal{H}$-divergence As pointed out by the authors, this bound is not directly comparable to Theorem 10, but involves similar terms and reflects a very common trade-off between them. Indeed, the first term of this bound stands for the same source risk function as that in the work of [Ben-David et al., 2010a]. The second term here captures the deviation between the two domains through the discrepancy distance similar to the $\mathcal{H}\Delta\mathcal{H}$-divergence used before. Finally, the last term ϵ can be interpreted as the capacity to adapt, and it is very close in spirit to the λ term seen previously.

Despite these similarities, the closer comparison made by [Mansour et al., 2009a] revealed that the bound based on the discrepancy distance can be tighter in some plausible scenarios. For instance, in a degenerate case where there is only one hypothesis $h \in \mathcal{H}$ and a single target function $f_{\mathcal{T}}$, the bounds of Theorem 15 and of Theorem 10 with true distributions give $\mathrm{R}^{\ell}_{\mathcal{T}}(h, f) + disc_{\ell}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ and $\mathrm{R}^{\ell}_{\mathcal{T}}(h, f) + 2\mathrm{R}^{\ell}_{\mathcal{S}}(h, f) + disc_{\ell}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$, respectively. In this case, the latter expression is obviously larger when $\mathrm{R}^{\ell}_{\mathcal{S}}(h, f) \leq \mathrm{R}^{\ell}_{\mathcal{T}}(h, f)$. The same kind of result can also be shown to hold under the following plausible assumptions:

1. When $h^* = h^*_{\mathcal{S}} = h^*_{\mathcal{T}}$, the bounds of Theorems 15 and 10 respectively boil down to
$$\mathrm{R}^{\ell}_{\mathcal{T}}(h) \leq \mathrm{R}^{\ell}_{\mathcal{T}}(h^*) + \mathrm{R}^{\ell}_{\mathcal{S}}(h, h^*) + disc_{\ell}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})\,, \tag{5}$$
and
$$\mathrm{R}^{\ell}_{\mathcal{T}}(h) \leq \mathrm{R}^{\ell}_{\mathcal{T}}(h^*) + \mathrm{R}^{\ell}_{\mathcal{S}}(h^*) + \mathrm{R}^{\ell}_{\mathcal{S}}(h) + disc_{\ell}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})\,, \tag{6}$$
where the right-hand side of Equation 6 includes the sum of three errors and is always larger than the right-hand side of Equation 5, due to the triangle inequality.
2. When $h^* = h^*_{\mathcal{S}} = h^*_{\mathcal{T}}$ and $disc_{\ell}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) = 0$, Theorems 15 and 10 give
$$\mathrm{R}^{\ell}_{\mathcal{T}}(h) \leq \mathrm{R}^{\ell}_{\mathcal{T}}(h^*) + \mathrm{R}^{\ell}_{\mathcal{S}}(h, h^*) \qquad \text{and} \qquad \mathrm{R}^{\ell}_{\mathcal{T}}(h) \leq \mathrm{R}^{\ell}_{\mathcal{T}}(h^*) + \mathrm{R}^{\ell}_{\mathcal{S}}(h^*) + \mathrm{R}^{\ell}_{\mathcal{S}}(h)\,,$$
where the former coincides with the standard generalization bound, while the latter does not.
3. Finally, when $f_{\mathcal{T}} \in \mathcal{H}$, Theorem 10 simplifies to
$$|\mathrm{R}^{\ell}_{\mathcal{T}}(h) - \mathrm{R}^{\ell}_{\mathcal{S}}(h)| \leq disc_{\ell_{01}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})\,,$$
which can be straightforwardly obtained from Theorem 15.

All of these results show a tight link that can be observed in different contributions of the domain adaptation theory. This relation illustrates that the results of [Mansour et al., 2009a] strengthen the previous contributions on the subject, but retain a tight connection to them.

3.4 Generalization bounds based on the discrepancy distance for regression

As mentioned at the beginning of this section, the discrepancy distance not only extends the first theoretical results obtained for domain adaptation, but also allows new point-wise guarantees to be derived for other learning scenarios, such as, for instance, the regression task, where contrary to classification, the output variable Y is continuous. The domain adaptation problem for regression is illustrated in Figure 6.

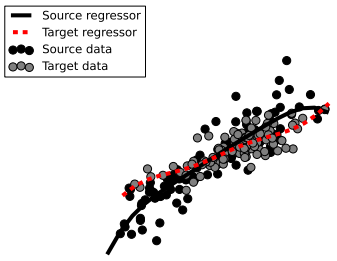

Figure 6: Domain adaptation problem for a regression task.

To address this scenario, another type of theoretical result based on the discrepancy distance was proposed by [Cortes and Mohri, 2011]. These authors considered the case where the hypothesis set $\mathcal{H}$ as a subset of the reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ associated to a positive definite symmetric kernel $K : \mathcal{H} = \{h \in \mathbb{H} : \|h\|_K \leq \Lambda\}$, where $\|\cdot\|_K$ denotes the norm defined by the inner product on $\mathbb{H}$ and $\Lambda \geq 0$. We shall assume that there exists $R > 0$ such that $K(\mathbf{x}, \mathbf{x}) \leq R^2$ for all $\mathbf{x} \in \mathbf{X}$. By the reproducing property, for any $h \in \mathcal{H}$ and $\mathbf{x} \in \mathbf{X}$, $h(\mathbf{x}) = \langle h, K(\mathbf{x}, \cdot)\rangle_K$, and thus this implies that $|h(\mathbf{x})| \leq \|h\|_K \sqrt{K(\mathbf{x}, \mathbf{x})} \leq \Lambda R$.

In this setting, the authors further presented point-wise loss guarantees in domain adaptation for a broad class of kernel-based regularization algorithms. Given a learning sample S, where $\forall (\mathbf{x}, y) \in S, \mathbf{x} \sim \mathcal{D}_{\mathbf{X}},\ y = f_{\mathcal{D}}(\mathbf{x})$, these algorithms are defined by the minimization of the following objective function:
$$F_{\hat{\mathcal{D}}_{\mathbf{X}}}(h) = \mathrm{R}^{\ell}_{\hat{\mathcal{D}}_{\mathbf{X}}}(h, f_{\mathcal{D}}) + \beta \|h\|^2_K,$$
where $\beta > 0$ is a trade-off parameter. This family of algorithms includes support vector machines, support vector regression [Vapnik, 1995], kernel ridge regression (KRR) [Saunders et al., 1998], and many other methods. Finally, the loss function ℓ is also assumed to be μ-admissible following the definition given below.

Definition 13 (μ-admissible loss)**.** *A loss function $\ell : Y \times Y \to \mathbb{R}$ is μ-admissible if it is symmetric and convex with respect to both of its arguments, and for all $\mathbf{x} \in \mathbf{X}$ and $y \in Y$ and $(h, h') \in \mathcal{H}^2$, it verifies the following Lipschitz condition for some $\mu > 0$:*

$$|\ell(h'(\mathbf{x}), y) - \ell(h(\mathbf{x}), y)| \leq \mu |h'(\mathbf{x}) - h(\mathbf{x})|.$$

The family of μ-admissible losses includes the hinge loss and all $\ell_q(y, y') = |y - y'|^q$ with $q \geq 1$, in particular the squared loss, when the hypothesis set and the set of output labels are bounded.

With the assumptions made previously, the following results can be proven.

Theorem 16 ([Cortes and Mohri, 2011, Cortes and Mohri, 2014])**.** *Let $\mathcal{S}$ and $\mathcal{T}$ be the source and target domains on $\mathbf{X} \times Y$, let $\mathcal{H}$ be a hypothesis class, and let ℓ be a μ-admissible loss. We assume that the target labeling function $f_{\mathcal{T}}$ belongs to $\mathcal{H}$, and let η denote $\max\{\ell(f_{\mathcal{S}}(\mathbf{x}), f_{\mathcal{T}}(\mathbf{x})) : \mathbf{x} \in \text{SUPP}(\hat{\mathcal{S}}_{\mathbf{X}})\}$. Let h' be the hypothesis that minimizes $F_{\hat{\mathcal{T}}_{\mathbf{X}}}$ and h the one returned when $F_{\hat{\mathcal{S}}_{\mathbf{X}}}$ is minimized. Then, for all $(\mathbf{x}, y) \in \mathbf{X} \times Y$, we have*

$$|\ell(h'(\mathbf{x}), y) - \ell(h(\mathbf{x}), y)| \leq \mu\, R \sqrt{\frac{disc_{\ell}(\hat{\mathcal{S}}_{\mathbf{X}}, \hat{\mathcal{T}}_{\mathbf{X}}) + \mu\eta}{\beta}}.$$

This theorem shows that the difference between the errors achieved by the optimal hypotheses learned on the source and target samples is proportional to the distance between the samples plus a term that reflects the worst value that a loss function can achieve for some instance that belongs to the support of $\hat{\mathcal{S}}_{\mathbf{X}}$.

A similar theorem can be proven when $f_{\mathcal{S}} \in \mathcal{H}$ and not $f_{\mathcal{T}}$ is assumed. Moreover, the authors indicated that these theorems can be extended to the case where neither the target function $f_{\mathcal{T}}$ nor $f_{\mathcal{S}}$ belong to $\mathcal{H}$, by replacing η in the statement of the theorem with

$$\eta' = \max_{\mathbf{x} \in \text{SUPP}(\hat{\mathcal{S}}_{\mathbf{X}})} \{\ell(h^*_{\mathcal{T}}(\mathbf{x}), f_{\mathcal{S}}(\mathbf{x}))\} + \max_{\mathbf{x} \in \text{SUPP}(\hat{\mathcal{T}}_{\mathbf{X}})} \{\ell(h^*_{\mathcal{T}}(\mathbf{x}), f_{\mathcal{T}}(\mathbf{x}))\},$$

where $h^*_{\mathcal{T}} \in \operatorname*{argmin}_{h \in \mathcal{H}} \ell(h(\mathbf{x}), f_{\mathcal{T}})$. In both cases, when η is assumed to be small, *i.e.* $\eta \ll 1$, the key term of the obtained bound is the empirical discrepancy distance $disc_{\ell}(\hat{\mathcal{S}}_{\mathbf{X}}, \hat{\mathcal{T}}_{\mathbf{X}})$. In the extreme case when $f_{\mathcal{T}} = f_{\mathcal{S}} = f$, we obtain $\eta = 0$, and the problem reduces to the covariate shift adaptation scenario that is characterized by the same labeling function in both domains, and is analyzed in more in detail in the following section. In general, a parallel can be drawn between the η term that appears in this bound and the other so-called adaptation capacity terms, such as the λ term in the bound of Ben-David *et* al. from Theorem 10.

The result given by Theorem 16 can be further strengthened when the considered loss function is assumed to be the squared loss $\ell_2 = (y - y')^2$ for some $(y, y') \in Y^2$, and when the kernel-based regularization algorithm described above coincides with the KRR. In what follows, the term η will be replaced by a finer quantity defined as

$$\delta_{\mathcal{H}}(f_{\mathcal{S}}, f_{\mathcal{T}}) = \inf_{h \in \mathcal{H}} \| \mathop{\mathbf{E}}_{\mathbf{x} \sim \hat{\mathcal{S}}_{\mathbf{X}}} [\Delta(h, f_{\mathcal{S}})] - \mathop{\mathbf{E}}_{\mathbf{x} \sim \hat{\mathcal{T}}_{\mathbf{X}}} [\Delta(h, f_{\mathcal{T}})] \|,$$

where $\Delta(h, f) = \big(f(\mathbf{x}) - h(\mathbf{x})\big)\Phi(\mathbf{x})$ with $\Phi(\mathbf{x})$ is associated to the kernel K feature vector, such that $K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$. Using this quantity, the following guarantee holds.

Theorem 17 ([Cortes and Mohri, 2014])**.** *Let ℓ be a squared loss bounded by some $M > 0$, and let h' be the hypothesis that minimizes $F_{\hat{\mathcal{T}}_{\mathbf{X}}}$, and h the one returned when $F_{\hat{\mathcal{S}}_{\mathbf{X}}}$ is minimized. Then, for all $(\mathbf{x}, y) \in \mathbf{X} \times Y$, we have:*

$$|\ell(h(\mathbf{x}), y) - \ell(h'(\mathbf{x}), y)| \leq \frac{R\sqrt{M}}{\beta} \left(\delta_{\mathcal{H}}(f_{\mathcal{S}}, f_{\mathcal{T}}) + \sqrt{\delta^2_{\mathcal{H}}(f_{\mathcal{S}}, f_{\mathcal{T}}) + 4\,\beta\, disc_{\ell}(\hat{\mathcal{S}}_{\mathbf{X}}, \hat{\mathcal{T}}_{\mathbf{X}})} \right).$$

As indicated by the authors, the main advantage of this result is its expression in terms of $\delta_{\mathcal{H}}(f_{\mathcal{S}}, f_{\mathcal{T}})$ instead of $\eta_{\mathcal{H}}(f_{\mathcal{S}}, f_{\mathcal{T}})$. It can be noted that $\delta_{\mathcal{H}}(f_{\mathcal{S}}, f_{\mathcal{T}})$ is defined as a difference, and thus it becomes zero for $\mathcal{S}_{\mathbf{X}} = \mathcal{T}_{\mathbf{X}}$, which does not hold for $\eta_{\mathcal{H}}(f_{\mathcal{S}}, f_{\mathcal{T}})$. Furthermore, when the covariate-shift assumption holds for some shared labeling function f such that $f_{\mathcal{S}} = f_{\mathcal{T}} = f$, $\delta_{\mathcal{H}}(f, f)$ can be upper-bounded using the following result.

Theorem 18 ([Cortes and Mohri, 2014])**.** *Assume that for all $\mathbf{x} \in \mathbf{X}$, $K(\mathbf{x}, \mathbf{x}) \leq R^2$ for some $R > 0$. Let $\mathcal{A}$ denote the union of the supports of $\hat{\mathcal{S}}_{\mathbf{X}}$ and $\hat{\mathcal{T}}_{\mathbf{X}}$. Then, for any $p > 1$ and $q > 1$, with $1/p + 1/q = 1$,*

$$\delta_{\mathcal{H}}(f, f) \leq d_p(f_{|\mathcal{A}}, \mathcal{H}_{|\mathcal{A}}) \ell_q(\hat{\mathcal{S}}_{\mathbf{X}}, \hat{\mathcal{T}}_{\mathbf{X}}),$$

where for any set $\mathcal{A} \subseteq \mathbf{X}$, $f_{|\mathcal{A}}$ (resp. $\mathcal{H}_{|\mathcal{A}}$) denote the restriction of f (resp. h) to $\mathcal{A}$ and $d_p(f_{|\mathcal{A}}, \mathcal{H}_{|\mathcal{A}}) = \inf_{h \in \mathcal{H}} \|f - h\|_p$.

In particular, the authors show that for a labeling function f that belongs to the closure of $\mathcal{H}_{|\mathcal{A}}$, $\delta_{\mathcal{H}}(f) = 0$ when the KRR algorithm is used with normalized Gaussian kernels. For this specific algorithm that is often used in practice, the bound of the theorem then reduces to the simpler expression:

$$|\ell(h(\mathbf{x}), y) - \ell(h'(\mathbf{x}), y)| \;\leq\; 2\,R\sqrt{\frac{Mdisc_{\ell}(\hat{\mathcal{S}}_{\mathbf{X}}, \hat{\mathcal{T}}_{\mathbf{X}})}{\beta}}.$$

Generalized discrepancy The above-mentioned results can be further strengthened using a recently proposed notation of the generalized discrepancy introduced by [Cortes et al., 2015]. To introduce this distance, we can first note that a regression task in the domain adaptation context can be seen as an optimal approximation of an ideal hypothesis $h^*_{\mathcal{T}} = \operatorname*{argmin}_{h\in\mathcal{H}} \mathrm{R}^{\ell}_{\mathcal{T}}(h, f_{\mathcal{T}})$ by another hypothesis h that ensures the closeness of the losses $\mathrm{R}^{\ell}_{\mathcal{T}}(h^*, f_{\mathcal{T}})$ and $\mathrm{R}^{\ell}_{\mathcal{T}}(h, f_{\mathcal{T}})$. As we do not have access to $f_{\mathcal{T}}$, but only to the labels of the source sample S, the main idea is to define for any $h \in \mathcal{H}$, a reweighting function $Q_h : S \to \mathbb{R}$ such that the objective function G that is defined for all $h \in \mathcal{H}$ by

$$G(h) = \mathrm{R}^{\ell}_{Q_h}(h) + \beta\|h\|^2_K,$$

remains uniformly close to $F_{\hat{\mathcal{T}}_{\mathbf{X}}}(h)$ defined over the target sample T_u. As indicated by the authors, this idea introduces a different learning concept, as instead of reweighting the training sample with some fixed set of weights, the weights are allowed to vary as a function of the hypothesis h, and are not assumed to sum to 1 or to be nonnegative. Based on this construction, the optimal reweighting can be obtained by solving:

$$Q_h = \operatorname*{argmin}_{q\in\mathcal{F}(\mathcal{S}_{\mathbf{X}},\mathbb{R})} |\mathrm{R}^{\ell}_{\hat{\mathcal{T}}_{\mathbf{X}}}(h, f_{\mathcal{T}}) - \mathrm{R}^{\ell}_{q}(h, f_{\mathcal{S}})|,$$

where $\mathcal{F}(\mathcal{S}_{\mathbf{X}}, \mathbb{R})$ is the set of real-valued functions defined over $\textsc{supp}(\mathcal{S}_{\mathbf{X}})$.

We can note that, in practice, we might not have access to labeled target instances, which implies that we cannot estimate $f_{\mathcal{T}}$. To solve this problem, the authors proposed to consider a nonempty convex set of candidate hypotheses $\mathcal{H}'' \subseteq \mathcal{H}$ that can contain a good approximation of $f_{\mathcal{T}}$. Using $\mathcal{H}''$ as a set of surrogate labeling functions, the previous optimization problem becomes:

$$Q_h = \operatorname*{argmin}_{q\in\mathcal{F}(\mathcal{S}_{\mathbf{X}},\mathbb{R})} \max_{h''\in\mathcal{H}} |\mathrm{R}^{\ell}_{\hat{\mathcal{T}}_{\mathbf{X}}}(h, h'') - \mathrm{R}^{\ell}_{q}(h, f_{\mathcal{S}})|.$$

The risk obtained using the solution of this optimization problem given by Q_h can be equivalently expressed as follows:

$$\mathrm{R}^{\ell}_{Q_h}(h, f_{\mathcal{S}}) = \frac{1}{2}\left(\max_{h''\in\mathcal{H}} \mathrm{R}^{\ell}_{\hat{\mathcal{T}}_{\mathbf{X}}}(h, h'') + \min_{h''\in\mathcal{H}} \mathrm{R}^{\ell}_{\hat{\mathcal{T}}_{\mathbf{X}}}(h, h'')\right).$$

This, in its turn, allows us to reformulate $G(h)$, which can now become:

$$G(h) = \frac{1}{2}\left(\max_{h''\in\mathcal{H}} \mathrm{R}^{\ell}_{\hat{\mathcal{T}}_{\mathbf{X}}}(h, h'') + \min_{h''\in\mathcal{H}} \mathrm{R}^{\ell}_{\hat{\mathcal{T}}_{\mathbf{X}}}(h, h'')\right) + \beta\|h\|^2_K.$$

The proposed optimization problem should have the same point-wise guarantees as those established in Theorem 17, but based on a new notation of the distance between the probability distributions that can be seen as a generalization of the discrepancy distance used before. To introduce this, we now define $A(\mathcal{H})$ as a set of functions $U : h \to U_h$ that map $\mathcal{H}$ to $\mathcal{F}(\mathcal{S}_{\mathbf{X}}, \mathbb{R})$, such that for all $h \in \mathcal{H}$, $h \to \ell_{U_h}(h, f_{\mathcal{S}})$ is a convex function. The set $A(\mathcal{H})$ contains all of the constant functions U such that $U_h = q$ for all $h \in \mathcal{H}$, where q is a distribution over $\mathcal{S}_{\mathbf{X}}$. The definition of the generalized discrepancy can thus be given as follows.

Definition 14. *For any $U \in A(\mathcal{H})$, the generalized discrepancy between U and $\hat{\mathcal{T}}_{\mathbf{X}}$ is defined as*

$$\mathit{DISC}(\hat{\mathcal{T}}_{\mathbf{X}}, U) = \max_{h\in\mathcal{H},h''\in\mathcal{H}''} \left|\mathrm{R}^{\ell}_{\hat{\mathcal{T}}_{\mathbf{X}}}(h, h'') - \mathrm{R}^{\ell}_{U_h}(h, f_{\mathcal{S}})\right|.$$

In addition, the authors defined the following distance of f to $\mathcal{H}''$ over the support of $\hat{\mathcal{T}}_{\mathbf{X}}$:

$$d^{\hat{\mathcal{T}}_{\mathbf{X}}}_{\infty}(f_{\mathcal{T}}, \mathcal{H}'') = \min_{h_0\in\mathcal{H}''} \max_{\mathbf{x}\in\textsc{supp}(\hat{\mathcal{T}}_{\mathbf{X}})} |h_0(\mathbf{x}) - f_{\mathcal{T}}(\mathbf{x})|.$$

Using the above-defined quantities, the following point-wise guarantees can be given.

Theorem 19 ([Cortes et al., 2015]). *Let h^* be a minimizer of $\mathrm{R}^{\ell}_{\hat{\mathcal{T}}_{\mathbf{X}}}(h, f_{\mathcal{T}}) + \beta\|h\|^2_K$, and h_Q be a minimizer of $\mathrm{R}^{\ell}_{Q_h}(h, f_{\mathcal{S}}) + \beta\|h\|^2_K$. Then, for $Q : h \to Q_h$ and $\forall \mathbf{x} \in \mathbf{X},\ y \in Y$, the following holds*

$$|\ell(h_Q(\mathbf{x}), y) - \ell(h^*(\mathbf{x}), y)| \leq \mu R \sqrt{\frac{\mu d_{\infty}^{\hat{\mathcal{T}}_{\mathbf{X}}}(f_{\mathcal{T}}, \mathcal{H}'') + \mathit{DISC}(Q, \hat{\mathcal{T}}_{\mathbf{X}})}{\beta}}.$$

Furthermore, this inequality can be equivalently written in terms of the risk functions as

$$\mathrm{R}^{\ell}_{\mathcal{T}}(h_{\mathrm{Q}}, f_{\mathcal{T}}) \leq \mathrm{R}^{\ell}_{\mathcal{T}}(h^*, f_{\mathcal{T}}) + \mu R \sqrt{\frac{\mu d_{\infty}^{\hat{\mathcal{T}}_{\mathbf{X}}}(f_{\mathcal{T}}, \mathcal{H}'') + \mathrm{DISC}(\mathrm{Q}, \hat{\mathcal{T}}_{\mathbf{X}})}{\beta}}.$$

The result of Theorem 19 suggests the selection of $\mathcal{H}''$ to minimize the right-hand side of the last inequality. In particular, the authors provided further evidence that if the space over which $\mathcal{H}''$ is searched is the family of all of the balls centered in $f_{\mathcal{S}}$ defined in terms of l_{q^*}, *i.e.*, $\mathcal{H}'' = \{h'' \in \mathcal{H} | l_q(h'', f_Q) \leq r\}$ for some distribution q over the space of the reweighted source samples, then the proposed algorithm based on the generalized discrepancy gives demonstrably better results compared to the original algorithm.

Semi-supervised case When labeled sample T from the target domain is available, part of it can actually be used to find an appropriate value of r. To support this statement, let us consider the following set $S' = S \cup T$ and an empirical distribution $\hat{\mathcal{S}}'_{\mathbf{X}}$ over it, and use q'^* to denote the distribution that minimizes the discrepancy between $\hat{\mathcal{S}}'_{\mathbf{X}}$ and $\hat{\mathcal{T}}_{\mathbf{X}}$. Now, as $\textsc{supp}(\hat{\mathcal{S}}_{\mathbf{X}})$ is included in that of $\textsc{supp}(\hat{\mathcal{S}}'_{\mathbf{X}})$, the following inequality can be obtained

$$\begin{aligned} \mathit{disc}_{\ell}(\hat{\mathcal{T}}_{\mathbf{X}}, q'^*) &= \min_{\textsc{supp}(q) \subseteq \textsc{supp}(\hat{\mathcal{S}}'_{\mathbf{X}})} \mathit{disc}_{\ell}(\hat{\mathcal{T}}_{\mathbf{X}}, q) \\ &\leq \min_{\textsc{supp}(q) \subseteq \textsc{supp}(\hat{\mathcal{S}}_{\mathbf{X}})} \mathit{disc}_{\ell}(\hat{\mathcal{T}}_{\mathbf{X}}, q) = \mathit{disc}_{\ell}(\hat{\mathcal{T}}_{\mathbf{X}}, q^*). \end{aligned}$$

Consequently, in view of Theorem 19, for an appropriate choice of $\mathcal{H}''$, the learning guarantee for adaptation algorithms based on the generalized discrepancy is more favorable when some labeled data from the target domain are used. Thus, use of the limited amount of labeled points from the target distribution can improve the performance of their proposed algorithm.

3.5 Other relevant contributions

[Mansour et al., 2008] In this paper, the authors considered the multi-source domain adaptation problem, and introduced the learning bounds in two different adaptation settings. For the first one, they assumed that $\mathcal{T}_{\mathbf{X}} = \sum_{i=1}^{N} \alpha_i \mathcal{S}^i_{\mathbf{X}}$, and studied the performance of a hypothesis defined as $h_{\alpha} = \sum_{i=1}^{N} \alpha_i h_i$, where $\mathcal{S}^i_{\mathbf{X}}$ is the marginal distributions of the i^{th} source domain, and $\forall i, \alpha_i \geq 0,\ \sum_{i=1}^{N} \alpha_i = 1$. In this scenario, the authors proved that there exists a domain adaptation problem such that $\mathrm{R}_{\mathcal{T}}(h_{\alpha}) = \frac{1}{2}$ even when $\forall i,\ \mathrm{R}_{\mathcal{S}^i_{\mathbf{X}}}(h_i) = 0$. This prompted them to consider a different combined hypothesis defined as

$$h^{\mathcal{D}}_{\alpha} = \sum_{i=1}^{N} \frac{\alpha_i \mathcal{S}^i_{\mathbf{X}}}{\sum_{i=1}^{N} \alpha_i \mathcal{S}^i_{\mathbf{X}}} h_i.$$

In this case, the authors proved that $\mathrm{R}_{\mathcal{T}}(h^{\mathcal{D}}_{\alpha}) \leq \varepsilon$ when $\forall i,\ \mathrm{R}_{\mathcal{S}^i_{\mathbf{X}}}(h_i) \leq \varepsilon$.

[Mansour et al., 2009b] This work extends the contribution of [Mansour et al., 2008] by analyzing arbitrary target distributions that are not necessarily represented by a weighted mixture of source distributions. The authors proposed domain adaptation learning bounds of the following form:

$$\mathrm{R}_{\mathcal{T}}(h^{\mathcal{D}}_{\alpha}) \leq (\varepsilon d_{\alpha}(\mathcal{T}_{\mathbf{X}} | \mathcal{S}_{\mathbf{X}}))^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}},$$

where $d_{\alpha}(\mathcal{T}_{\mathbf{X}} | \mathcal{S}_{\mathbf{X}}) = \left(\int_{\mathbf{X}} \frac{\mathcal{T}^{\alpha}_{\mathbf{X}}}{\mathcal{S}^{(\alpha-1)}_{\mathbf{X}}} \right)^{\frac{1}{\alpha-1}}$ is the exponential of the α-Rényi divergence, $\mathrm{R}_{\mathcal{S}^i_{\mathbf{X}}}(h_i) \leq \varepsilon$, and $M \geq 0$ is a constant that bounds the loss function used in the definition of $\mathrm{R}_{\mathcal{D}}$.

[Hoffman et al., 2018] In this work, the authors extend the analysis of [Mansour et al., 2009b] to account for cross-entropy and other similar losses not considered in previous work. They also propose a principal way of determining the coefficients α_i ensuring efficient adaptation and extend their analysis to the scenario of non-deterministic labeling.

[Dhouib and Redko, 2018] In this work, the authors proposed a learning bound for hypotheses associated to a general family of similarity functions introduced in [Balcan et al., 2008]. The proposed bounds rely on L^1 and χ^2 divergences and similar to [Mansour et al., 2009b] present a multiplicative dependence of the source error on the divergence term.

[Redko et al., 2019a] Finally, in this work the authors introduced a bound for the multi-source domain adaptation based on the discrepancy of [Mansour et al., 2009a] for the target shift scenario where the inequality between $\mathcal{S}$ and $\mathcal{T}$ is due to the drift between the marginal distributions of Y in each domain.

[Kuroki et al., 2019] This paper proposes source-guided discrepancy (S-disc) that has a virtue of being much easier to estimate in case of ℓ_{01} than the discrepancy proposed by [Mansour et al., 2009a]. The authors also derive a generalization error bound based on S-disc and show that it is never looser than the original bound proposed by [Mansour et al., 2009a].

3.6 Summary

This section presents several cornerstone results of the domain adaptation theory, including those proposed by Ben-David *et al.* based on the $\mathcal{H}\Delta\mathcal{H}$-divergence, and a variety of results based on the discrepancy distance proposed by Mansour *et al.* and Cortes *et al.* for the tasks of classification and regression. As can be noted, the general ideas used to prove generalization bounds for domain adaptation are based on the definition of a relation between the source and target domains through a divergence that allows us to upper-bound the target risk by the source risk, and on the theoretical results presented in Section 2, and their properties. Unsurprisingly, this trend is usually maintained regardless of the considered domain adaptation scenario or the learning algorithm analyzed. The overall form of the presented generalization bound on the error of a hypothesis calculated with respect to the target distribution appears to contain, inevitably, the following important terms:

1. The source error of the hypothesis measured with respect to some loss function;
2. The divergence term between the marginal distributions of the source and target domains. In the case of Ben-David *et al.*, this term is explicitly linked to the hypothesis space that induces a complexity term that is related to its Vapnik-Chervonenkis dimension; in the case of Mansour *et al.* and Cortes *et al.*, the divergence term depends on the hypothesis space, but the complexity term is data dependent and is linked to the Rademacher complexity of the hypothesis space;
3. The nonestimable term that reflects the *a-priori* hardness of the domain adaptation problem. This last usually requires at least some target labeled data to be quantified.

The terms that appear in the bounds show us that in the case where two domains are almost indistinguishable, the performance of a given hypothesis across these will remain largely similar. When this is not the case, the divergence between the source and target domain marginal distributions starts to have a crucial role in the assessment of the proximity of two domains. For both of the set of results presented, the actual value of this divergence can be consistently calculated using the available finite (unlabeled) samples, thus providing us with a first estimate of the potential success of adaptation. Finally, the last term tells us that even when the divergence between the marginal distributions is taken to zero across two domains, this might not suffice for efficient adaptation. This last point can be summarized by the following statement, as made by Ben-David in [Ben-David et al., 2010a]:

> "When the combined error of the ideal joint hypothesis is large, then there is no classifier that performs well on both the source and target domains, so we cannot hope to find a good target hypothesis by training only on the source domain."

This statement brings us to another important question regarding the conditions that need to be verified to make sure that the adaptation is successful. This question stimulates a cascade of other relevant questions, such as what is the actual size of the source and target unlabeled samples needed for the adaptation to be efficient? Are target labeled data needed for an efficient adaptation, and if yes, can we prove formally that it leads to better results? And finally, what are the pitfalls of domain adaptation when even strong prior knowledge regarding the adaptation problem does not guarantee that it has a solution? All these question are answered by the so-called "hardness theorems" that we present in the following section.

4 Hardness results for domain adaptation

This section is devoted to a series of results that prove the so-called "hardness or impossibility theorems" for domain adaptation. These latter statements show the extent to which the domain adaptation problem can be hard to solve, or the conditions when it is provably unsolvable under some common assumptions. These theorems are very important, as they highlight that in some cases it will not be possible to adapt well even with a prohibitively large amount of data from both domains, or when the adaptation task might be trivial.

4.1 Problem set-up

Before presenting the main theoretical results, we first introduce the necessary preliminary definitions that formalize the concepts used afterwards. These definitions are then followed by a set of assumptions that are commonly considered to have a direct influence on the potential success of domain adaptation.

Definitions We have seen from the previous sections that the adaptation efficiency is directly correlated with two main terms that inevitably appear in almost all analyses: one term that depicts the divergence between the domains, and the other term that stands for the existence and the error achieved by the best hypothesis across the source and target domains. The authors of [Ben-David et al., 2010b] proposed to analyze the presence of these two terms in the bounds by answering the following questions:

1. Is the presence of these two terms inevitable in the domain adaptation bounds?
2. Is there a way to design a more intelligent domain adaptation algorithm that uses not only the labeled training sample, but also the unlabeled sample of the target data distribution?

These two questions are very important, as answering them can help us to obtain an exhaustive set of conditions that theoretically ensure efficient adaptation with respect to a given domain adaptation algorithm. Before proceeding to the presentation of the main results, the authors first defined several quantities that they used later. The first one is the formalization of an unsupervised domain adaptation algorithm [Ben-David et al., 2010b].

Definition 15 (domain adaptation learner)**.** *A domain adaptation learner is a function*

$$\mathcal{A} : \bigcup_{m=1}^{\infty} \bigcup_{n=1}^{\infty} (\mathbf{X} \times \{0,1\})^m \times \mathbf{X}^n \to \{0,1\}^{\mathbf{X}}.$$

As before, the standard notation for the performance of the learner is given by the error function used. When the error is measured with respect to the best hypothesis in some hypothesis class $\mathcal{H}$, we use the notation $\mathrm{R}_{\mathcal{D}}(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathrm{R}_{\mathcal{D}}(h)$. Using this notation, the authors further defined the learnability, as follows.

Definition 16 ($(\varepsilon, \delta, m, n)$-learnability)**.** *Let $\mathcal{S}$ and $\mathcal{T}$ be distributions over $\mathbf{X} \times \{0,1\}$, $\mathcal{H}$ a hypothesis class, $\mathcal{A}$ a domain adaptation learner, $\varepsilon > 0$, $\delta > 0$, and m, n positive integers. We say that $\mathcal{A}(\varepsilon, \delta, m, n)$-learns $\mathcal{T}$ from $\mathcal{S}$ relative to $\mathcal{H}$, if when given access to a labeled sample S of size m, generated i.i.d.by $\mathcal{S}$, and an unlabeled sample T_u of size n, generated i.i.d. by $\mathcal{T}_{\mathbf{X}}$, with probability of at least $1 - \delta$ (over the choice of the samples S and T_u), the learned classifier does not exceed the error of the best classifier in $\mathcal{H}$ by more than ε, i.e.,*

$$\Pr_{\substack{S \sim (\mathcal{S})^m \\ T_u \sim (\mathcal{T}_{\mathbf{X}})^n}} \left[\mathrm{R}_{\mathcal{T}}(\mathcal{A}(S, T_u)) \leq \mathrm{R}_{\mathcal{T}}(\mathcal{H}) + \varepsilon \right] \geq 1 - \delta.$$

This definition gives us a criterion that we can use to judge whether a particular algorithm has strong learning guarantees, which consists in finding an optimal trade-off between both ε and δ in the above definition. We further introduce two alternative definitions of domain adaptation learnability for the proper learning setting and when the best error of a classifier in $\mathcal{H}$ is scaled by an additional constant c.

Definition 17 ($(c, \varepsilon, \delta, m, n)$-proper learnability)**.** *With the notations from Definition 16, we say that $A(c, \varepsilon, \delta, m, n)$-solves a proper domain adaptation for the class $\mathcal{W}$ relative to $\mathcal{H}$, if $\mathcal{A}$ outputs an element h of $\mathcal{H}$ with*

$$\Pr_{\substack{S \sim (\mathcal{S})^m \\ T_u \sim (\mathcal{T}_{\mathbf{X}})^n}} \left[\mathrm{R}_{\mathcal{T}}(\mathcal{A}(S, T_u)) \leq c\mathrm{R}_{\mathcal{T}}(\mathcal{H}) + \varepsilon \right] \geq 1 - \delta.$$

In other words, this definition says that the proper solving of the domain adaptation problem is achieved when the error of the returned hypothesis *from a fixed hypothesis class w.r.t.* the target distribution is bounded by c times the error of the best hypothesis on the target distribution plus a constant ε. Obviously, efficient solving of the proper domain adaptation is characterized by small δ and ϵ, and c close to 1. We also note that for both of the definitions given above, the inequality event can be reduced to $\mathrm{R}_{\mathcal{T}}(\mathcal{A}(S, T_u)) \leq \varepsilon$ when the hypothesis class $\mathcal{H}$ contains a zero-error hypothesis, *i.e.*, $\mathrm{R}_{\mathcal{T}}(\mathcal{H}) = 0$.

Finally, we will also need a definition that was introduced in [Ben-David and Urner, 2012] that expresses the capacity of a hypothesis class to produce a zero-error classifier with margin γ.

Definition 18. *Let $\mathbf{X} \subseteq \mathbb{R}^d$, $\mathcal{D}_{\mathbf{X}}$ be a distribution over $\mathbf{X}$, $h : \mathbf{X} \to \{0, 1\}$ be a classifier, and $B_\gamma(\mathbf{x})$ be the ball of radius γ around some domain point $\mathbf{x}$. We say that h is a γ-margin classifier with respect to $\mathcal{D}_{\mathbf{X}}$ if for all $\mathbf{x} \in \mathbf{X}$ whenever $\mathcal{D}_{\mathbf{X}}(B_\gamma(\mathbf{x})) > 0$, then $h(y) = h(z)$ holds for all $y, z \in B_\gamma(\mathbf{x})$.*

In [Ben-David and Urner, 2012], it was also noted that when h is a γ-margin classifier with respect to $\mathcal{D}_{\mathbf{X}}$, this is equivalent to h satisfying the Lipschitz-property with Lipschitz constant $\frac{1}{2\gamma}$ on the support of $\mathcal{D}_{\mathbf{X}}$. Thus, we can refer to this assumption as the Lipschitzness assumption. For the sake of completeness, we present the original definition of the probabilistic Lipschitzness below.

Definition 19. *Let $\phi : \mathbb{R} \to [0, 1]$. We say that $f : \mathbf{X} \to \mathbb{R}$ is ϕ-Lipschitz with respect to a distribution $\mathcal{D}_{\mathbf{X}}$ over $\mathbf{X}$ if, for all $\lambda > 0$, we have*

$$\Pr_{\mathbf{x} \sim \mathcal{D}_{\mathbf{X}}} \left[\exists \mathbf{x}' : |f(\mathbf{x}) - f(\mathbf{x}')| > \lambda \mu(\mathbf{x}, \mathbf{x}') \right] \leq \phi(\lambda),$$

where $\mu : \mathbf{X} \times \mathbf{X} \to \mathbb{R}_+$ is some metric over $\mathbf{X}$.

Common assumptions in domain adaptation We now proceed to recall the most common assumptions that were considered in the literature as those that ensure efficient adaptation.

1. **Covariate shift.** This assumption is among the most popular ones, and it has been extensively studied in a series of theoretical studies on the subject (see, for instance, [Sugiyama et al., 2008], and the references therein). While in domain adaptation we generally assume $\mathcal{S} \neq \mathcal{T}$, this can be further understood as $\mathcal{S}_{\mathbf{X}}(\mathbf{X})\mathcal{S}(Y|\mathbf{X}) \neq \mathcal{T}_{\mathbf{X}}(\mathbf{X})\mathcal{T}(Y|\mathbf{X})$, where $\mathcal{S}(Y|\mathbf{X}) = \mathcal{T}(Y|\mathbf{X})$ while $\mathcal{S}_{\mathbf{X}} \neq \mathcal{T}_{\mathbf{X}}$ is generally called the covariate shift assumption.

2a. **Similarity of the (unlabeled) marginal distributions.** [Ben-David et al., 2010b] considered the $\mathcal{H}\Delta\mathcal{H}$-distance between $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$ to assess the impossibility of domain adaptation, and assumed that it remains low between these two domains. This is the most straightforward assumption that directly follows from all of the proposed generalization bounds for domain adaptation. We refer the reader to Section 3 for the details.

2b. **Weight-ratio of the (unlabeled) marginal distributions.** The weight-ratio assumption was introduced in [Cortes et al., 2010], and further studied in [Ben-David and Urner, 2012] as a stronger concept of similarity between two marginal distributions. This is defined as:

$$C_{\mathcal{B}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) = \inf_{\substack{b \in \mathcal{B} \\ \mathcal{T}_{\mathbf{X}}(b) \neq 0}} \frac{\mathcal{S}_{\mathbf{X}}(b)}{\mathcal{T}_{\mathbf{X}}(b)}$$

with respect to a collection of input space subsets $\mathcal{B} \subseteq 2^{\mathbf{X}}$.

3. **Ideal joint error.** Finally, this last important assumption is the one that states that there should exist a low-error hypothesis for both domains. As explained in Section 3, this error can be defined as a so-called $\lambda_{\mathcal{H}}$ term, as follows:

$$\lambda_{\mathcal{H}} = \min_{h \in \mathcal{H}} \mathrm{R}_{\mathcal{S}}(h) + \mathrm{R}_{\mathcal{T}}(h).$$

These three assumptions are at the heart of impossibility theorems, where they are usually analyzed in a pair-wise fashion.

4.2 Constructive impossibility theorems

In what follows, we present a series of so-called impossibility results related to the domain adaptation problem. These results are then illustrated based on some concrete examples that highlight the pitfalls of domain adaptation algorithms.

To proceed, we present a theorem showing that some of the intuitive assumptions presented above do not suffice to guarantee the success of domain adaptation. More precisely, among the three assumptions that have been rapidly

discussed – covariate shift, small $\mathcal{H}\Delta\mathcal{H}$-distance between the unlabeled distributions, and the existence of the hypothesis that achieves low error on both the source and target domains (small $\lambda_{\mathcal{H}}$) – these last two are both necessary (and, as we know from previous results, are also sufficient).

Theorem 20 (Necessity of a small $\mathcal{H}\Delta\mathcal{H}$-distance [Ben-David et al., 2010b])**.** *Let $\mathbf{X}$ be some domain set, and $\mathcal{H}$ a class of functions over $\mathbf{X}$. Assume that, for some $\mathcal{A} \subseteq \mathbf{X}$, we have that $\{h^{-1}(1) \cap \mathcal{A} : h \in \mathcal{H}\}$ contains more than two sets and is linearly ordered by inclusion. Then, the conditions "covariate shift" plus "small $\lambda_{\mathcal{H}}$" do not suffice for domain adaptation. In particular, for every $\epsilon > 0$, there exists probability distributions $\mathcal{S}$ over $\mathbf{X} \times \{0, 1\}$, and $\mathcal{T}_{\mathbf{X}}$ over $\mathbf{X}$ such that for every domain adaptation learner $\mathcal{A}$, every integer $m > 0$, $n > 0$, there exists a labeling function $f : \mathbf{X} \to \{0, 1\}$ such that*

1. *$\lambda_{\mathcal{H}} \leq \epsilon$ is small;*
2. *$\mathcal{S}$ and $\mathcal{T}_f$ satisfy the covariate shift assumption;*
3. $\underset{\substack{S\sim(\mathcal{S})^m \\ T_u\sim(\mathcal{T}_{\mathbf{X}})^n}}{\mathbf{Pr}} \left[\mathrm{R}_{\mathcal{T}_f}(\mathcal{A}(S, T_u)) \geq \frac{1}{2}\right] \geq \frac{1}{2}$,

where the distribution $\mathcal{T}_f$ over $\mathbf{X} \times \{0, 1\}$ is defined as $\mathcal{T}_f\{1|\mathbf{x} \in \mathbf{X}\} = f(\mathbf{x})$.

This result highlights the importance of the need for small divergence between the marginal distributions of the domains, as even when the covariate shift assumption is satisfied and $\lambda_{\mathcal{H}}$ is small, the error of the classifier returned by a domain adaptation learner can be larger than $\frac{1}{2}$ with a probability that exceeds this same value. We now proceed to the symmetric result that shows the necessity for a small joint error between the two domains expressed by the $\lambda_{\mathcal{H}}$ term.

Theorem 21 (Necessity for a small $\lambda_{\mathcal{H}}$ [Ben-David et al., 2010b])**.** *Let $\mathbf{X}$ be some domain set, and $\mathcal{H}$ a class of functions over $\mathbf{X}$ where the VC dimension is much smaller than $|\mathbf{X}|$ (for instance, any $\mathcal{H}$ with a finite VC dimension over an infinite $\mathbf{X}$). Then, the conditions covariate shift plus small $\mathcal{H}\Delta\mathcal{H}$-divergence do not suffice for domain adaption. In particular, for every $\epsilon > 0$ there exist probability distributions $\mathcal{S}$ over $\mathbf{X} \times \{0, 1\}$, $\mathcal{T}_{\mathbf{X}}$ over $\mathbf{X}$, such that for every domain adaptation learner $\mathcal{A}$, every integer $m, n > 0$, there exists a labeling function $f : \mathbf{X} \to \{0, 1\}$ such that*

1. *$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{T}_{\mathbf{X}}, \mathcal{S}_{\mathbf{X}}) \leq \epsilon$ is small;*
2. *The covariate shift assumption holds;*
3. $\underset{\substack{S\sim\mathcal{S}^m \\ T_u\sim(\mathcal{T}_{\mathbf{X}})^n}}{\mathbf{Pr}} \left[\mathrm{R}_{\mathcal{T}_f}(\mathcal{A}(S, T_u)) \geq \frac{1}{2}\right] \geq \frac{1}{2}$.

Once again, this theorem shows that small divergence combined with a satisfied covariate shift assumption can lead to an error of the hypothesis returned by a domain adaptation learner that exceeds $\frac{1}{2}$ with high probability. Consequently, the main conclusion of these two theorems can be summarized as follows: among the studied assumptions, neither the assumption combination 1. and 3., nor 2a. and 3., suffice for successful domain adaptation in the unsupervised case. Another important conclusion that should be underlined here is that all generalization bounds for domain adaptation with a distance term and a joint error term introduced throughout this survey indeed imply learnability, even with the most straightforward learning algorithm. On the other hand, the covariate shift assumption is not really necessary: it cannot replace any of the other assumptions, and it becomes redundant when the other two assumptions hold. This study, however, needs further investigation, as in the case of semi-supervised domain adaptation, the situation can be drastically different.

Case of proper domain adaptation learning Below, we turn our attention to the impossibility results established in [Ben-David et al., 2012] for the case where the output of the given domain adaptation algorithm should be a hypothesis that belongs to some predefined hypothesis class. This particular constraint easily justifies itself in practice, where we may need to find a hypothesis as quickly as possible from a predefined set of hypotheses, at the expense of a higher error rate. The following result was obtained by [Ben-David and Urner, 2012] in this setting.

Theorem 22 ([Ben-David et al., 2012])**.** *Let domain $\mathbf{X} = [0, 1]^d$, for some d. Consider the class $\mathcal{H}$ of half-spaces as the target class. Let $\mathbf{x}$ and $\mathbf{z}$ be a pair of antipodal points on the unit sphere, and let $\mathcal{W}$ be a set that contains two pairs $(\mathcal{S}, \mathcal{T})$ and $(\mathcal{S}', \mathcal{T}')$ of distributions with:*

1. *both pairs satisfy the covariate shift assumption;*
2. *$f(\mathbf{x}) = f(\mathbf{z}) = 1$ and $f(\overline{0}) = 0$ for their common labeling function f;*

3. $\mathcal{S}_{\mathbf{X}}(\mathbf{x}) = \mathcal{T}_{\mathbf{X}}(\mathbf{z}) = \mathcal{S}_{\mathbf{X}}(\overline{0}) = \frac{1}{3}$;

4. $\mathcal{T}_{\mathbf{X}}(\mathbf{x}) = \mathcal{T}_{\mathbf{X}}(\overline{0}) = \frac{1}{2}$ *or* $\mathcal{T}'_{\mathbf{X}}(\mathbf{z}) = \mathcal{T}'_{\mathbf{X}}(\overline{0}) = \frac{1}{2}$.

Then, for any number m, any constant c, no proper domain adaptation learning algorithm can $(c, \varepsilon, \delta, m, 0)$ solve the domain adaptation learning task for $\mathcal{W}$ with respect to $\mathcal{H}$, if $\varepsilon < \frac{1}{2}$ and $\delta < \frac{1}{2}$. In other words, every learner that ignores unlabeled target data fails to produce a zero-risk hypothesis with respect to $\mathcal{W}$.

This theorem shows that having some amount of data generated by the target distribution is crucial for the learning algorithm to estimate whether the support of the target distribution is $\mathbf{x}$ and $\overline{0}$, or $\mathbf{z}$ and $\overline{0}$. Surprisingly, there is no possible way to obtain this information without having access to a sample drawn from the target distribution event if the point-wise weight-ratio is assumed to be as large as $\frac{1}{2}$. Thus, no amount of labeled source data can compensate for having a sample from the target marginal distribution.

Illustrative examples Now as the main impossibility theorems are stated, it can be useful to give an illustrative example of situations where different assumptions and different learning strategies might fail or succeed. To this end, [Ben-David et al., 2010b] considered several examples that showed the inadequacy of the covariate shift assumption explained above, as well as the limits of the reweighting scheme.

In what follows, the considered hypothesis class is restricted to the space of threshold functions on $[0, 1]$, where a threshold function $h_t(\mathbf{x})$ is defined for any $t \in [0, 1]$ as $h_t(\mathbf{x}) = 1$ if $\mathbf{x} < t$, and 0 otherwise. In this case, the set $\mathcal{H}\Delta\mathcal{H}$ becomes the class of half-open intervals.

Inadequacy of the covariate shift. Let us consider the following construction: for some small fixed $\xi \in \{0; 1\}$, let $\mathcal{T}$ be a uniform distribution over $\{2k\xi : k \in \mathbb{N}, 2k\xi \leq 1\} \times \{1\}$, and let the source distribution $\mathcal{S}$ be the uniform distribution over $\{(2k+1)\xi : k \in \mathbb{N}, (2k+1)\xi \leq 1\} \times \{0\}$. The illustration of these distributions is given in Figure 7.

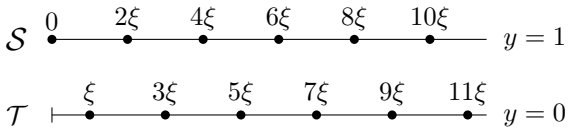

Figure 7: This scheme illustrates the considered source and target distributions that satisfy the covariate shift assumption with $\xi = \frac{2}{23}$.

For this construction, the following holds.

1. The covariate shift assumption holds for $\mathcal{T}$ and $\mathcal{S}$;
2. The distance $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) = \xi$, and thus it can be arbitrarily small;
3. The errors $\mathrm{R}_{\mathcal{S}}(\mathcal{H})$ and $\mathrm{R}_{\mathcal{T}}(\mathcal{H})$ are zero;
4. $\lambda_{\mathcal{H}}(\mathcal{S}, \mathcal{T}) = 1 - \xi$ and $\mathrm{R}_{\mathcal{T}}(h^*_{\mathcal{S}}) \geq 1 - \xi$ are large.

From this example it can instantly be seen that even if the covariate shift assumption is combined with a small $\mathcal{H}\Delta\mathcal{H}$-divergence between domains, this still results in a large joint error, and consequently in complete failure of the best source classifier when applied to the target distribution.

Reweighting method. A reweighting method in domain adaptation consists of the determination of a vector of weights $\mathbf{w} = \{w_1, w_2, \ldots, w_m\}$ that are used to reweight the unlabeled source sample S_u generated by $\mathcal{S}_{\mathbf{X}}$, to built a new distribution $\mathcal{T}_{\mathbf{w}}^{S_u}$ such that $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{T}_{\mathbf{w}}^{S_u}, \mathcal{T}_{\mathbf{X}})$ is as small as possible. In what follows, we denote this reweighted distribution $\mathcal{T}^{\mathcal{S}}$. This new sample is then fed to any available supervised learning algorithm at hand, to produce a classifier that is expected to have a good performance when applied subsequently in the target domain. As this method has a very important role in the domain adaptation, the authors also gave two intrinsically close examples that show both its success and failure under the standard domain adaptation assumptions.

We first consider the following scheme: for some small $\epsilon \in \left(0, \frac{1}{4}\right)$, we assume that the covariate shift assumption holds; *i.e.*, for any $\mathbf{x} \in \mathbf{X}$, $\mathcal{T}(y = 1|\mathbf{x}) = \mathcal{S}(y = 1|\mathbf{x}) = f(\mathbf{x})$. We define $f : \mathbf{X} \to [0, 1]$ as follows: for $\mathbf{x} \in [1 - 3\epsilon, 1 - \epsilon]$, we set $f(\mathbf{x}) = 0$, and otherwise we set $f(\mathbf{x}) = 1$. To define $\mathcal{S}$ and $\mathcal{T}$, we only have to specify their marginals $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$. To this end, we let $\mathcal{S}_{\mathbf{X}}$ be the uniform distribution over $[0, 1]$, and we let $\mathcal{T}_{\mathbf{X}}$ be the uniform distribution over $[1 - \epsilon, 1]$. This particular setting is shown in Figure 8.

The following observations follow from this construction.

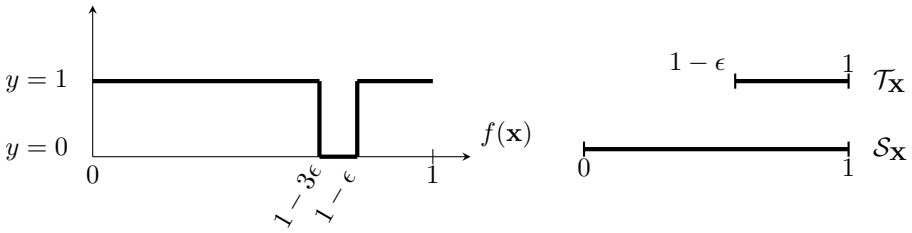

Figure 8: Illustration of the reweighting scenario. The source and target distributions satisfy the covariate shift assumption where f is their common conditional distribution. The marginal $\mathcal{S}_\mathbf{X}$ is the uniform distribution over $[0, 1]$, and the marginal $\mathcal{T}_\mathbf{X}$ is the uniform distribution over $[1 - \epsilon, 1]$.

1. For the given construction, the best joint hypothesis that defines $\lambda_\mathcal{H}$ is given by the function $h_{t=1}$; This function commits 0 errors on the target distribution and 2ϵ errors on the source distribution, thus giving $\lambda_\mathcal{H}(\mathcal{S}, \mathcal{T})$ equal to 2ϵ.
2. From the definition of $\mathcal{H}\Delta\mathcal{H}$-divergence, we obtain that $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}) = 1 - \epsilon$;
3. $\mathrm{R}_\mathcal{T}(h^*_\mathcal{S}) = 1$, $\mathrm{R}_\mathcal{T}(\mathcal{H}) = 0$, and $\mathrm{R}_\mathcal{S}(\mathcal{H}) = \epsilon$ achieved by the threshold functions $h_{t=1-3\epsilon}$, $h_{t=1}$ and $h_{t=1-3\epsilon}$, respectively.

On the other hand, it is possible to find a reweighting distribution that will produce a sample such that $\mathrm{R}_\mathcal{T}(h^*_{\mathcal{T}^\mathcal{S}}) \to 0$ in the probability when m and n tend towards infinity and $h^*_{\mathcal{T}^\mathcal{S}} = \operatorname{argmin}_{h \in \mathcal{H}} \mathrm{R}_\mathcal{T}(h_{\mathcal{T}^\mathcal{S}})$. This happens along with the probability of the source error tending to 1 when m grows to infinity. This example is a clear illustration of when a simple reweighting scheme can be efficient for adaptation. This, however, is not the case when we consider different labeling of the target data points. Let us now assume that the source distribution remains the same, while for the target distribution $f(\mathbf{x}) = 1$ for any $\mathbf{x} \in \mathbf{X}$. This slight change gives the following results:

1. $\lambda_\mathcal{H}(\mathcal{S}, \mathcal{T}) = \epsilon$;
2. $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}) = 1 - \epsilon$;
3. $\mathrm{R}_\mathcal{T}(h^*_\mathcal{S}) = 0$, $\mathrm{R}_\mathcal{T}(\mathcal{H}) = 0$ and $\mathrm{R}_\mathcal{S}(\mathcal{H}) = \epsilon$.

We can observe that the $\lambda_\mathcal{H}$ term has now become smaller, and that the best source hypothesis achieves a 0 error on the target distribution. However, the result that we obtain with the reweighting method is completely different: it is not difficult to see that $\mathrm{R}_\mathcal{T}(h^*_{\mathcal{T}^\mathcal{S}}) \to 1$ in the probability when m and n tend towards infinity, while the error of $h^*_\mathcal{S}$ will tend to zero.

We conclude by saying that the bound from [Ben-David et al., 2010a] recalled in Section 3 implies that $\mathrm{R}_\mathcal{T}(h^*_\mathcal{S})$ is bounded by $\mathrm{R}_\mathcal{T}(\mathcal{H}) + \lambda_\mathcal{H}(\mathcal{S}, \mathcal{T}) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X})$, and thus it can be hoped that by reweighting the sample S to reflect the distribution $\mathcal{T}_\mathbf{X}$, the term $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X})$ in that bound would be diminished. The last example, however, shows that this might not be the case, as $\mathrm{R}_{\mathcal{T}^L_\mathbf{w}\mathbf{X}}$ might be as bad as that bound allows.

4.3 Impossibility theorems based on sample complexity

We now present several results that assess the hardness of the domain through the lens of its sample complexity, which is usually defined as the number of training instances required to achieve a low-error classifier for a certain distribution $\mathcal{D}$. This setting in the context of the adaptation problem was studied by [Ben-David and Urner, 2012], where their first theorem established the sample complexity of solving a domain adaptation problem formulated as follows.

Theorem 23 ([Ben-David and Urner, 2012])**.** *For every finite domain $\mathbf{X}$, for every ε and δ with $\varepsilon + \delta < \frac{1}{2}$, no algorithm can $(\varepsilon, \delta, |S_u|, |T_u|)$-solve the domain adaptation problem for the class $\mathcal{W}$ of triples $(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}, f)$ with $C_\mathcal{B}(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}) \geq \frac{1}{2}$, $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}) = 0$, and $\mathrm{R}_\mathcal{T}(\mathcal{H}) = 0$ if*

$$|S_u| + |T_u| < \sqrt{(1 - 2(\varepsilon + \delta))|\mathbf{X}|},$$

, where $\mathcal{H}$ is the hypothesis class that contains only the all-1 and all-0 labeling functions, and $\mathrm{R}_\mathcal{T}(\mathcal{H}) = \min_{h \in \mathcal{H}} \mathrm{R}_\mathcal{T}(h, f)$.

This result is interesting in many ways. First, the assumptions used in the theorem are extremely simplified, which means that the *a-priori* knowledge about the target task is so strong that a zero error classifier for the given hypothesis

class can be obtained using only one labeled target instance. Secondly, we can also note that the considered setting is extremely favorable for adaptation, as the marginal distributions of the source and target domains are close both in terms of the $\mathcal{H}\Delta\mathcal{H}$-divergence and the weight-ratio $C_{\mathcal{B}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$. For the latter, this roughly means that the probability to encounter a source point is at least half of the probability of finding it in the target domain. These assumptions further spur the following surprising conclusions:

1. The sample complexity of domain adaptation cannot be bounded only in terms of the VC dimension of the class that can produce a hypothesis that achieves a zero error on it. This statement agrees well with the previous results, which shows the need for the existence of a good hypothesis for both domains;
2. Some data drawn from the target distribution should be available, to obtain a bound with an exclusive dependency on the VC dimension of the hypothesis class;
3. This result implies that the sample sizes that are needed to obtain useful approximations of the weight-ratio are prohibitively high.

We now provide another result provided by Ben-David and Urner that shows that the same lower bound can be obtained using the Lipschitzness assumption imposed on the labeling function f.

Theorem 24 ([Ben-David and Urner, 2012])**.** *Let $\mathbf{X} = [0,1]^d$, $\varepsilon > 0$ and $\delta > 0$ be such that $\varepsilon + \delta < \frac{1}{2}$, let $\lambda > 1$ and let $\mathcal{W}_\lambda$ be the set of triples $(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}, f)$ of distributions over $\mathbf{X}$ with $\mathrm{R}_{\mathcal{T}}(\mathcal{H}) = 0$, $C_{\mathcal{B}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \geq \frac{1}{2}$, $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) = 0$, and λ-Lipschitz labeling functions f. Then no domain adaptation-learner can $(\varepsilon, \delta, |S_u| + |T_u|)$-solve the domain adaptation problem for the class $\mathcal{W}_\lambda$, unless*

$$|S_u| + |T_u| \geq \sqrt{(\lambda+1)^d(1 - 2(\varepsilon + \delta))}.$$

4.4 Hardness results for sample complexity

So far we have presented theorems that show what conditions provably lead to the failure of domain adaptation. These results show that even in some extremely simple settings, successful adaptation might require an abundant amount of labeled source data, or at least a reasonable amount of labeled target data. In spite of this, a natural question that might be asked is to what extent the target domain unlabeled data can help to adapt when traded against some labeled source domain data. Before answering this question, we first turn our attention to the sample complexity results presented by [Ben-David et al., 2012], who investigated the existence of a learning method that can efficiently learn a good hypothesis for a target task provided that the target sample from its corresponding probability distribution is replaced by a (possibly larger) generated sample from a different probability distribution. The efficiency of such a learning method requires that it does not worsen the generalization guarantee of the learned classifier in the target domain. As an example of the considered classifier, we can take a popular nearest-neighbor classifier $h_{\mathrm{NN}}(\mathbf{x})$ that given a metric μ defined over the input space $\mathbf{X}$, assigns a label to a point $\mathbf{x}$ as $h_{\mathrm{NN}}(\mathbf{x}) = y(N_S(\mathbf{x}))$, with $N_S(\mathbf{x}) = \operatorname{argmin}_{\mathbf{z} \in S} \mu(\mathbf{x}, \mathbf{z})$ being the nearest neighbor of $\mathbf{x}$ in the labeled source sample S, and $y(N_S(\mathbf{x}))$ is the label of this nearest neighbor. The theorems obtained are proven under the covariate shift condition and the assumption of a bound on the weight-ratio between the two domains, as explained before. We now present below the first such theorem below.

Theorem 25 ([Ben-David et al., 2012])**.** *Let domain $\mathbf{X} = [0,1]^d$ and for some $C > 0$, let $\mathcal{W}$ be a class of pairs of source and target distributions $\{(\mathcal{S}, \mathcal{T}) | C_{\mathcal{B}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \geq C\}$ with a bounded weight-ratio and their common labeling function $f : \mathbf{X} \to [0,1]$, satisfying the ϕ-probabilistic-Lipschitz property with respect to the target distribution, for some function ϕ. Then, for all λ,*

$$\mathop{\mathbf{E}}_{S \sim \mathcal{S}^m} \left[\mathrm{R}_{\mathcal{T}}(h_{NN})\right] \;\leq\; 2\mathrm{R}^*_{\mathcal{T}}(\mathcal{H}) + \phi(\lambda) + 4\lambda \frac{\sqrt{d}}{C} m^{-\frac{1}{d-1}}.$$

This theorem suggests that under covariate shift and bounded weight-ratio assumptions, the expected target error of a NN classifier learned on a sample drawn from the source distribution is bounded by twice the optimal risk over the whole considered hypothesis space, plus several constants related to the nature of the labeling function and the dimension of the input space. Regarding these latter, it can be noted that if the labeling function is λ-Lipschitz in the standard sense of Lipschitzness, and the labels are deterministic, then we have $\mathrm{R}^*_{\mathcal{T}}(\mathcal{H}) = 0$ and $\phi(a) = 0$ for all $a \geq \lambda$. Applying Markov's inequality then yields the following corollary on the sample size bound which further strengthens the previous result.

Corollary 26. *Let domain $\mathbf{X} = [0,1]^d$ and for some $C > 0$, let $\mathcal{W}$ be a class of pairs of source and target distributions $\{(\mathcal{S}, \mathcal{T}) | C_{\mathcal{B}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \geq C\}$ with a bounded weight-ratio and their common labeling function $f : \mathbf{X} \to [0,1]$ satisfying the ϕ-probabilistic-Lipschitz property with respect to the target distribution, for some function ϕ. Then, for all $\varepsilon > 0$,*

$\delta > 0$, $m \geq \left(\frac{4\lambda\sqrt{d}}{C\varepsilon\delta}\right)^{d+1}$, *the nearest neighbor algorithm applied to a sample of size* m *has, with probability of at least* $1-\delta$, *error of at most* ε *w.r.t. the target distribution for any pair* $(\mathcal{S},\mathcal{T}) \in \mathcal{W}$.

This corollary provides the first positive result to establish the number of samples required for efficient adaptation in cases where no target data is available to the learner. A natural question that arises is then to quantify the utility of the additional unlabeled target data in the adaptation process, and the conditions required for it to succeed. To answer this question, the authors of [Ben-David and Urner, 2012] considered a particular adaptation algorithm $\mathcal{A}$, as summarized below.

Input: An *i.i.d.* sample $S_u \sim \mathcal{S}_{\mathbf{X}}$ labeled by f, an unlabeled *i.i.d.* sample $T_u \sim \mathcal{T}_{\mathbf{X}}$, and margin parameter γ.

Step 1. Partition $[0,1]^d$ into a collection $\mathcal{B}$ of boxes (axis-aligned rectangles) with side length $\gamma/\sqrt{d}$.
Step 2. Obtain sample S' by removing every point in S_u, which is sitting in a box that is not hit by T_u.
Step 3. Output an optimal risk-minimizing classifier from $\mathcal{H}$ for the sample S'.

The following theorem provides lower bounds for both the size of the source labeled and the target unlabeled samples required by algorithm $\mathcal{A}$, to learn well when a prior knowledge is assumed to be available to the learner in the form of a hypothesis class that realizes $\mathcal{T}_{\mathbf{X}}$ with margins, as in the definition above.

Theorem 27 ([Ben-David and Urner, 2012])**.** *Let* $\mathbf{X} = [0,1]^d$, $\gamma > 0$ *be a margin parameter,* $\mathcal{H}$ *be a hypothesis class of finite VC dimension, and* $\mathcal{W}$ *be the set of triples* $(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}, f)$ *of source distribution, target distribution, and labeling function with*

1. $C_{\mathcal{I}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \geq \frac{1}{2}$ *for the class* $\mathcal{I} = (\mathcal{H}\Delta\mathcal{H}) \cap \mathcal{B}$, *where* $\mathcal{B}$ *is a partition of* $[0,1]^d$ *into boxes of side length* $\frac{\gamma}{\sqrt{d}}$;
2. $\mathcal{H}$ *contains a hypothesis that has* γ*-margin on* $\mathcal{T}$;
3. *the labeling function* f *is a* γ*-margin classifier with respect to* $\mathcal{T}$.

Then there is a constant $c > 1$, *such that for all* $\varepsilon > 0$, $\delta > 0$, *and for all* $(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}, f) \in \mathcal{W}$, *when given an i.i.d. sample* S_u *from* $\mathcal{S}_{\mathbf{X}}$, *labeled by* f *of size*

$$|S_u| \geq c\left[\frac{VC(\mathcal{H}) + \log\frac{1}{\delta}}{C_{\mathcal{I}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})(1-\varepsilon)\varepsilon} \log\left(\frac{VC(\mathcal{H})}{C_{\mathcal{I}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})(1-\varepsilon)\varepsilon}\right)\right],$$

and an i.i.d. sample T_u *from* $\mathcal{T}_{\mathbf{X}}$ *of size*

$$|T_u| \geq \frac{1}{\epsilon}\left(2\left[\frac{\sqrt{d}}{\gamma}\right]^d \ln\left(3\left[\frac{\sqrt{d}}{\gamma}\right]^d \delta\right)\right),$$

then $\mathcal{A}$ *outputs a classifier* h *with* $\mathrm{R}_{\mathcal{T}}(h,f) \leq \epsilon$ *with probability of at least* $1-\delta$.

It is worth noting that these bounds follow the standard bounds from statistical learning theory, where the size of the learning sample required for successful learning is given as a function of the VC dimension of the hypothesis class. In domain adaptation, this dependency is further extended to the weight-ratio and the accuracy parameters of the learnability model. Moreover, we observe that this theorem considers the input space that might contain an infinite number of points. This assumption can lead to a vacuous bound, as in reality the input space often presents a finite domain, and the dependency of the sample size should be given in its terms. The following theorem covers this case.

Theorem 28. *Let* $\mathbf{X}$ *be some finite domain,* $\mathcal{H}$ *be a hypothesis class of finite VC dimension, and* $\mathcal{W} = \{(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}, f) | \mathrm{R}_{\mathcal{T}}(\mathcal{H}) = 0, C(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \geq 0\}$ *be a class of pairs of source and target distributions with bounded weight-ratio where* $\mathcal{H}$ *contains the zero-error hypothesis on* $\mathcal{T}$. *Then there is a constant* $c > 1$, *such that for all* $\varepsilon > 0$, $\delta > 0$, *and all* $(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}, f) \in \mathcal{W}$, *when given an i.i.d. sample* S_u *from* $\mathcal{S}_{\mathbf{X}}$, *labeled by* f *of size*

$$|S_u| \geq c\left[\frac{VC(\mathcal{H}) + \log\frac{1}{\delta}}{C(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})(1-\varepsilon)\varepsilon} \log\left(\frac{VC(\mathcal{H})}{C(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})(1-\varepsilon)\varepsilon}\right)\right],$$

and an i.i.d. sample T_u *from* $\mathcal{T}_{\mathbf{X}}$ *of size*

$$|T_u| \geq \frac{1}{\epsilon}\left(\frac{2|\mathbf{X}|\ln 3|\mathbf{X}|}{\delta}\right),$$

then algorithm $\mathcal{A}$ *outputs a classifier* h *with* $\mathrm{R}_{\mathcal{T}}(h,f) \leq \epsilon$ *with probability of at least* $1-\delta$.

To conclude, we note that both hardness results that state under which conditions the domain adaptation fails, and the results of the analysis of the sample sizes required from the source and target domains for the adaptation to succeed, fall into the category of the so-called impossibility theorems. They essentially draw the limits of the domain adaptation problem under various common assumptions, and provide insights into the hardness of solving this.

The case of agnostic proper domain adaptation We presented above an impossibility result for proper domain adaptation that shows that a conservative learner that is fed with a large labeled sample from the source domain might fail to produce a low-error classifier in the target domain, even under high weight-ratio and covariate shift assumptions. Below, we define a two-stage paradigm suggested by [Ben-David et al., 2012] that allows successful learning in this scenario. The proposed two-stage procedure consists of: 1) using a labeled source sample to learn an arbitrary hypothesis with decent performance on the target domain; and 2) applying the learned hypothesis to the unlabeled examples from the target domain, and feeding them to a standard agnostic learner. For the sake of clarity, the definition of an agnostic learning is given below.

Definition 20 ([Ben-David et al., 2012])**.** *For $\varepsilon > 0, \delta > 0$, $m \in \mathbb{N}$, we say that an algorithm (ε, δ, m) (agnostically) learns a hypothesis class $\mathcal{H}$, if for all distributions $\mathcal{D}$, when given an i.i.d. sample of size at least m, it outputs a classifier of error at most $\mathrm{R}_{\mathcal{D}}(\mathcal{H}) + \varepsilon$ with probability of at least $1 - \delta$. If the output of the algorithm is always a member of $\mathcal{H}$, we call it an agnostic proper learner for $\mathcal{H}$.*

This definition can now be used to prove the following theorem for the proposed two-stage procedure.

Theorem 29 ([Ben-David et al., 2012])**.** *Let $\mathbf{X}$ be some domain and $\mathcal{W}$ be a class of pairs $(\mathcal{S}, \mathcal{T})$ of distributions over $\mathbf{X} \times \{0, 1\}$ with $\mathrm{R}_{\mathcal{T}}(\mathcal{H}) = 0$, such that there is an algorithm $\mathcal{A}$ and functions $m : (0,1)^2 \to \mathbb{N}$, $n : (0,1)^2 \to \mathbb{N}$ such that $\mathcal{A}(0, \varepsilon, \delta, m(\varepsilon, \delta), n(\varepsilon, \delta))$-solves the domain adaptation learning task for $\mathcal{W}$ for all $\varepsilon, \delta > 0$. Let $\mathcal{H}$ be some hypotheses class for which there exists an agnostic proper learner. Then, the $\mathcal{H}$-proper domain adaptation problem can be $((0, \varepsilon, \delta, m(\varepsilon/3, \delta/2), n(\varepsilon/3, \delta/2)) + m'(\varepsilon/3, \delta/2))$-solved with respect to the class $\mathcal{W}$, where m' is the sample complexity function for agnostically learning $\mathcal{H}$.*

As in the previous case, the algorithm $\mathcal{A}$ in the statement of this theorem can be considered to be the nearest neighbor classifier $\mathrm{NN}(\mathcal{S})$, if the class $\mathcal{W}$ satisfies the conditions from the theorem. To summarize, the presented theorems for the proper domain adaptation learning show that with a domain adaptation algorithm that takes into account the unlabeled instances from the target marginal distribution, it might be possible to solve the proper domain adaptation problem, while in the contrary case, it is provably unsolvable.

4.5 Other relevant contributions

[Redko et al., 2019b] In this study, the authors provide a first analysis for consistent estimation of the adaptability term λ when some target label data is available. The main construction used in their study is to express the ideal joint hypothesis $h^* = \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{R}_{\mathcal{S}}(h, f_{\mathcal{S}}) + \mathrm{R}_{\mathcal{T}}(h, f_{\mathcal{T}})$ as a barycenter of the source and target labeling functions $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$. These latter are then considered to be probability measures over $\mathbf{X}$, so that the barycenter is defined over the space of probability distributions without requiring a hypothesis space to be picked in advance.

[Zhao et al., 2019] In this paper, the authors provide an example similar to that given in [Ben-David et al., 2010b], to show that small $\mathcal{H}$-divergence between marginal distributions and low source error do not guarantee good performance in the target domain. They further argue that this is mainly explained by the shift in the conditional distributions over the two domains that is accounted for by the inestimable adaptability term.

[Johansson et al., 2019] This paper proceeds in a spirit similar to that of [Zhao et al., 2019], by first showing an example where finding an invariant representation decreasing the shift between the two domains while minimizing the source error leads to poor performance in the target domain. This is attributed to the unobserved adaptability term and lack of invertability of the learned representation, and it is dealt with by taking into account the performance of a hypothesis in the source domain in regions where the source density is sufficiently high. The authors then provide a tight learning bound based on a weighted source error, a support discrepancy, and an unobservable term that characterizes the invertability of the invariant representation.

[Hanneke and Kpotufe, 2019] In this paper, the authors consider a semi-supervised setting where the goal is to learn a hypothesis from a mixture of labeled source and target samples, and to bound the excess risk of this hypothesis, *i.e.*, $\mathrm{R}_{\mathcal{D}}(h) - \mathrm{R}_{\mathcal{D}}(\mathcal{H})$ in each domain. The paper further introduces the novel concept of discrepancy between the two domains, called "transfer-exponents", and provides the first minimax-rates, in terms of both source and target sample size and of the latter divergence, similar to the work of [Ben-David and Urner, 2012].

4.6 Summary

In this section, we covered a series of results that establish the conditions required to make a domain adaptation problem solvable. As shown, these necessary conditions might take on different forms, and depend on the value of certain terms presented in the generalization bound and on the size of the available source and target learning samples. The take-away messages of this section can be summarized as follows:

1. Solving a domain adaptation problem requires two independent conditions to be fulfilled. First, there is the need to properly minimize the divergence between the source and target marginal distributions. Secondly, there is the need to ensure simultaneously that the *a-priori* adaptability of the two domains is high (which is reflected by the small ideal joint error term $\lambda_{\mathcal{H}}$);
2. Even under some strong assumptions that make the adaptation problem appear to be easy to solve, there might still be the need for a certain amount of unlabeled source and target data that in the most general case, can be prohibitively large;
3. A certain amount of labeled source and unlabeled target data can ensure efficient adaptation, and can produce a hypothesis with a small target error. In both cases, this amount depends on the general characteristics of the adaptation problem given by the weight-ratio and the complexity of the hypothesis space represented by its VC dimension;
4. In proper domain adaptation, ignoring unlabeled target data leads to provably unsolvable adaptation problems, where the domain adaptation learner fails to produce a zero-error hypothesis for the target domain.

All these conclusions provide us with a more general view on the learning properties of the adaptation phenomenon, and essentially provide a list of conditions that need to be verified to make sure that the adaptation problem at hand can be solved efficiently. Apart from that, the established results also provide us with an understanding that some adaptation tasks are harder when compared to others, and that this hardness can be quantified by not one, but several, criteria that take into account both the data distribution and the labeling of instances. Finally, they also show that successful adaptation requires a certain amount of data to be available during the adaptation step, and that this amount might directly depend on the proximity of the marginal distributions of the two domains. This last feature is quite important, as it is added to the dependence on the complexity of the hypothesis class considered previously in the standard supervised learning described in Section 2.

5 Learning bounds with integral probability metrics

In the previous sections, we presented several seminal results regarding the generalization bounds for domain adaptation and the impossibility theorems for some of them. We have shown that the basic shape of generalization bounds in the context of domain adaptation remains more or less the same, and mainly differs only in the divergence used to measure the distance between the source and the target marginal distributions. In this section, we consider a large family of metrics on the space of probability measures known as IPMs that present a well-studied topic in probability theory. In particular, we show that depending on the chosen functional class, some instances of IPMs can have interesting properties that are completely different from those shown by both the $\mathcal{H}\Delta\mathcal{H}$-divergence and the discrepancy distance seen previously.

5.1 Problem set-up

Integral probability metrics represent a large class of distances defined on the space of probability measures that have found applications in many machine-learning algorithms. The general definition of IPMs can be given as follows.

Definition 21 ([Zolotarev, 1984])**.** *Given two probability measures $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$ defined on a measurable space $\mathbf{X}$, the IPM is defined as*

$$D_{\mathcal{F}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) = \sup_{f \in \mathcal{F}} \left| \int_{\mathbf{X}} f d\mathcal{S}_{\mathbf{X}} - \int_{\mathbf{X}} f d\mathcal{T}_{\mathbf{X}} \right|,$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on $\mathbf{X}$.

As mentioned by [Müller, 1997], the quantity $D_{\mathcal{F}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ is a semimetric, and it is a metric if and only if the function class $\mathcal{F}$ separates the set of all signed measures with $\mu(\mathbf{X}) = 0$. It then follows that for any non-trivial function class $\mathcal{F}$, the quantity $D_{\mathcal{F}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ is zero if $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$ are the same. Several important special cases of IPMs can be obtained by specifically choosing the functional class $\mathcal{F}$. We present those that were used for the analysis of the domain adaptation problem below.

Maximum mean discrepancy Let $\mathcal{F} = \{f : \|f\|_{\mathcal{H}_k} \leq 1\}$ where $\mathcal{H}_k$ is a RKHS with its associated kernel k. Then, the maximum mean discrepancy (MMD) distance is defined as follows:

$$d_{\text{MMD}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left| \int f d(\mathcal{S}_{\mathbf{X}} - \mathcal{T}_{\mathbf{X}}) \right| = \left\| \int_{\mathbf{X}} k(\mathbf{x}, \cdot) d(\mathcal{S}_{\mathbf{X}} - \mathcal{T}_{\mathbf{X}}) \right\|_{\mathcal{H}_k}.$$

From a practical point of view, we observe that numerous domain adaptation and transfer learning approaches have been based on MMD minimization [Pan et al., 2009, Geng et al., 2011, Huang et al., 2006, Pan et al., 2008, Chen et al., 2009], and thus a theoretical analysis of the domain adaptation problem with this is of high scientific interest.

Wasserstein distance Let $\mathcal{F} = \{f : \|f\|_L \leq 1\}$ where

$$\|f\|_L = \sup_{\mathbf{x} \neq \mathbf{x}' \in \mathbf{X}} \frac{|f(\mathbf{x}) - f(\mathbf{x}')|}{c(\mathbf{x}, \mathbf{x}')}$$

is the Liptschitz semi-norm for real-valued continuous f on $\mathbf{X}$ and some metric $c(\cdot, \cdot) : \mathbf{X} \times \mathbf{X} \to \mathbb{R}_+$.

In this case, the Kantorovich-Rubinstein theorem [Dudley, 2002] yields the following result, with the Wasserstein distance W_1 defined as follows:

$$W_1(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) = \sup_{\|f\|_L \leq 1} \left| \int f d(\mathcal{S}_{\mathbf{X}} - \mathcal{T}_{\mathbf{X}}) \right| = \inf_{\gamma \in \Pi(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})} \int_{\mathbf{X} \times \mathbf{X}} c(\mathbf{x}, \mathbf{x}') d\gamma(\mathbf{x}, \mathbf{x}'),$$

where $\Pi(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ is a space of all joint probability measures on $\mathbf{X} \times \mathbf{X}$ with marginals $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$.

The original optimal transportation problem was introduced by [Monge, 1781] to study the problem of resource allocation. Its modern formulation, which led to the introduction of the Wasserstein distance, is due to [Kantorovich, 1942], who proposed a relaxation of the Monge's problem allowing to prove the existence of a unique minimizer for it. Despite being a very powerful tool for comparing and aligning probability distributions, the Wasserstein distance has become an emerging topic in machine learning only recently due to [Cuturi, 2013], where an efficient regularization scheme that allowed the solving of the optimal transportation problem was introduced.

5.2 Generalization bound with IPMs

We start this section with a general result that introduces IPMs to the domain adaptation generalization bounds provided by [Zhang et al., 2012]. In this paper, the authors considered a general multi-source scenario where not one, but $K \geq 2$ source domains are available. To be consistent with the rest of the survey, we present the main result of [Zhang et al., 2012] that introduces the IPMs in the context of domain adaptation specified for the case of one source and one target domain below.

Theorem 30. *For a labeling function $f \in \mathcal{G}$, let $\mathcal{F} = \{(\mathbf{x}, y) \to \ell(f(\mathbf{x}), y)\}$ be a loss function class that consists of the bounded functions with the range $[a, b]$ for a space of labeling functions $\mathcal{G}$. Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ be a labeled sample drawn from $\mathcal{S}$ of size m. Then, given any arbitrary $\xi \geq D_{\mathcal{F}}(\mathcal{S}, \mathcal{T})$, we have for any $m \geq \frac{8(b-a)}{\xi'^2}$ and any $\epsilon > 0$, with probability of at least $1 - \epsilon$, the following holds*

$$\sup_{f \in \mathcal{F}} \left| \mathrm{R}^{\ell}_{\hat{\mathcal{S}}} f - \mathrm{R}^{\ell}_{\mathcal{T}} f \right| \leq D_{\mathcal{F}}(\mathcal{S}, \mathcal{T}) + \left(\frac{\ln \mathcal{N}_1(\xi'/8, \mathcal{F}, 2m) - \ln(\epsilon/8)}{\frac{m}{32(b-a)^2}} \right)^{\frac{1}{2}},$$

where $\xi' = \xi - D_{\mathcal{F}}(\mathcal{S}, \mathcal{T})$.

Here the quantity $\mathcal{N}_1(\xi, \mathcal{F}, 2m)$ is defined in terms of the uniform entropy number (see Definition 8), and it is given by the following equation

$$\mathcal{N}_1(\xi, \mathcal{F}, 2m) = \sup_{\{S^{2m}\}} \log N\big(\xi, \mathcal{F}, \ell_1(\{S^{2m}\})\big),$$

where for the source sample S and its associated ghost sample $S' = \{(\mathbf{x}'_1, y'_1), \ldots, (\mathbf{x}'_m, y'_m)\}$ drawn from $\mathcal{S}$, the quantity $S^{2m} = \{S, S'\}$ and the metric ℓ_1 are a variation of the ℓ_1 metric defined for some $f \in \mathcal{F}$ based on the following norm

$$\|f\|_{\ell_1(\{S^{2m}\}))} = \frac{1}{m} \sum_{i=1}^{m} \Big(|f(\mathbf{x}_i, y_i)| + |f(\mathbf{x}'_i, y_i)| \Big).$$

It can be noted that there are several peculiarities related to this result. First, it is different from other generalization bounds provided before, as the divergence term here is defined for the joint distributions $\mathcal{S}$ and $\mathcal{T}$, and not for the marginal distributions $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$. Note that, in general, the joint target distribution $\mathcal{T}$ cannot be estimated in the classical scenario of unsupervised domain adaptation, as this can be done only when target labels are known, thus making the application of this bound quite uninformative in practice. Secondly, the proposed bound is very general, as it does not specify explicitly the functional class $\mathcal{F}$ considered in the definition of the IPM. On the one hand, this allows this bound to be adjusted to any instance of IPMs that can be obtained by choosing the appropriate functional class, while on the other hand, it also requires the uniform entropy number for this to be determined. Finally, the authors established a link between the discrepancy distance seen before and the $D_{\mathcal{F}}(\mathcal{S}, \mathcal{T})$ that allows us to obtain a bound with a more "traditional" shape. More precisely, the authors proved that the following inequality holds in the case of one source and one target domain for any ℓ and functional class $\mathcal{F}$:

$$D_{\mathcal{F}}(\mathcal{S}, \mathcal{T}) \leq disc_{\ell}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \sup_{g \in \mathcal{G}} \left| \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{T}_{\mathbf{X}}} [\ell(g(\mathbf{x}), f_{\mathcal{T}}(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{T}_{\mathbf{X}}} [\ell(g(\mathbf{x}), f_{\mathcal{S}}(\mathbf{x})] \right| .$$

Note that the second term of the right-hand side is basically a disagreement between the labeling functions $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$ that is zero only when they are equal. Using this inequality, it can be shown that the proposed theorem can be reduced to the following shape:

$$\sup_{f \in \mathcal{F}} |\mathrm{R}^{\ell}_{\hat{\mathcal{S}}} f - \mathrm{R}^{\ell}_{\mathcal{T}} f| \leq disc_{\ell}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \lambda + \left(\frac{\ln \mathcal{N}_1(\xi'/8, \mathcal{F}, 2m) - \ln(\epsilon/8)}{\frac{m}{32(b-a)^2}} \right)^{\frac{1}{2}}, \tag{7}$$

where $\lambda = \sup_{g \in \mathcal{G}} \left| \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{T}_{\mathbf{X}}} [\ell(g(\mathbf{x}), f_{\mathcal{T}}(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{T}_{\mathbf{X}}} [\ell(g(\mathbf{x}), f_{\mathcal{S}}(\mathbf{x})] \right|$, and the last term is the complexity term that depends on the covering number of the space $\mathcal{F}$, similar to the bounds based on the algorithmic robustness presented by Section 2. To this end, Equation (7) now looks similar to the generalization bounds from the previous sections.

To show that for a finite complexity term the difference between the empirical source risk and the target risk never exceeds the divergence between the two domains with the increasing number of available source examples, the authors proved the following theorem.

Theorem 31. *For a labeling function $f \in \mathcal{G}$, let $\mathcal{F} = \{(\mathbf{x}, y) \to \ell(f(\mathbf{x}), y)\}$ be a loss function class that consists of the bounded functions with the range $[a, b]$ for a space of labeling functions $\mathcal{G}$. If the following holds*

$$\lim_{m \to \infty} \frac{\ln \mathcal{N}_1(\xi'/8, \mathcal{F}, 2m)}{\frac{m}{32(b-a)^2}} < \infty ,$$

with $\xi' = \xi - D_{\mathcal{F}}(\mathcal{S}, \mathcal{T})$, then we have for any $\xi \geq D_{\mathcal{F}}(\mathcal{S}, \mathcal{T})$,

$$\lim_{m \to \infty} \mathbf{Pr} \left\{ \sup_{f \in \mathcal{F}} |\mathrm{R}^{\ell}_{\hat{\mathcal{S}}} f - \mathrm{R}^{\ell}_{\mathcal{T}} f| > \xi \right\} = 0.$$

It can be noted here that the probability of event $\{\sup_{f \in \mathcal{F}} |\mathrm{R}^{\ell}_{\hat{\mathcal{S}}} f - \mathrm{R}^{\ell}_{\mathcal{T}} f| > \xi\}$ is taken with respect to the threshold $\xi \geq D_{\mathcal{F}}(\mathcal{S}, \mathcal{T})$, while in standard learning theory this guarantee is usually stated for any $\xi > 0$ given that $\lim_{m \to \infty} \frac{\ln \mathcal{N}_1(\xi, \mathcal{F}, m)}{m} < \infty$. This highlights an important difference between the classic generalization bounds for supervised learning and the result given by Theorem 30.

As we mentioned above, the general setting for generalization bounds with IPMs proposed by Zhang *et al.* suffers from two major drawbacks: (1) the function class in the definition of the IPM is not specified, making it intractable to compute; (2) the proposed bounds are established for joint distributions rather than marginal distributions, making them not very informative in practice. To this end, we present below two different lines of research that tackle these drawbacks, and establish the generalization bounds for domain adaptation by explicitly considering a particular function class with a divergence term that takes into account the discrepancy between the marginal distributions of the source and target domains. These lines lead to two important particular cases of IPMs that were used to derive generalization bounds in domain adaptation: the Wasserstein distance and the MMD. We take a closer look at both of these in what follows.

5.3 Learning bounds with the Wasserstein distance

Despite many important theoretical insights presented previously, the above-mentioned divergence measures, such as the $\mathcal{H}\Delta\mathcal{H}$-divergence and the discrepancy, do not directly take into account the geometry of the data distribution when

estimating the discrepancy between two domains. Recently, [Courty et al., 2014] proposed to tackle this drawback by solving the domain adaptation using the Wasserstein distance. To justify domain adaptation algorithms based on the minimization of the Wasserstein distance, the generalization bounds for the three domain adaption settings involving this latter were presented by [Redko et al., 2017]. According to [Villani, 2009], the Wasserstein distance is relatively strong and can be combined with smoothness bounds to obtain convergences in other distances. As mentioned by the authors, this important advantage of the Wasserstein distance leads to tighter bounds in comparison to other state-of-the-art results, and it is more computationally attractive, as explained below.

To proceed, let $\mathcal{F} = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1\}$, where $\mathcal{H}_k$ is a RKHS with its associated kernel k. Let $\ell_{h,f} : \mathbf{x} \to \ell(h(\mathbf{x}), f(\mathbf{x}))$ be a convex loss-function defined $\forall h, f \in \mathcal{F}$, and assume that ℓ obeys the triangle inequality. As before, $h(\mathbf{x})$ corresponds to the hypothesis and $f(\mathbf{x})$ to the true labeling functions. Considering that $(h, f) \in \mathcal{F}^2$, the loss function ℓ is a non-linear mapping of the RKHS $\mathcal{H}_k$ for the family of ℓ_q losses defined previously[2]. Using results from [Saitoh, 1997], it can be shown that $\ell_{h,f}$ also belongs to the RKHS $\mathcal{H}_{k^q}$, admitting the reproducing kernel k^q, and that its norm obeys the following inequality:

$$||\ell_{h,f}||^2_{\mathcal{H}_{k^q}} \leq ||h - f||^{2q}_{\mathcal{H}_k}.$$

This result gives us two important properties of $\ell_{f,h}$ that are used further:

1. the function $\ell_{h,f}$ belongs to the RKHS, which allows us to use the reproducing property via some feature map $\phi(\mathbf{x})$ associated to kernel k^q;
2. the norm $||\ell_{h,f}||_{\mathcal{H}_{k^q}}$ is bounded.

Thus, the error function defined above can be also expressed in terms of the inner product in the corresponding Hilbert space, *i.e*[3],

$$\mathrm{R}^{\ell}_{\mathcal{D}}(h, f_{\mathcal{D}}) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}}[\ell(h(\mathbf{x}), f_{\mathcal{D}}(\mathbf{x}))] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}}[\langle\phi(\mathbf{x}), \ell\rangle_{\mathcal{H}_{k^q}}].$$

Now the following lemma that relates the Wasserstein metric with the source and target error functions for an arbitrary pair of hypotheses can be proved.

Lemma 32 ([Redko et al., 2017])**.** *Let $\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}} \in \mathcal{P}(\mathbf{X})$ be two probability measures on $\mathbb{R}^d$. Assume that the cost function $c(\mathbf{x}, \mathbf{x}') = \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_{\mathcal{H}_{k_\ell}}$, where $\mathcal{H}$ is a RKHS equipped with kernel $k_\ell : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ induced by $\phi : \mathbf{X} \to \mathcal{H}_{k_\ell}$ and $k_\ell(\mathbf{x}, \mathbf{x}') = \langle\phi(\mathbf{x}), \phi(\mathbf{x}')\rangle_{\mathcal{H}_{k_\ell}}$. Assume further that the loss function $\ell_{h,f} : \mathbf{x} \longrightarrow \ell(h(\mathbf{x}), f(\mathbf{x}))$ is convex, symmetric, bounded, obeys triangle equality, and has the parametric form $|h(\mathbf{x}) - f(\mathbf{x})|^q$ for some $q > 0$. Assume also that the kernel k_ℓ in the RKHS $\mathcal{H}_{k_\ell}$ is square-root integrable w.r.t. both $\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}$ for all $\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}} \in \mathcal{P}(\mathbf{X})$ where $\mathbf{X}$ is separable and $0 \leq k_\ell(\mathbf{x}, \mathbf{x}') \leq K, \forall\, \mathbf{x}, \mathbf{x}' \in \mathbf{X}$. If $\|\ell\|_{\mathcal{H}_{k_\ell}} \leq 1$, then the following holds*

$$\forall (h, h') \in \mathcal{H}^2_{k_\ell}, \quad \mathrm{R}^{\ell_q}_{\mathcal{T}}(h, h') \leq \mathrm{R}^{\ell_q}_{\mathcal{S}}(h, h') + W_1(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}).$$

This lemma makes use of the Wasserstein distance to relate the source and target errors. The assumption made here is to specify for the cost function that $c(\mathbf{x}, \mathbf{x}') = \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_{\mathcal{H}}$. While it might appear too restrictive, this assumption is, in fact, not that strong. Using the properties of the inner-product, we have

$$\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_{\mathcal{H}} = \sqrt{\langle\phi(\mathbf{x}) - \phi(\mathbf{x}'), \phi(\mathbf{x}) - \phi(\mathbf{x}')\rangle_{\mathcal{H}}} = \sqrt{k(\mathbf{x}, \mathbf{x}) - 2k(\mathbf{x}, \mathbf{x}') + k(\mathbf{x}, \mathbf{x}')}.$$

As the authors noted, it is possible to further show that for any given positive-definite kernel k there is a distance c (used as a cost function in our case) that generates this, and *vice versa* (see Lemma 12 from [Sejdinovic et al., 2013]).

The following generalization bound was proven by the authors using a result that showed the convergence of the empirical measure $\hat{\mu}$ to its true associated measure *w.r.t.* the Wasserstein metric provided by [Bolley et al., 2007].

Theorem 33. *Under the assumptions of Lemma 32, let S_u and T_u be two samples of size N_S and N_T drawn i.i.d. from $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$, respectively. Let $\hat{\mathcal{S}}_{\mathbf{X}} = \frac{1}{N_S}\sum_{i=1}^{N_S}\delta_{\mathbf{x}_i^S}$ and $\hat{\mathcal{T}}_{\mathbf{X}} = \frac{1}{N_T}\sum_{i=1}^{N_T}\delta_{\mathbf{x}_i^T}$ be the associated empirical measures. Then for any $d' > d$ and $\varsigma' < \sqrt{2}$, there exists some constant N_0 depending on d', such that for any $\delta > 0$ and $\min(N_S, N_T) \geq N_0 \max(\delta^{-(d'+2)}, 1)$ with probability of at least $1 - \delta$ for all h, we have*

$$\mathrm{R}^{\ell_q}_{\mathcal{T}}(h) \leq \mathrm{R}^{\ell_q}_{\mathcal{S}}(h) + W_1(\hat{\mathcal{S}}_{\mathbf{X}}, \hat{\mathcal{T}}_{\mathbf{X}}) + \sqrt{2\log\left(\frac{1}{\delta}\right)/\varsigma'}\left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right) + \lambda\,,$$

where λ is the combined error of the ideal hypothesis h^* that minimizes the combined error of $\mathrm{R}^{\ell_q}_{\mathcal{S}}(h) + \mathrm{R}^{\ell_q}_{\mathcal{T}}(h)$.

[2] If $(h, f) \in \mathcal{F}^2$ then $h - f \in \mathcal{F}$, which implies that $\ell(h(\mathbf{x}), f(\mathbf{x})) = |h(\mathbf{x}) - f(\mathbf{x})|^q$ is a nonlinear transform for $h - f \in \mathcal{F}$.

[3] For simplicity, we further write ℓ meaning $\ell_{f,h}$.

A first immediate consequence of this theorem is that it justifies the use of the optimal transportation in the domain adaptation context when combined with the minimization of the source error, and assuming the joint error given by the λ term is small. For this latter, [Courty et al., 2014] proposed a class-labeled regularization term added to the original optimal transport formulation to restrict source examples of different classes to be transported to the same target example, by promoting group sparsity in the matrix γ due to $\|\cdot\|_q^p$ with $q = 1$ and $p = \frac{1}{2}$. In some way, this regularization term influences the capability term, by ensuring the existence of a good hypothesis that will be discriminant on both source and target domain data.

Semi-supervised case To remain consistent with the previous sections, we also provide the generalization bound for the Wasserstein distance in the semi-supervised setting below.

Theorem 34 ([Redko et al., 2017]). *Let S_u, T_u be unlabeled samples of size N_S and N_T each, drawn independently from $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$, respectively. Let S be a labeled sample of size m generated by drawing $\beta\, m$ points from $\mathcal{T}_\mathbf{X}$ ($\beta \in [0,1]$) and $(1-\beta)\, m$ points from $\mathcal{S}_\mathbf{X}$ and labeling them according to $f_\mathcal{S}$ and $f_\mathcal{T}$, respectively. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\mathrm{R}^\alpha_{\hat{\mathcal{S}}}(h)$ on S and $h^*_T = \underset{h\in\mathcal{H}}{\operatorname{argmin}} \mathrm{R}^{\ell_q}_\mathcal{T}(h)$, then for any $\delta \in (0,1)$ with probability of at least $1-\delta$ (over the choice of samples),*

$$\mathrm{R}^{\ell_q}_\mathcal{T}(\hat{h}) \;\leq\; \mathrm{R}^{\ell_q}_\mathcal{T}(h^*_T) + c_1 + 2(1-\alpha)(W_1(\hat{\mathcal{S}}_\mathbf{X}, \hat{\mathcal{T}}_\mathbf{X}) + \lambda + c_2),$$

where

$$c_1 = 2\sqrt{\frac{2K\left(\frac{(1-\alpha)^2}{1-\beta} + \frac{\alpha^2}{\beta}\right)\log(2/\delta)}{m}} + 4\sqrt{K/m}\left(\frac{\alpha}{m\beta\sqrt{\beta}} + \frac{(1-\alpha)}{m(1-\beta)\sqrt{1-\beta}}\right),$$

$$c_2 = \sqrt{2\log\left(\frac{1}{\delta}\right)/\varsigma'}\left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right).$$

In line with the results obtained previously, this theorem shows that the best hypothesis that takes into account both source and target labeled data (*i.e.*, $0 \leq \alpha < 1$) performs at least as good as the best hypothesis learned on target data instances alone ($\alpha = 1$). This result agrees well with the intuition that semi-supervised domain adaptation approaches should be at least as good as unsupervised ones.

5.4 Generalization bound with MMD

Based on the results with the Wasserstein distance, we now introduce learning bounds for the target error where the divergence between the task distributions is measured by the MMD distance. As before, we start with a lemma that relates the source and target errors in terms of the introduced discrepancy measure for an arbitrary pair of hypotheses. Then, we show how the target error can be bounded by the empirical estimate of the MMD plus the complexity term.

Lemma 35 ([Redko, 2015]). *Let $\mathcal{F} = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1\}$ where $\mathcal{H}_k$ is a RKHS with its associated kernel k. Let $\ell_{h,f} : \mathbf{x} \to \ell(h(\mathbf{x}), f(\mathbf{x}))$ be a convex loss-function with a parametric form $|h(\mathbf{x}) - f(\mathbf{x})|^q$ for some $q > 0$, and defined $\forall h, f \in \mathcal{F}$ such that ℓ obeys the triangle inequality. Then, if $\|l\|_{\mathcal{H}_{kq}} \leq 1$, we have :*

$$\forall (h, h') \in \mathcal{F}, \quad \mathrm{R}^{\ell_q}_\mathcal{T}(h, h') \leq \mathrm{R}^{\ell_q}_\mathcal{S}(h, h') + d_{MMD}(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}).$$

This lemma is proved in a similar way to Lemma 32 from [Redko et al., 2017], as presented before in this section. Using this and the result that relates the true and the empirical MMD distances [Song, 2008], we can prove the following theorem.

Theorem 36. *With the assumptions from Lemma 35, let S_u and T_u be two samples of size m drawn i.i.d. from $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$, respectively. Then, with probability of at least $1-\delta (\delta \in (0,1))$ for all $h \in \mathcal{F}$, the following holds:*

$$\mathrm{R}^{\ell_q}_\mathcal{T}(h) \leq \mathrm{R}^{\ell_q}_\mathcal{S}(h) + d_{MMD}(\hat{\mathcal{S}}_\mathbf{X}, \hat{\mathcal{T}}_\mathbf{X}) + \frac{2}{m}\left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_\mathbf{X}}\left[\sqrt{\mathrm{tr}(K_\mathcal{S})}\right] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}_\mathbf{X}}\left[\sqrt{\mathrm{tr}(K_\mathcal{T})}\right]\right) + 2\sqrt{\frac{\log(\frac{2}{\delta})}{2m}} + \lambda,$$

where $d_{MMD}(\hat{\mathcal{S}}_\mathbf{X}, \hat{\mathcal{T}}_\mathbf{X})$ is an empirical counterpart of $d_{MMD}(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X})$, $K_\mathcal{S}$ and $K_\mathcal{T}$ are the kernel functions calculated on samples from $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$, respectively, and λ is the combined error of the ideal hypothesis h^* that minimizes the combined error of $\mathrm{R}^{\ell_q}_\mathcal{S}(h) + \mathrm{R}^{\ell_q}_\mathcal{T}(h)$.

We can see that this theorem is similar in shape to Theorem 33 and Theorem 10. The main difference, however, is that the complexity term does not depend on the Vapnik-Chervonenkis dimension. In our case, the loss function between two errors is bounded by the empirical MMD between distributions and two terms that correspond to the empirical Rademacher complexities of $\mathcal{H}$ *w.r.t.* the source and target samples. In both theorems, λ has the role of the combined error of the ideal hypothesis. Its presence in the bound comes from the use of triangle inequality for the classification error.

This result is particularly useful, as an unbiased estimate of the squared MMD distance $d^2_{\text{MMD}}(\hat{\mathcal{S}}_{\mathbf{X}}, \hat{\mathcal{T}}_{\mathbf{X}})$ can be calculated in linear time. We also note that the bound obtained can be further simplified with the use of, for instance, Gaussian, exponential or Laplacian kernels, to calculate the kernel functions $K_{\mathcal{S}}$ and $K_{\mathcal{T}}$, as these have 1s on the diagonal, thus facilitating the calculation of the trace. Finally, it can be seen that the bound from Theorem 36 has the same terms as Theorem 10, while the MMD distance is estimated as in Corollary 14.

Semi-supervised case Similar to the case considered by [Ben-David et al., 2010a], we can also derive similar bounds for the MMD distance in the case of combined error. To this end, we present the following analog of Theorem 11.

Theorem 37. *With the assumptions from Lemma 35, let S_u, T_u be unlabeled samples of size m', each drawn independently from $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$, respectively. Let S be a labeled sample of size m generated by drawing $\beta\, m$ points from $\mathcal{T}_{\mathbf{X}}$ ($\beta \in [0,1]$) and $(1-\beta)\, m$ points from $\mathcal{S}_{\mathbf{X}}$, and labeling them according to $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$, respectively. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\mathrm{R}^{\alpha}(h)$ on S and $h^*_T = \operatorname*{argmin}_{h\in\mathcal{H}} \mathrm{R}^{\ell_q}_{\mathcal{T}}(h)$, then for any $\delta \in (0,1)$, with probability of at least $1-\delta$ (over the choice of samples),*

$$\mathrm{R}^{\ell_q}_{\mathcal{T}}(\hat{h}) \;\leq\; \mathrm{R}^{\ell_q}_{\mathcal{T}}(h^*_T) + c_1 + c_2,$$

$$c_1 = 2\sqrt{\frac{2K\left(\frac{(1-\alpha)^2}{1-\beta} + \frac{\alpha^2}{\beta}\right)\log\frac{2}{\delta}}{m}} + 2\left(\sqrt{\frac{\alpha}{\beta}} + \sqrt{\frac{1-\alpha}{1-\beta}}\right)\sqrt{\frac{K}{m}},$$

$$c_2 = \hat{d}_{MMD}(S_u, T_u) + \frac{2}{m'}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbf{X}}}\sqrt{\mathrm{tr}(K_{\mathcal{S}})} + \frac{2}{m'}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbf{X}}}\sqrt{\mathrm{tr}(K_{\mathcal{T}})} + 2\sqrt{\frac{\log\frac{2}{\delta}}{2m'}} + \lambda.$$

Several observations can be made from this theorem. First of all, the main quantities that define the potential success of domain adaptation according to [Ben-David et al., 2010a] (*i.e.*, the distance between the distributions and the combined error of the joint ideal hypothesis) are preserved in the bound. This is an important point that indicates that the two results are not contradictory or supplementary. Secondly, rewriting the approximation of the bound as a function of α and omitting additive constants can lead to a similar result as for Theorem 11. This observation might indicate the existence of a strong connection between these.

The generalization guarantees obtained for domain adaptation based on the MMD distance allow another step forward to be made in domain adaptation theory, and the results presented in the previous sections to be extended in two different ways. Similar to discrepancy-based results, the bounds with the MMD distance allow any arbitrary loss function to be considered, and thus applications of domain adaptation other than binary classification can be studied. On the other hand, similar to the entropic-regularized Wasserstein distance, the MMD distance has some very useful estimation guarantees that are unavailable for both the $\mathcal{H}\Delta\mathcal{H}$ and $disc$ divergences. This feature can be very important in accessing both the *a-priori* hardness of adaptation and its *a-posteriori* success, to understand whether a given adaptation algorithm manages to correctly reduce the discrepancy between the domains.

5.5 Relationship between the Wasserstein and the the MMD distances

Here, we have just presented two results that introduced the Wasserstein and the MMD distances to the domain adaptation generalization bounds for both semi-supervised and unsupervised cases. As both results are built on the same construction, there might be the need to explore the link between the Wasserstein and the MMD distances. To do this, we first observe that in some particular cases, the latter can be bounded by the former. Indeed, if we assume that the

ground metric in the Wasserstein distance is $c(\mathbf{x}, \mathbf{x}') = \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_{\mathcal{H}}$, then the following results can be obtained:

$$\begin{aligned}
\left\| \int_{\mathbf{X}} f d(\mathcal{S}_{\mathbf{X}} - \mathcal{T}_{\mathbf{X}}) \|_{\mathcal{H}} &= \left\| \int_{\mathbf{X}\times\mathbf{X}} (f(\mathbf{x}) - f(\mathbf{x}')) d\gamma(\mathbf{x}, \mathbf{x}') \right\|_{\mathcal{H}} \\
&\leq \int_{\mathbf{X}\times\mathbf{X}} \|f(\mathbf{x}) - f(\mathbf{x}')\|_{\mathcal{H}} d\gamma(\mathbf{x}, \mathbf{x}') \\
&= \int_{\mathbf{X}\times\mathbf{X}} \| \langle f(\mathbf{x}), \phi(\mathbf{x}) \rangle - \langle f(\mathbf{x}'), \phi(\mathbf{x}') \rangle \|_{\mathcal{H}} d\gamma(\mathbf{x}, \mathbf{x}') \\
&\leq \|f\|_{\mathcal{H}} \int_{\mathbf{X}\times\mathbf{X}} \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_{\mathcal{H}} d\gamma(\mathbf{x}, \mathbf{x}').
\end{aligned}$$

Now taking the supremum over f *w.r.t.* $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$, and the *infimum* over $\gamma \in \Pi(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$, this gives

$$d_{\text{MMD}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \leq W_1(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}). \tag{8}$$

This result holds under the hypothesis that $c(\mathbf{x}, \mathbf{x}') = \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_{\mathcal{H}}$. On the other hand, in [Gao and Galvao, 2014], the authors showed that $W_1(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ with this particular ground metric can be further bounded, as follows

$$W_1(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \leq \sqrt{d^2_{\text{MMD}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + C},$$

where $C = \|\mu[\mathcal{S}_{\mathbf{X}}]\|_{\mathcal{H}} + \|\mu[\mathcal{T}_{\mathbf{X}}]\|_{\mathcal{H}}$. This result is quite strong for multiple reasons. First, it allows the squared MMD distance to be introduced to the domain adaptation bounds using [Redko et al., 2017, Lemma 1], which leads to the following result for two arbitrary hypotheses $(h, h') \in \mathcal{H}^2$

$$\mathrm{R}_{\mathcal{T}}(h, h') \leq \mathrm{R}_{\mathcal{S}}(h, h') + \sqrt{d^2_{\text{MMD}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + C}.$$

On the other hand, the unified inequality

$$d_{\text{MMD}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \leq W_1(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \leq \sqrt{d^2_{\text{MMD}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \|\mu[\mathcal{S}_{\mathbf{X}}]\|_{\mathcal{H}} + \|\mu[\mathcal{T}_{\mathbf{X}}]\|_{\mathcal{H}}} \tag{9}$$

suggests that the MMD distance establishes an interval bound for the Wasserstein distance. This point is very interesting, because originally the calculation of the Wasserstein distance (also known as the Earth Mover's distance) requires the solving of a linear programming problem that can be quite time consuming due to the computational complexity of $\mathcal{O}(n^3 \log(n))$, where n is the number of instances.

This result, however, is true only under the assumption that $c(\mathbf{x}, \mathbf{x}') = \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_{\mathcal{H}}$. While in most applications, the Euclidean distance $c(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|$ is used as a ground metric, this assumption can represent an important constraint. Luckily, it can be circumvented due to the duality between the RKHS-based and distance-based metric representations studied by [Sejdinovic et al., 2013]). Let us first rewrite the ground metric as

$$\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_{\mathcal{H}} = \sqrt{\langle \phi(\mathbf{x}) - \phi(\mathbf{x}'), \phi(\mathbf{x}) - \phi(\mathbf{x}') \rangle_{\mathcal{H}}} = \sqrt{k(\mathbf{x}, \mathbf{x}) - 2k(\mathbf{x}, \mathbf{x}') + k(\mathbf{x}', \mathbf{x}')}.$$

Now, to obtain the standard Euclidean distance in the expression of the ground metric, we can pick a kernel given by the covariance function of the fractional Brownian motion, *i.e.*, $k(\mathbf{x}, \mathbf{x}') = \frac{1}{2}(\|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2 - 2\|\mathbf{x} - \mathbf{x}'\|^2)$. Inserting this expression into the definition of $c(\mathbf{x}, \mathbf{x}')$ gives the desired Euclidean distance, and thus allows the Wasserstein distance to be calculated with the standard ground metric.

5.6 Other relevant contributions

[Zhang et al., 2019] In this work, the authors generalized the seminal bounds to the multi-class setting, and introduced a classification margin $\beta > 0$ into their results. This was done by introducing a definition of the error function $\mathrm{R}^{\beta}_{\mathcal{D}}$ that takes into account the classification margin, as follows:

$$\mathrm{R}^{\beta}_{\mathcal{D}} = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}} [l^{\beta}(h(\mathbf{x}), f_{\mathcal{D}}(\mathbf{x}))],$$

wherel^{β} is the ramp loss ([Shalev-Shwartz and Ben-David, 2014, Section 15.2.3]), defined as:

$$l^{\beta}(t) := \begin{cases} 1 - \frac{t}{\beta}, & \text{if } 0 \leq t \leq \beta \\ [t < 0], & \text{otherwise} \end{cases} \tag{10}$$

Their main contribution for the case of binary classification with labels encoded in $\{-1, 1\}$ can then be stated as follows:

$$\mathrm{R}_{\mathcal{T}}(h) \leq \mathrm{R}^{\beta}_{\mathcal{S}}(h) + \sup_{h' \in \mathcal{H}} \left| \mathrm{R}^{\beta}_{\mathcal{S}}(\mathrm{sgn}\,(h), h') - \mathrm{R}^{\beta}_{\mathcal{T}}(\mathrm{sgn}\,(h), h') \right| + \lambda^{(\beta)}, \tag{11}$$

where

$$\lambda^{(\beta)} = \inf_{h \in \mathcal{H}} \mathrm{R}^{\beta}_{\mathcal{S}}(h) + \mathrm{R}^{\beta}_{\mathcal{T}}(h).$$

The alignment term in Equation (11) was termed the margin disparity discrepancy. As can be noted, this involves a supremum over one hypothesis instead of two, making it lower than $\mathcal{H}\Delta\mathcal{H}$-divergence defined previously, which corresponds to the case of $\beta = 0$ with the definition of the error given above. This also offers new insights into the 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.*, $\epsilon^{0,0}_{\mathcal{T}}(h, f)$, their bound does not indicate the behavior of the margin violation rate on this latter. For $\lambda^{(\beta)}$, this remains conceptually similar to the λ term of the other bounds, with the only difference consisting in the definition of the error terms.

[Dhouib et al., 2020b] This work provides a generalization bound using a translated version of the ramp loss given in Equation (10) and defined as $l^{\rho,\beta} := l^{\beta}(\cdot - \rho)$ for some $\rho > 0$. The authors first prove a bound that is analogous to Equation (11), but concerning the margin violation loss $\mathrm{R}^{\rho,0}_{\mathcal{T}}(h)$ on the target domain, as follows:

$$\mathrm{R}^{\rho,0}_{\mathcal{T}}(h) \leq \mathrm{R}^{\frac{\rho+\beta}{\alpha},0}_{\mathcal{S}}(h) + \sup_{h' \in \mathcal{H}'} \left| \mathrm{R}^{\rho,\beta}_{\mathcal{S}}(h, h') - \mathrm{R}^{\rho,\beta}_{\mathcal{T}}(h, h') \right| + \lambda^{(\alpha)}, \tag{12}$$

where

$$\lambda^{(\alpha)} = \inf_{h \in \mathcal{H}'} \mathrm{R}_{\mathcal{S}}(h) + \mathrm{R}_{\mathcal{T}}(h) + \Pr_{\mathbf{x} \sim \mathcal{S}_{\mathbf{X}}} [|h(\mathbf{x})| < \alpha].$$

Compared to the bound from Equation (11), this bound is more informative on the separation quality between classes in the target domain, assessed by the margin violation risk $\mathrm{R}^{\rho,0}_{\mathcal{T}}(h)$. Also, the divergence term is continuous in both h and h' for $\beta > 0$, which makes it more suitable for optimization algorithms. The non estimable term $\lambda^{(\alpha)}$ is non symmetric w.r.t to $\mathcal{T}$ and $\mathcal{S}$ as it involves an absolute margin violation risk only for $\mathcal{S}_{\mathbf{X}}$. Finally, hypothesis space $\mathcal{H}'$ used to define the divergence and the $\lambda^{(\alpha)}$ term on the one hand, and the one concerning h, *i.e.*$\mathcal{H}$, are not necessarily equal.

[Shen et al., 2018, Courty et al., 2017] Several studies have presented generalization bounds for domain adaptation based on the Wasserstein distance, similar to those presented in this section. To this end, [Shen et al., 2018] gave a learning bound with the exact same form as the bound in Theorem 33, but without imposing any additional assumptions on the ground metric used in the definition of the Wasserstein distance. On the other hand, [Courty et al., 2017] proposed a learning bound for an adaptation scenario between joint source and target probability distributions $\mathcal{S}$ and $\mathcal{T}$, similar to that of [Zhang et al., 2012]. Their bound introduced $W(\mathcal{S}, \mathcal{T})$ with an additional term related to the probabilistic transfer Lipschitzness assumption introduced in the latter paper for the labeling function with respect to the optimal coupling. Also, the work of [Dhouib et al., 2020b] mentioned above proposed a generalization DA bound with an adversarial (minimax) version of the Wasserstein distance between the marginal distributions analyzed extensively in [Dhouib et al., 2020a].

Finally, we also note that the study of [Johansson et al., 2019] mentioned in the previous section also introduces learning bounds for domain adaptation based on the concept of IPM.

5.7 Summary

In this section, we presented several theoretical results that use IPMs as a measure of divergence between the marginal source and the target domain distributions in the domain adaptation generalization bounds. We argued that this particular choice of a distance provides a number of advantages compared to the $\mathcal{H}\Delta\mathcal{H}$-distance and the discrepancy distances considered before. First, both the Wasserstein distance and the MMD distance can be calculated from available finite samples in a computationally attractive way, due to linear time estimators for their entropy-regularized and quadratic versions, respectively. Secondly, the Wasserstein distance allows geometrical information to be taken into account when calculating the divergence between the two domain distributions, while the MMD distance is calculated based on the distance between the embeddings of two distributions in some (possibly) richer space. This feature is relatively interesting, as it provides more flexibility when it comes to incorporating the prior knowledge into the domain adaptation problem on the one hand, and allows a potentially richer characterization of the divergence between the domains, on the other. This might explain the abundance of domain adaptation algorithms based on the MMD distance, and some recent

domain adaptation techniques developed based on optimal transportation theory. Finally, we note that in general, the presented bounds are similar in shape to those described in Section 3, and they preserve their main terms, thus remaining consistent with these. This shows that despite the large variety of ways that can be used to formally characterize the generalization phenomenon in domain adaptation, the intuition behind this process and the main factors defining its potential success remain the same.

6 PAC-Bayesian theory for domain adaptation

In this section, we recall the results from [Germain et al., 2016, Germain et al., 2013, Germain et al., 2020], where PAC-Bayesian theory was used to theoretically understand domain adaptation through the weighted majority vote learning point of view.

6.1 Problem set-up

In the traditional PAC-Bayesian setting, we consider a π distribution over the hypothesis set $\mathcal{H}$, and the objective is to learn a ρ distribution over $\mathcal{H}$, by taking into account the information captured by the learning sample S. In the domain adaptation setting, the goal is different, and it consists of learning the ρ-weighted majority vote

$$\forall \mathbf{x} \in \mathbf{X}, \quad B_\rho(\mathbf{x}) = \operatorname{sign}\left[\mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x})\right],$$

, with the best performance on the target domain $\mathcal{T}$. Note that, here, we consider the $0-1$ loss function. As in the nonadaptation setting, PAC-Bayesian domain adaptation generalization bounds do not directly upper-bound $\mathrm{R}_{\mathcal{T}}^{\ell_{01}}(B_\rho)$, but upper-bound the expectation according to ρ of the individual risks of the functions from $\mathcal{H}$: $\mathbf{E}_{h\sim\rho}\,\mathrm{R}^{\ell_{01}}(h)$, which is closely related to B_ρ (see Equation (2)). Let us introduce a tight relation between $\mathrm{R}_{\mathcal{D}}(B_\rho)$ and $\mathbf{E}_{h\sim\rho}\,\mathrm{R}^{\ell_{01}}(h)$, known as the C-bound [Lacasse et al., 2006], and defined for all distribution $\mathcal{D}$ on $\mathbf{X}\times Y$ as

$$\mathrm{R}_{\mathcal{D}}^{\ell_{01}}(B_\rho) \;\le\; 1 - \frac{\left(1 - 2\mathop{\mathbf{E}}_{h\sim\rho}\mathrm{R}_{\mathcal{D}}^{\ell_{01}}(h)\right)^2}{1 - 2d_{\mathcal{D}_{\mathbf{X}}}(\rho)}. \tag{13}$$

where

$$d_{\mathcal{D}_{\mathbf{X}}}(\rho) \;=\; \mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\;\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}} \ell_{01}\big(h(\mathbf{x}), h'(\mathbf{x})\big)$$

is the expected disagreement between pairs of voters on the marginal distribution $\mathcal{D}_{\mathbf{X}}$. It is important to highlight that the expected disagreement $d_{\mathcal{D}_{\mathbf{X}}}(\rho)$ is closely related to the concept of expected joint error $e_{\mathcal{D}}(\rho)$ between pairs of voters:

$$e_{\mathcal{D}}(\rho) \;=\; \mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\;\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \ell_{01}\big(h(\mathbf{x}), y)\big) \times \ell_{0-1}\big(h'(\mathbf{x}), y)\big).$$

Indeed, for all distribution $\mathcal{D}$ on $\mathbf{X}\times Y$, we have

$$\mathop{\mathbf{E}}_{h\sim\rho}\mathrm{R}_{\mathcal{D}}^{\ell_{01}}(h) \;=\; \frac{1}{2}d_{\mathcal{D}_{\mathbf{X}}}(\rho) + e_{\mathcal{D}}(\rho). \tag{14}$$

In the following, we present the two PAC-Bayesian generalization bounds for domain adaptation presented in [Germain et al., 2013, Germain et al., 2016], through the point of view of [Catoni, 2007].

6.2 In the spirit of Ben-David et al. and Mansour et al.

The authors of [Germain et al., 2013] proposed to define a divergence measure that follows the idea underlying the C-bound of Equation (13). More precisely, if $\mathbf{E}_{h\sim\rho}\,\mathrm{R}_{\mathcal{S}}^{\ell_{01}}(h)$ and $\mathbf{E}_{h\sim\rho}\,\mathrm{R}_{\mathcal{T}}^{\ell_{01}}(h)$ are similar, then $\mathrm{R}_{\mathcal{S}}^{\ell_{01}}(B_\rho)$ and $\mathrm{R}_{\mathcal{T}}^{\ell_{01}}(B_\rho)$ are similar when $d_{\mathcal{S}_{\mathbf{X}}}(\rho)$ and $d_{\mathcal{T}_{\mathbf{X}}}(\rho)$ are also similar. Thus, the domains $\mathcal{S}$ and $\mathcal{T}$ are close according to ρ if the expected disagreement over the two domains tends to be close. This intuition led the authors to the following domain disagreement pseudometric.

Definition 22 (Domain disagreement [Germain et al., 2013])**.** *Let $\mathcal{H}$ be a hypothesis class. For any marginal distributions $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$ over $\mathbf{X}$, and any distribution ρ on $\mathcal{H}$, the domain disagreement $\mathrm{dis}_\rho(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ between $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$ is defined by*

$$\mathrm{dis}_\rho(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \;=\; \Big|\, d_{\mathcal{T}_{\mathbf{X}}}(\rho) - d_{\mathcal{S}_{\mathbf{X}}}(\rho)\,\Big|.$$

It is worth noting that the value of $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X})$ is always lower than the $\mathcal{H}\Delta\mathcal{H}$-distance between $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$. Indeed, for every $\mathcal{H}$ and ρ over $\mathcal{H}$, we have

$$\begin{aligned}
\tfrac{1}{2}\, d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}) &= \sup_{(h,h')\in\mathcal{H}^2} \left| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_\mathbf{X}} \ell_{01}\big(h(\mathbf{x}), h'(\mathbf{x})\big) - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}} \ell_{01}\big(h(\mathbf{x}), h'(\mathbf{x})\big) \right| \\
&\geq \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_\mathbf{X}} \ell_{01}\big(h(\mathbf{x}), h'(\mathbf{x})\big) - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}} \ell_{01}\big(h(\mathbf{x}), h'(\mathbf{x})\big) \right| \\
&\geq \Big| d_{\mathcal{T}_\mathbf{X}}(\rho) - d_{\mathcal{S}_\mathbf{X}}(\rho) \Big| \\
&= \mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}).
\end{aligned}$$

Using this domain divergence, the authors proved the following domain adaptation bound.

Theorem 38 ([Germain et al., 2013])**.** *Let $\mathcal{H}$ be a hypothesis class. We have*

$$\forall \rho \text{ on } \mathcal{H},\ \mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^{\ell_{01}}_{\mathcal{T}}(h) \ \leq\ \mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^{\ell_{01}}_{\mathcal{S}}(h) + \frac{1}{2}\,\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}) + \lambda_\rho\,,$$

where λ_ρ is the deviation between the expected joint errors between pairs for voters on the target and source domains, defined as

$$\lambda_\rho \ =\ \Big| e_{\mathcal{T}}(\rho) - e_{\mathcal{S}}(\rho) \Big|. \tag{15}$$

The above theorem can be used to prove different kinds of PAC-Bayesian generalization bounds. Below, we present only one such generalization bound, which was used to derive an adaptation algorithm in [Germain et al., 2013].

Theorem 39. *For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X}\times Y$, any set of voters $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any $\delta\in(0,1]$, any real numbers $\omega>0$ and $a>0$, with a probability of at least $1-\delta$ over the random choice of $S\times T_u\sim(\mathcal{S}\times\mathcal{T}_\mathbf{X})^m$, for every posterior distribution ρ on $\mathcal{H}$, we have*

$$\begin{aligned}
\mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^{\ell_{01}}_{\mathcal{T}}(h) \ \leq\ & \omega' \mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^{\ell_{01}}_{S}(h) + a'\,\tfrac{1}{2}\,\mathrm{dis}_\rho(S, T_u) \\
& + \left(\frac{\omega'}{\omega} + \frac{a'}{a}\right) \frac{\mathrm{KL}(\rho|\pi) + \ln\frac{3}{\delta}}{m} + \lambda_\rho + \tfrac{1}{2}(a'-1)\,,
\end{aligned}$$

where $\mathrm{dis}_\rho(S, T_u)$ is the empirical estimate of the the domain disagreement; λ_ρ is defined by Equation (15); $\omega' = \frac{\omega}{1-e^{-\omega}}$ and $a' = \frac{2a}{1-e^{-2a}}$.

Similarly to the bounds of Theorems 6 and 15, this bound can be seen as a trade-off between different quantities. The terms $\mathbf{E}_{h\sim\rho}\,\mathrm{R}^{\ell_{01}}_S(h)$ and $\mathrm{dis}_\rho(S,T)$ are akin to the first two terms of the bound of Theorem 6: $\mathbf{E}_{h\sim\rho}\,\mathrm{R}^{\ell_{01}}_S(h)$ is the ρ-average risk over $\mathcal{H}$ on the source sample, and $\mathrm{dis}_\rho(S,T_u)$ measures the ρ-average disagreement between the marginals, although it is specific to the current model depending on ρ. The last term λ_ρ measures the deviation between the expected joint target and source errors of the individual hypothesis from $\mathcal{H}$ (according to ρ). A successful domain adaptation is possible if this deviation is low, although when no labels in the target sample are available, this term cannot be controlled or estimated.

Despite the same underlying philosophy, the authors note that this bound is in general incomparable with those ones of Theorems 6 and 15 due to the dependence of $\mathrm{dis}_\rho(S,T)$ and λ_ρ on the learned posterior.

6.3 A different philosophy

In [Germain et al., 2016], the authors introduce another domain divergence to provide an original bound for the PAC-Bayesian setting. They take advantage of Equation (14), which expresses the risk of the Gibbs classifier in terms of two quantities:

$$\mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^{\ell_{01}}_{\mathcal{T}}(h) \ =\ \tfrac{1}{2} d_{\mathcal{T}_\mathbf{X}}(\rho) + e_{\mathcal{T}}(\rho) \tag{16}$$

It can be noted that the latter expression consists of half of the expected disagreement, which does not require labeled data to be estimated, and the inestimable expected joint error. To deal with the latter, the authors designed a divergence to link $e_{\mathcal{T}}(\rho)$ to $e_{\mathcal{S}}(\rho)$, called the β-divergence, which is defined by

$$\forall q>0, \quad \beta_q \ =\ \left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left(\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}\right)^q\right]^{\frac{1}{q}}. \tag{17}$$

The β-divergence is parametrized by the value of $q > 0$, and allows well-known distribution divergence to be recovered, such as the χ^2-distance and the Rényi divergence mentioned at the end of Section 3. When $q \to \infty$, we have

$$\beta_\infty \;=\; \sup_{(\mathbf{x},y)\in \text{SUPP}(\mathcal{S})} \left(\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}\right), \tag{18}$$

where $\text{SUPP}(\mathcal{S})$ denotes the support of the domain $\mathcal{S}$. This β-divergence leads to the following bound.

Theorem 40 ([Germain et al., 2016])**.** *Let $\mathcal{H}$ be a hypothesis space, $\mathcal{S}$ and $\mathcal{T}$ be the source and target domains on $\mathbf{X} \times Y$, and $q > 0$ be some positive constant. Then, for all posterior distributions ρ on $\mathcal{H}$, we have*

$$\mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^{\ell_{01}}_{\mathcal{T}}(h) \;\le\; \frac{1}{2}\, d_{\mathcal{T}_\mathbf{X}}(\rho) + \beta_q \times \Big[e_{\mathcal{S}}(\rho)\Big]^{1-\frac{1}{q}} + \eta_{\mathcal{T}\setminus\mathcal{S}},$$

where

$$\eta_{\mathcal{T}\setminus\mathcal{S}} = \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{T}} \Big((\mathbf{x},y) \notin \text{SUPP}(\mathcal{S})\Big) \sup_{h\in\mathcal{H}} \mathrm{R}_{\mathcal{T}\setminus\mathcal{S}}(h)$$

with $\mathcal{T}\setminus\mathcal{S}$ the distribution of $(\mathbf{x},y)\sim\mathcal{T}$ conditional to $(\mathbf{x},y) \in \text{SUPP}(\mathcal{T})\setminus\text{SUPP}(\mathcal{S})$.

The last term of the bound, $\eta_{\mathcal{T}\setminus\mathcal{S}}$, which cannot be estimated without target labels, captures the worst possible risk for the target area not included in $\text{SUPP}(\mathcal{S})$, similar to the idea used by [Johansson et al., 2019]. Note that we have

$$\eta_{\mathcal{T}\setminus\mathcal{S}} \le \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{T}} \big((\mathbf{x},y) \notin \text{SUPP}(\mathcal{S})\big).$$

An interesting property of Theorem 40 is that when domain adaptation is not required (*i.e.*, $\mathcal{S} = \mathcal{T}$), the bound is still sound and nondegenerate. Indeed, in this case we have

$$\mathrm{R}_{\mathcal{S}}(G_\rho) \;=\; \mathrm{R}_{\mathcal{T}}(G_\rho) \;\le\; \tfrac{1}{2}\, d_{\mathcal{T}_\mathbf{X}}(\rho) + 1 \times [e_{\mathcal{S}}(\rho)]^1 + 0 = \;\tfrac{1}{2}\, d_{\mathcal{S}_\mathbf{X}}(\rho) + e_{\mathcal{S}}(\rho) \;=\; \mathrm{R}_{\mathcal{S}}(G_\rho)\,.$$

Below, we present the PAC-Bayesian generalization bound obtained from the above theorem for the case $q\to\infty$.

Theorem 41. *For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X} \times Y$, any set of voters $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any $\delta \in (0,1]$, any real numbers $b > 0$ and $c > 0$, with a probability of at least $1-\delta$ over the random choices of $S \sim (\mathcal{S})^{m_S}$ and $T_u \sim (\mathcal{T}_\mathbf{X})^{m_T}$, for every posterior distribution ρ on $\mathcal{H}$, we have*

$$\mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^{\ell_{01}}_{\mathcal{T}}(h) \;\le c'\, \tfrac{1}{2}\, d_T(\rho) + b'\, e_S(\rho) + \eta_{\mathcal{T}\setminus\mathcal{S}} + \left(\frac{c'}{m_T \times c} + \frac{b'}{m_S \times b}\right)\left(2\ \mathrm{KL}(\rho|\pi) + \ln\tfrac{2}{\delta}\right),$$

where $d_T(\rho)$ and $e_S(\rho)$ are the empirical estimations of the target voters' disagreement and the source joint error, and $b' = \frac{b}{1-e^{-b}}\,\beta_\infty$, and $c' = \frac{c}{1-e^{-c}}$.

Similarly to the first bound, the above theorem upper-bounds the target risk by a trade-off of different terms given by the following atypical quantities:

1. The expected disagreement $d_T(\rho)$ that captures second degree information about the target domain;
2. The divergence between the domains, captured by the β_q-divergence is not an additive term any more: it weights the influence of the expected joint source error $e_S(\rho)$ where the parameter q allows different instances of the β_q-divergence to be considered;
3. The term $\eta_{\mathcal{T}\setminus\mathcal{S}}$ quantifies the worst feasible target error on the regions where the source domain is not informative for the target task.

6.4 Comparison of the two domain adaptation bounds

The main difference between the bounds of Theorems 38 and 40 lies in the estimable terms that the latter relies on. In Theorem 40, the nonestimable terms are the β-divergence β_q and the term $\eta_{\mathcal{T}\setminus\mathcal{S}}$. Contrary to the noncontrollable term λ_ρ of Theorem 38, these terms do not depend on the *learned* posterior distribution ρ: for every ρ on $\mathcal{H}$, β_q and $\eta_{\mathcal{T}\setminus\mathcal{S}}$ are constant values that measure the relation between the domains for the considered task. Moreover, the β-divergence is not an additive term but a multiplicative one (as opposed to $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X},\mathcal{T}_\mathbf{X}) + \lambda_\rho$ in Theorem 38), which is an important contribution of this new perspective. This is similar to the studies of [Mansour et al., 2009b] and [Dhouib and Redko, 2018], who also introduced such a multiplicative dependence. Consequently, β_q can be viewed as a hyperparameter, which allows us to tune the trade-off between the target voters' disagreement $d_{\mathcal{T}_\mathbf{X}}(\rho)$ and the source joint error $e_{\mathcal{S}}(\rho)$.

Note that, when $e_{\mathcal{T}}(\rho) \geq e_{\mathcal{S}}(\rho)$, we can upper-bound the term λ_ρ of Theorem 38 by using the same trick as in the proof of Theorem 40. This leads to

$$e_{\mathcal{T}}(\rho) \geq e_{\mathcal{S}}(\rho) \quad \Longrightarrow \quad \lambda_\rho = e_{\mathcal{T}}(\rho) - e_{\mathcal{S}}(\rho) \leq \beta_q \times \left[e_{\mathcal{S}}(\rho)\right]^{1-\frac{1}{q}} + \eta_{\mathcal{T}\setminus\mathcal{S}} - e_{\mathcal{S}}(\rho).$$

Thus, in this particular case, we can rewrite the Theorem 38 statement for all ρ on $\mathcal{H}$, as

$$\mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^{\ell_{01}}_{\mathcal{T}}(h) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}^{\ell_{01}}_{\mathcal{S}}(h) + \frac{1}{2}\,\mathrm{dis}_\rho(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \beta_q \times \left[e_{\mathcal{S}}(\rho)\right]^{1-\frac{1}{q}} - e_{\mathcal{S}}(\rho) + \eta_{\mathcal{T}\setminus\mathcal{S}}.$$

It turns out that, if $d_{\mathcal{T}_{\mathbf{X}}}(\rho) \geq d_{\mathcal{S}_{\mathbf{X}}}(\rho)$ in addition to $e_{\mathcal{T}}(\rho) \geq e_{\mathcal{S}}(\rho)$, the above statement reduces to that of Theorem 40. In all other cases, Theorem 40 is tighter, thus confirming that following the seminal works of Section 3, with introduction of absolute values in Theorem 38, gives a very rough approximation. Finally, one of the key points of the generalization bounds of Theorems 39 and 41 is that they suggest algorithms for tackling majority vote learning in the domain adaptation context. Similar to what was done in traditional supervised learning [Langford and Shawe-Taylor, 2002, Ambroladze et al., 2006], [Germain et al., 2013, Germain et al., 2016, Germain et al., 2020] specialized these theorems to linear classifiers, and derived adaptation algorithms based on this specialization.

6.5 Other relevant contributions

[McNamara and Balcan, 2017] In this study, the authors made use of the PAC-Bayesian framework to derive a generalization bound for fine tuning in deep learning in a spirit close to that of analysing a domain adaptation problem. Their considered setting corresponded to a scenario where there is the need to adapt a network trained for a given domain to a similar one. The authors obtained a bound that does not directly involve the concept of divergence between the domains, but a function that measures a transferability property between the two domains.

6.6 Summary

In this section, we recalled the two domain adaptation analyses for the PAC-Bayesian framework presented in [Germain et al., 2013, Germain et al., 2016, Germain et al., 2020] for models taking the form of a majority vote over a set of classifiers. More precisely, the first result of this section follows the underlying philosophy of the seminal works of Ben-David *et al.* and Mansour *et al.* of Section 3, by upper-bounding the target risk by a source risk and a domain divergence measure suitable for the PAC-Bayesian setting. This divergence is expressed as the average deviation between the disagreement over a set of classifiers on the source and target domains, contrary to $\mathcal{H}\Delta\mathcal{H}$-divergence and discrepancy distance, which are defined in terms of the worst-case deviation. Then, we recalled another domain adaptation bound that takes advantage of the inherent behavior of the target risk in the PAC-Bayesian setting. The upper bound obtained is different from the original one, as it expresses a trade-off between the disagreement on the target domain only, the joint errors of the classifiers on the source domain only, and a term that reflects the worst-case error in regions where the source domain is noninformative. Contrary to the first bound and those of the previous sections, the divergence is not an additive term, but is a factor that weights the importance of the source information. These analyses were combined with PAC-Bayesian generalization bounds of Section 2, and involved an additional term that measures the deviation of the learned majority vote to the *a-priori* knowledge we have on the majority vote.

7 Domain adaptation theory based on algorithmic properties

In this section, we first review the work of [Mansour and Schain, 2014], where they derived a domain adaptation generalization bound in terms of the algorithmic robustness of [Xu and Mannor, 2010] recalled in Section 2. Then, we present the works of [Kuzborskij and Orabona, 2013] based on a closely related concept of algorithmic stability. Note that this last contribution is proved for a setting different from the domain adaptation problem considered so far, as in this case there is no access to the source examples, but rather to a hypothesis learned from them.

7.1 Robust domain adaptation

Definition of λ-shift [Mansour and Schain, 2014] used the concept of algorithmic robustness [Xu and Mannor, 2010] to define the λ-shift that encodes prior knowledge of the deviation between the source and target domains. The goal of their definition was to capture the proximity of the loss associated to a hypothesis on the source and target domains in the regions defined by partitioning the joint space $\mathbf{X} \times Y$. As there is usually no access to target labels, the authors proposed to consider the conditional distribution of the label in a given region, and the relation to its sampled value over the given labeled sample S. To proceed, let ρ be a distribution over the label space Y, and let σ^y and $\sigma^{-y} = 1 - \sigma^y$ denote the probability of a given label $y \in Y$ and the total probability of the other labels, respectively. The definition of the λ-shift is then given as follows.

Definition 23 ([Mansour and Schain, 2014]). *Let σ and ρ be two distributions over Y. ρ is the λ-shift with respect to σ, denoted by $\rho \in \lambda(\sigma)$, if for all $y \in Y$ we have $\rho^y \leq \sigma^y + \lambda\sigma^{-y}$ and $\rho^y \geq \sigma^y(1-\lambda)$. If for some $y \in Y$ we have $\rho^y = \sigma^y + \lambda\sigma^{-y}$, we say that ρ is strict-λ-shift with respect to σ.*

Note that, for the sake of simplicity, for $\rho \in \lambda(\sigma)$, the upper bound and the lower bound of the probability ρ^y are respectively denoted by:

$$\bar{\lambda}^y(\sigma) = \sigma^y + \lambda(1-\sigma^y), \qquad \text{and} \qquad \underline{\lambda}^y(\sigma) = \sigma^y(1-\lambda)\,.$$

The above definition means that λ-shift between two distributions on Y implies a restriction on the deviation between the probability of a label on the distributions: this shift might be at most a λ portion of the probability of the other labels or of the probability of the label. Note that $\lambda = 1$, respectively $\lambda = 0$, corresponds to the no restriction and the total restriction cases, respectively.

Learning bounds based on algorithmic robustness To analyze the domain adaptation setting, the authors assumed that $\mathbf{X} \times Y$ can be partitioned into M disjoint subsets, defined as $\mathbf{X} \times Y = \bigcup_{i,j} \mathbf{X}_i \times Y_j$, where the input space is partitioned as $\mathbf{X} = \bigcup_{i=1}^{M_\mathbf{X}}$, and the output space as $Y = \bigcup_{j=1}^{M_Y} Y_j$ and $M = M_\mathbf{X} M_Y$. Note that, an (M, ϵ)-robust algorithm outputs a hypothesis that has an ϵ variation in the loss in each region $\mathbf{X}_i \times Y_j$. We now present the following theorem.

Theorem 42 ([Mansour and Schain, 2014]). *Let $\mathcal{A}$ be an (M, ϵ)-robust algorithm with respect to a loss function $\ell : \mathbf{X} \times Y$, such that $0 \leq \ell(h(\mathbf{x}, y) \leq M_l$, for all $(\mathbf{x}, y) \in (\mathbf{X} \times Y)$ and $h \in \mathcal{H}$. If $\mathcal{S}$ is λ-shift of $\mathcal{T}$ with respect to the partition of $\mathbf{X}$ for any $\delta \in (0, 1]$, the following bound holds with probability of at least $1 - \delta$, over the random draw of the sample S from $\mathcal{S}$, and of the sample T from $\mathcal{T}$ of size m,*

$$\forall h \in \mathcal{H},\ \mathrm{R}^\ell_\mathcal{T}(h) \ \leq \ \sum_{i=1}^{M_\mathbf{x}} T(\mathbf{X}_i)\ell^\lambda_S(h, \mathbf{X}_i) + \epsilon + M_\ell\sqrt{\frac{2M\ln 2 + 2\ln\frac{1}{\delta}}{m}}\,,$$

where $T(\mathbf{X}_i) = \frac{1}{m}\big|\{\mathbf{x} \in T \cap \mathbf{X}_i\}\big|$ is the ratio of target points in the region $\mathbf{X}_i$, and

$$\forall i \in \{1, \ldots, M_\mathbf{X}\}, \quad \ell^\lambda_S(h, \mathbf{X}_i) \leq \max_{y \in Y}\left\{\ell_i(h, y)\bar{\lambda}^y(\mathcal{S}_i) + \sum_{y' \neq y}\ell_i(h, y')\underline{\lambda}^{y'}(\mathcal{S}_i)\right\},$$

with

$$\ell_i(h, y) = \begin{cases} \max_{\mathbf{x} \in S \cap \mathbf{X}_i \times y} \ell(h(\mathbf{x}), y) & \text{if } S \cap \mathbf{X}_i \times y \neq \emptyset \\ M_\ell & \text{otherwise.} \end{cases}$$

The main difference between this domain adaptation result and the original robustness bound of Theorem 9 of Section 2 is seen in the first term. In the latter case, which is an upper bound on the source risk, the first term $\frac{1}{m}\sum_{(x,y)\in S}\ell(h_S(x), y)$ simply corresponds to the empirical error of the model learned on the source sample. In the former bound, which upper-bounds the target risk, the first term $\sum_{i=1}^{M_\mathbf{x}} T(\mathbf{X}_i)\ell^\lambda_S(h, \mathbf{X}_i)$ depends also on the empirical risk on the source sample, which is a combination of the λ-shifted source risk of each region weighted by the ratio of target points in the region. This is reminiscent of the multiplicative dependence between the source error and the divergence term already mentioned in previous sections.

7.2 Hypothesis transfer learning

In this section we review theoretical results for the *hypothesis transfer learning (HTL)* setting where only a hypothesis learned in the source domain, and not the source (labeled) data, is available in addition to a *small* training sample from the target domain. As a direct consequence of this, HTL does not introduce any assumptions about the relatedness of the source and target distributions, and it has an advantage in that it avoids the need to store abundant source data.

More formally, let $h_{\text{src}} \in \mathcal{H}_\mathcal{S}$ be a hypothesis learned from labeled source data, and let $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim (\mathcal{T})^m$ be a (labeled) target sample. The goal of HTL is then to learn a target model using h_{src} and T that is better than the one we can learn from T only. This goal is formalized using the following definition of a HTL algorithm $\mathcal{A}$:

$$\mathcal{A} : (\mathbf{X} \times Y)^m \times \mathcal{H}_\mathcal{S} \mapsto \mathcal{H}\,,$$

where $\mathcal{A}$ maps any (labeled) target sample $T \sim (\mathcal{T})^m$ and a source hypothesis $h_{\text{src}} \in \mathcal{H}_\mathcal{S}$ onto a target hypothesis $h \in \mathcal{H}$. We now use this formalization to present several key definitions in HTL.

Definition 24 (Usefulness and Collaboration [Kuzborskij, 2018]). *A hypothesis $h_{src} \in \mathcal{H}_{\mathcal{S}}$ is useful for $\mathcal{A}$ with respect to the distribution $\mathcal{S}$ and a training sample S of size m if*

$$\mathop{\mathbf{E}}_{S\sim(\mathcal{S})^m}\left[\mathrm{R}_{\mathcal{D}}(\mathcal{A}(S,h_{src}))\right] < \mathop{\mathbf{E}}_{S\sim(\mathcal{S})^m}\left[\mathrm{R}_{\mathcal{D}}(\mathcal{A}(S,\mathbf{0}))\right].$$

A hypothesis $h_{src} \in \mathcal{H}_{\mathcal{S}}$ and a distribution $\mathcal{D}$ collaborate [Ben-David and Urner, 2013] for $\mathcal{A}$, with respect to a training sample S of size m, if

$$\mathop{\mathbf{E}}_{S\sim(\mathcal{S})^m}\left[\mathrm{R}_{\mathcal{S}}(\mathcal{A}(S,h_{src}))\right] < \min\left\{\mathrm{R}_{\mathcal{S}}(\mathcal{A}(\emptyset,h_{src}), \mathop{\mathbf{E}}_{S\sim(\mathcal{S})^m}\left[\mathrm{R}_{\mathcal{D}}(\mathcal{A}(S,\mathbf{0}))\right]\right\}.$$

This definition provides two interesting properties for a hypothesis transfer learning algorithm. The concept of usefulness corresponds to the case where the algorithm $\mathcal{A}$ allows a model to be inferred with a lower risk by using the source hypothesis. The collaboration refers to the case where the access to both the source hypothesis h_{src} and the sample S used together helps to increase the performance in comparison to the case where they are used separately. If any one of these two properties is not satisfied, then the resulting learning procedure leads to higher target error. The authors further analyzed a regularized least squares algorithm (RLS) for HTL, as presented below.

A biased RLS algorithm for HTL We first begin with a quick recap of the classic RLS algorithm. For a learning sample $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim (\mathcal{T})^m$ such that $y_i \in [-B, B]$ with $B \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbf{R}^d$ with $\|\mathbf{x}\| \leq 1$, the RLS algorithm aims to solve the following optimization problem:

$$\min_{\mathbf{w}\in\mathbf{R}^d}\left\{\frac{1}{m}\sum_{i=1}^m(\mathbf{w}^T\mathbf{x}_i - y_i)^2 + \lambda\|\mathbf{w}\|^2\right\}.$$

It is well-known that RLS has useful theoretical properties and its solution can be expressed in a closed form. Now, we consider a source hypothesis of the form $h_{\text{src}}(\mathbf{x}) = \mathbf{x}^T\mathbf{w}_0$, where $\mathbf{w}_0$ corresponds to the parameters of h_{src} in the same space as $\mathbf{w}$. In [Orabona et al., 2009], the authors suggested to use a biased regularization with respect to $\mathbf{w}_0$, as

$$\min_{\mathbf{w}\in\mathbf{R}^d}\left\{\frac{1}{m}\sum_{i=1}^m(\mathbf{w}^T\mathbf{x}_i - y_i)^2 + \lambda\|\mathbf{w} - \mathbf{w}_0\|^2\right\}.$$

In this formulation, we can see that the source hypothesis represented by $\mathbf{w}_0$ acts as a bias that tends to make the learned model closer to $\mathbf{w}_0$ if the learning sample is compatible with it. Following the result of [Kuzborskij and Orabona, 2013], we present a more general version, where the target hypothesis to be learned is defined by

$$h_T(\mathbf{x}) = \mathrm{tr}_C\left(\mathbf{x}^\top\hat{\mathbf{w}}_T\right) + h_{\text{src}}(\mathbf{x}), \tag{19}$$

where

$$\hat{\mathbf{w}}_T = \mathop{\mathrm{argmin}}_{\mathbf{w}}\frac{1}{m}\sum_{i=1}^m\left(\mathbf{w}^\top\mathbf{x}_i - y_i + h_{\text{src}}(\mathbf{x}_i)\right)^2 + \lambda\|\mathbf{w}\|^2,$$

and the truncation function $\mathrm{tr}_C(a)$ is defined as

$$\mathrm{tr}_C(a) = \min\left[\max\left(a, -C\right), C\right].$$

This formulation is a generalization of the usual biased RLS algorithm that allows consideration of any type of source model h_{src}. In particular, we can retrieve the usual formulation when $C = \infty$ and $h_{\text{src}}(\mathbf{x}) = \mathbf{x}^\top\mathbf{w}_0$, where $\mathbf{w}_0$ and $\mathbf{w}_T$ belong to the same space.

From the theoretical standpoint, the goal of the authors was then to bound the expected risk associated with this algorithm, in terms of the characteristics of the source model h_{src}. The proposed result is based upon the *leave-one-out* risk over a sample T, defined as

$$\mathrm{R}^{\text{loo}}_{\hat{T}}(\mathcal{A}, T) = \frac{1}{m}\sum_{i=1}^m\ell(\mathcal{A}_{T^{\setminus i}}, (\mathbf{x}_i, y_i)),$$

where $\mathcal{A}_{T^{\setminus i}}$ represents the model learned by algorithm $\mathcal{A}$ from sample T, without the example $(\mathbf{x}_i, y_i)$. The first result related to HTL can be now presented in the following theorem.

Theorem 43 ([Kuzborskij and Orabona, 2013]). *Set $\lambda \geq \frac{1}{m}$. If $C \geq B + \|h_{src}\|_\infty$, then for any hypothesis learned by the algorithm presented in Equation (19), with probability of at least $1-\delta$ over any sample T of size m i.i.d. from $\mathcal{T}$, we have*

$$\mathrm{R}_{\mathcal{T}}(h_T) - \mathrm{R}^{loo}_{\hat{\mathcal{T}}}(h_T, T) = \mathcal{O}\left(C \frac{\sqrt[4]{\mathrm{R}_{\mathcal{T}}(h_{src}) \mathrm{tr}_{C^2}\left(\frac{\mathrm{R}_{\mathcal{T}}(h_{src})}{\lambda}\right) + \mathrm{R}^2_{\mathcal{T}}(h_{src})}}{\sqrt{m}\delta\lambda^{3/4}} \right).$$

If $C = \infty$, then we have

$$\mathrm{R}_{\mathcal{T}}(h_T) - \mathrm{R}^{loo}_{\hat{\mathcal{T}}}(h_T, T) = \mathcal{O}\left(\frac{\sqrt{\mathrm{R}_{\mathcal{T}}(h_{src})}(\|h_{src}\|_\infty + B)}{\sqrt{m}\delta\lambda} \right).$$

According to [Kuzborskij, 2018], we can draw the following implications.

1. For the null source hypothesis, *i.e.*, $h_{\text{src}} = \mathbf{0}$, we fall into a classic supervised learning setting, while for $C = \infty$, the generalization bound is bounded by $\mathcal{O}\left(\frac{B}{\sqrt{m}\lambda}\right)$, similar to the results obtained for classic RLS algorithms [Bousquet and Elisseeff, 2002];
2. If $h_{\text{src}} \neq \mathbf{0}$ and $\frac{1}{\lambda}\mathrm{R}_{\mathcal{T}}(h_{\text{src}})$ tend to zero, then the target true risk converges to the leave-one-out risk. This means that when the source hypothesis is good enough on the target domain, then transfer learning helps to learn a better hypothesis on the target domain, even with small training samples.
3. If $\frac{1}{\lambda}\mathrm{R}(h_{\text{src}})$ is high, then more target labeled data are needed to provide a reliable hypothesis on the target. The domains are then considered to be unrelated, so the source hypothesis does not bring any useful information.

Multi-source scenario Here, we consider the setting of [Kuzborskij and Orabona, 2017], where the source hypothesis is expressed as a weighted combination of different source hypotheses

$$h^{\boldsymbol{\beta}}_{\text{src}}(\mathbf{x}) = \sum_{i=1}^{n} \beta_i h^i_{\text{src}}(\mathbf{x}),$$

and where the target hypothesis is defined as

$$h_{\boldsymbol{w},\boldsymbol{\beta}}(\mathbf{x}) = \langle \boldsymbol{w}, \mathbf{x} \rangle + h^{\boldsymbol{\beta}}_{\text{src}}(\mathbf{x}).$$

The relevance of the different source hypotheses is then characterized by their associated weight given by the vector $\boldsymbol{\beta}$.

Let $\ell : Y \times Y \to \mathbb{R}_+$ be an H-smooth loss function, such that $\forall y_1, y_2 \in Y,\ |\nabla_{y_1}\ell(y_1, y) - \nabla_{y_2}\ell(y_2, y)| \leq H|y_1 - y_2|$, and let $\Omega : \mathcal{H} \to \mathbb{R}_+$ be a σ-strongly convex function with respect to a norm $\|\cdot\|$ and to a hypothesis space $\mathcal{H}$. Given a target training set $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, $\lambda \in \mathbb{R}_+$, n source hypotheses $\{h^i_{\text{src}}\}_{i=1}^n$ and a parameter vector $\boldsymbol{\beta}$ verifying $\Omega(\boldsymbol{\beta}) \leq \rho$, the transfer algorithm generates a target hypothesis $h_{\hat{\boldsymbol{w}},\boldsymbol{\beta}}$ such that

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell(\langle \boldsymbol{w}, \mathbf{x}_i \rangle + h^{\boldsymbol{\beta}}_{\text{src}}(\mathbf{x}_i), y_i) + \lambda\Omega(\mathbf{w}) \right\}.$$

In this formulation, the loss function is only minimized with respect to $\mathbf{w}$, and not specifically with respect to $\boldsymbol{\beta}$. However, it is assumed that $\Omega(\boldsymbol{\beta}) \leq \rho$ makes $\boldsymbol{\beta}$ constrained by a strongly convex function, which allows regularized algorithms to be covered that consider an additional regularization with respect to $\boldsymbol{\beta}$. As in the previous analysis, the key quantity $\mathrm{R}_{\mathcal{T}}(h^{\boldsymbol{\beta}}_{\text{src}})$ that measures the relevance of the source hypothesis on the target domain will have a crucial role in the analysis of the generalization properties of $h_{\hat{\boldsymbol{w}},\boldsymbol{\beta}}$. To illustrate the types of algorithms covered by this analysis, we can consider the least-squares-based regularization that given source hypotheses $\{\mathbf{w}^i_{\text{src}}\} \subset \mathcal{H}$, the parameters $\boldsymbol{\beta} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}_+$ outputs the target hypothesis

$$h(\mathbf{x}) = \langle \hat{\boldsymbol{w}}, \mathbf{x} \rangle,$$

where

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^{m} (\langle \boldsymbol{w}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \|\mathbf{w} - \sum_{j=1}^{n} \beta_j \mathbf{w}^j_{\text{src}}\|_2^2 \right\}. \tag{20}$$

The problem defined by Equation (20) presents a special case of the classic regularized empirical risk minimization (ERM), and can be interpreted as the minimization of the empirical error on the target sample while keeping the solution

close to the (best) linear combination of source hypotheses. Note that while such a formulation is limited to a linear combination of the source hypotheses in the same space as the target predictor, it can be generalized by allowing the source hypotheses to be treated as "black box" predictors. The results presented below correspond to generalization bounds for such an RLS multi-source algorithm.

Theorem 44 ([Kuzborskij and Orabona, 2017]). *Let $h_{\hat{\boldsymbol{w}},\boldsymbol{\beta}}$ be a hypothesis output by a regularized ERM algorithm from an m-sized training set T i.i.d. from the target domain $\mathcal{T}$, n source hypotheses $\{h^i_{src} : \|h^i_{src}\|_\infty \leq 1\}_{i=1}^n$, any source weights $\boldsymbol{\beta}$ obeying $\Omega(\boldsymbol{\beta}) \leq \rho$ and $\lambda \in \mathbb{R}_+$. Assume that the loss is bounded by M: $\ell(h_{\hat{\boldsymbol{w}},\boldsymbol{\beta}}(\mathbf{x}), y) \leq M$ for any $(\mathbf{x}, y)$ and any training set. Then, denoting $\kappa = \frac{H}{\sigma}$ and assuming that $\lambda \leq \kappa$, we have with probability of at least $1 - e^{-\eta}$, for all $\eta \geq 0$*

$$\begin{aligned}
\mathrm{R}_{\mathcal{T}}(h_{\hat{\boldsymbol{w}},\boldsymbol{\beta}}) &\leq \mathrm{R}_{\hat{\mathcal{T}}}(h_{\hat{\boldsymbol{w}},\boldsymbol{\beta}}) + \mathcal{O}\left(\frac{\mathrm{R}^{src}_{\mathcal{T}}\kappa}{\sqrt{m}\lambda} + \sqrt{\frac{\mathrm{R}^{src}_{\mathcal{T}}\rho\kappa^2}{m\lambda}} + \frac{M\eta}{m\log\left(1+\sqrt{\frac{M\eta}{u^{src}}}\right)}\right) \\
&\leq \mathrm{R}_{\hat{\mathcal{T}}}(h_{\hat{\boldsymbol{w}},\boldsymbol{\beta}}) + \mathcal{O}\left(\frac{\kappa}{\sqrt{m}}\left(\frac{\mathrm{R}^{src}_{\mathcal{T}}}{\lambda} + \sqrt{\frac{\mathrm{R}^{src}_{\mathcal{T}}\rho}{\lambda}}\right) + \frac{\kappa}{m}\left(\frac{\sqrt{\mathrm{R}^{src}_{\mathcal{T}}M\eta}}{\lambda} + \sqrt{\frac{\rho}{\lambda}}\right)\right),
\end{aligned}$$

where $u^{src} = \mathrm{R}^{src}_{\mathcal{T}}\left(m + \frac{\kappa\sqrt{m}}{\lambda}\right) + \kappa\sqrt{\frac{\mathrm{R}^{src}_{\mathcal{T}}m\rho}{\lambda}}$ and $\mathrm{R}^{src}_{\mathcal{T}} = \mathrm{R}_{\mathcal{T}}(h^{\boldsymbol{\beta}}_{src})$ is the risk of the source hypothesis combination.

The following conclusions can be drawn from this result.

1. If $\mathrm{R}^{\mathrm{src}}_{\mathcal{T}}$ is high, then $h^{\boldsymbol{\beta}}_{\mathrm{src}}$ has no use for transfer, and would only hurt the performance in the target domain;
2. If $m = \mathcal{O}(1/\mathrm{R}^{\mathrm{src}}_{\mathcal{T}})$, then a small value $\mathrm{R}^{\mathrm{src}}_{\mathcal{T}}$ allows a faster convergence rate of $\mathcal{O}(\sqrt{\rho}/m\sqrt{\lambda})$ when making use of the information coming from the source hypotheses combination.

Comparison with standard theory of domain adaptation Recall that the seminal results presented in Section 3 have the following general form

$$\mathrm{R}_{\mathcal{T}}(h) \leq \mathrm{R}_{\mathcal{S}}(h) + d(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \lambda\,,$$

where d is some divergence between the source and target marginal distributions and λ refers to the adaptation capability of the hypothesis class $\mathcal{H}$ from where h is taken.

In general, domain adaptation bounds cannot be directly compared to the result of Theorem 44, even though the term R^{src} can be interpreted as $\mathcal{H}\Delta\mathcal{H}$-divergence by defining $\mathcal{H} = \{\mathbf{x} \mapsto \langle \boldsymbol{\beta}, \boldsymbol{h}_{\mathrm{src}}(\mathbf{x})\,\rangle | \; \Omega(\boldsymbol{\beta}) \leq \tau\}$ where $\boldsymbol{h}_{\mathrm{src}}(\mathbf{x}) = [h^1_{\mathrm{src}}(\mathbf{x}), \ldots, h^n_{\mathrm{src}}(\mathbf{x})]^\top$, and fixing $h = h^{\boldsymbol{\beta}}_{\mathrm{src}} \in \mathcal{H}$, such that

$$R^{\mathrm{src}} = \mathrm{R}_{\mathcal{T}}(h^{\boldsymbol{\beta}}_{\mathrm{src}}) \leq \mathrm{R}_{\mathcal{S}}(h^{\boldsymbol{\beta}}_{\mathrm{src}}) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \lambda_{\mathcal{H}}.$$

If we insert this inequality into the result presented above, then for any hypothesis h and $\lambda \leq 1$, and $\rho \leq 1/\lambda$, we have

$$\mathrm{R}_{\mathcal{T}}(h) \leq \mathrm{R}_{\hat{\mathcal{S}}}(h) + \mathcal{O}\left(\frac{\mathrm{R}_{\mathcal{S}}(h^{\boldsymbol{\beta}}_{\mathrm{src}}) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \lambda_{\mathcal{H}}}{\sqrt{m}\lambda} + \frac{1}{m\lambda}\right). \tag{21}$$

The two results agree that the divergence between the domains has to be small to generalize well. The divergence is actually controlled by the choice of $\boldsymbol{h}_{\mathrm{src}}$, while the complexity of the hypothesis class $\mathcal{H}$ is controlled by τ. In traditional domain adaptation, a hypothesis h performs well on the target domain only if it performs well on the source domain, under the condition that $\mathcal{H}$ is expressive enough to ensure adaptation, or in other words that the $\lambda_{\mathcal{H}}$ term should be small. In HTL, however, this condition can be relaxed, as highlighted by Equation (21), which implies that a good source model has to perform well on its own domain. Additionally, while in traditional domain adaptation the λ-term is assumed to be small – otherwise there is no hypothesis that can perform well on both domains at the same time, and the adaptation cannot be effective – in HTL, the transfer can still be beneficial even for large λ, due to the availability of the labeled target samples.

7.3 Other relevant contributions

[Li and Bilmes, 2007] The authors of this study investigated HTL from the Bayesian perspective, by proposing a PAC-Bayesian study and deriving bounds that capture the relationship between domains by an additive KL-divergence term, which is classic in a PAC-Bayesian setting. In the particular case of logistic regression, they showed that the divergence term is upper-bounded by $\|h - h_{\mathrm{src}}\|^2$, which motivated the biased regularization term in logistic regression and the interest of incorporating the source hypothesis into the adaptation model.

[Morvant et al., 2012] As in [Dhouib and Redko, 2018], the authors of this paper considered learning with a particular family of similarity functions introduced in [Balcan et al., 2008], and provided a generalization bound for this using the algorithmic robustness framework.

[Habrard et al., 2013] This paper presented a study on iterative self-labeling for domain adaptation, where at each iteration a hypothesis h is learned from the current sample S, some target samples are pseudo-labeled from T_u by h, and these are incorporated into the source sample S to progressively modify the current classifier. Their analysis suggested that such a procedure theoretically solves a domain adaptation problem when the hypothesis obtained at each iteration improves upon the hypothesis obtained without self-labeling.

[Perrot and Habrard, 2015] The theoretical results of this paper made use of an extension of the concept of algorithmic stability (see Subsection 2.6) to similarity learning, and provided generalization bounds for this in the HTL framework presented above. In particular, instead of learning a set of weights $\mathbf{w}$ that parameterize the hypothesis function h, the authors learned a similarity matrix $\mathbf{M}$ that is regularized with respect to a similarity matrix $\mathbf{M}_S$ learned in a related source domain.

[Habrard et al., 2016] In this study, the authors analyzed a setting that consisted of learning N weak hypotheses[4] using the labeled source sample and reweights them differently by taking into account the data from the unlabeled target domain. Their theoretical analysis proves that the proportion of target examples having a margin γ decreases exponentially with the number of iterations, but does not benefit from any generalization guarantees given by an upper bound on the risk with respect to the target distribution.

[Du et al., 2017] In this study, the authors considered an extension of the original HTL setting through a general form of transfer defined by transformation functions that can be provided as input to the HTL algorithm. These transformation functions include, for instance, the offset transfer and scale transfer, thus generalizing the study of [Kuzborskij and Orabona, 2013].

Also, we note that the study of [Hanneke and Kpotufe, 2019] mentioned in Section 4 analyzed the HTL-based adaptation approach, and showed its efficiency in improving the target performance.

7.4 Summary

In this section, we have presented theoretical results that allow algorithmic properties of adaptation algorithms to be taken into consideration. First, we recalled how the algorithmic robustness can be extended to the domain adaptation setting, with relaxation of the covariate-shift assumption. Secondly, we focused on a different domain adaptation setting called hypothesis transfer learning, where there is no access to source samples, but to source model(s) given by the learned hypotheses. In this setting, we presented theoretical results obtained in the case of regularized ERM-based algorithms that rely on the algorithmic stability framework.

In general, we can highlight several important differences of this framework with respect to the results seen in the previous sections. These are the following:

1. Contrary to the divergence-based bounds, the learning guarantees presented in this section do not include a term that measures the discrepancy between the marginal distributions of the two domains. This is as expected, as in the HTL scenario we do not have access to a learning sample from the source domain, but only to a hypothesis learned on it;
2. The potential success of adaptation in the HTL framework depends on the performance of the source hypothesis on the target distribution, and allows a better hypothesis to be learned, even on small samples when some assumptions are fulfilled;
3. Contrary to the majority of the results seen so far, the adaptability term is absent from the bounds related to the HTL setting, as in this case, the learner has access to some target labeled data.

8 Conclusions and discussion

In this survey, we have presented an overview of the existing theoretical guarantees that have been proven for the domain adaptation problem, a learning setting that extends traditional learning paradigms to the case where the model is learned and deployed on samples coming from different, yet related, probability distributions. The cited theoretical

[4] A weak hypothesis for $\mathcal{D}$ is a hypothesis such that $\mathrm{R}_{\mathcal{D}}(h) = \frac{1}{2} - \varepsilon$, where $\varepsilon > 0$ is a small constant.

results often take the shape of learning bounds, where the goal is to relate the error of a model on the training domain (also called the source domain) to that of the test domain (also called the target domain). To this end, we note that the results presented are highly intuitive, as they explicitly introduce the dependence of the relationship between the two errors mentioned above to the similarity of their data-generating probability distributions and that of their corresponding labeling functions. Consequently, this two-way relatedness between the source and target domains characterizes both the unsupervised proximity of two domains, by comparing their marginal distributions, and the possible labelings of their samples, by looking for a good model with a low error with respect to these. This general trade-off is preserved, in one way or another, in the majority of published results on the subject, and thus this can be considered as a cornerstone of modern domain adaptation theory.

As any survey that gives an overview of a certain scientific field, this one would have been incomplete without identification of the problems that remain open. In the context of domain adaptation theory, these problems can be arguably split into two main categories, where the first is related to the domain adaptation problem itself, and the second is related to other learning scenarios similar to domain adaptation. For the first category, one important open problem is that of characterizing the *a-priori* adaptability of the adaptation given by the joint error term. Indeed, this term is often assumed to be small for domain adaptation to be possible, although only one previous study [Redko et al., 2019b] suggested an actual way for its consistent estimation from a handful of labeled target data. On the other hand, domain adaptation has been recently extended to open-set and heterogeneous settings, where for the former both source and target domains are allowed to have nonoverlapping classes, while for the latter the input space of the two domains might differ. To the best of our knowledge, there are still no theoretical results that analyze these scenarios. This point brings us to the second category of open problems related to learning scenarios similar to that of domain adaptation, such few-shot learning problems, where there is the need to learn on a sample that contains no or only a few examples of certain classes appearing in the test data. Intuitively, this problem is tightly related to domain adaptation and might naturally inherit some of its theoretical guarantees, although there have been no studies that make this link explicit in the literature to date.

Finally, this survey has not discussed such closely related topics as multitask learning, learning-to-learn, and lifelong learning, to name but a few. This particular choice was made to remain focused on one particular problem, as this is vast enough on its own. We also admit that there are certainly other relevant papers that provide guarantees for domain adaptation that are not included in this survey[5]. This field, however, is so large and recent advances have been published at such a great pace that it is simply not possible to keep up with it and to report all possible results, without breaking the general structure and the narrative of our survey.

References

[Ambroladze et al., 2006] Ambroladze, A., Parrado-Hernández, E., and Shawe-Taylor, J. (2006). Tighter PAC-Bayes bounds. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 9–16.

[Balcan et al., 2008] Balcan, M., Blum, A., and Srebro, N. (2008). Improved guarantees for learning via similarity functions. In *COLT*, pages 287–298.

[Bartlett and Mendelson, 2002] Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

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

[Ben-David et al., 2007] Ben-David, S., Blitzer, J., Crammer, K., and Pereira, O. (2007). Analysis of representations for domain adaptation. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 137–144.

[Ben-David et al., 2010b] Ben-David, S., Lu, T., Luu, T., and Pál, D. (2010b). Impossibility theorems for domain adaptation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9, pages 129–136.

[Ben-David et al., 2012] Ben-David, S., Shalev-Shwartz, S., and Urner, R. (2012). Domain adaptation–can quantity compensate for quality? In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*.

[Ben-David and Urner, 2012] Ben-David, S. and Urner, R. (2012). On the hardness of domain adaptation and the utility of unlabeled target samples. In *Proceedings of the conference on Algorithmic Learning Theory (ALT)*, pages 139–153.

[5] If your paper does not appear in this survey, but seems relevant to its contents, please let us know, and we will try to include it in the revised versions.

[Ben-David and Urner, 2013] Ben-David, S. and Urner, R. (2013). Domain adaptation as learning with auxiliary information. In *Workshop@NIPS New Directions in Transfer and Multi-Task*.

[Blitzer et al., 2008] Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. (2008). Learning bounds for domain adaptation. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 129–136.

[Bolley et al., 2007] Bolley, F., Guillin, A., and Villani, C. (2007). Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3-4):541–593.

[Bousquet and Elisseeff, 2002] Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526.

[Catoni, 2007] Catoni, O. (2007). *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56. Inst. of Mathematical Statistic.

[Chen et al., 2009] Chen, B., Lam, W., Tsang, I., and Wong, T.-L. (2009). Extracting discriminative concepts for domain adaptation in text mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 179–188.

[Cortes et al., 2010] Cortes, C., Mansour, Y., and Mohri, M. (2010). Learning bounds for importance weighting. In *NIPS*, pages 442–450.

[Cortes and Mohri, 2011] Cortes, C. and Mohri, M. (2011). Domain adaptation in regression. In *Proceedings of the conference on Algorithmic Learning Theory (ALT)*, pages 308–323.

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

[Cortes et al., 2015] Cortes, C., Mohri, M., and Muñoz Medina, A. (2015). Adaptation algorithm and theory based on generalized discrepancy. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 169–178. ACM.

[Courty et al., 2017] Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. In *NIPS*, pages 3730–3739.

[Courty et al., 2014] Courty, N., Flamary, R., Rakotomamonjy, A., and Tuia, D. (2014). Optimal transport for domain adaptation. In *Workshop@NIPS on Optimal Transport and Machine Learning*.

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

[Dhouib and Redko, 2018] Dhouib, S. and Redko, I. (2018). Revisiting (\epsilon, \gamma, \tau)-similarity learning for domain adaptation. In *NeurIPS*, pages 7408–7417.

[Dhouib et al., 2020a] Dhouib, S., Redko, I., Kerdoncuff, T., Emonet, R., and Sebban, M. (2020a). A swiss army knife for minimax optimal transport. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 7613–7622.

[Dhouib et al., 2020b] Dhouib, S., Redko, I., and Lartizien, C. (2020b). Margin-aware adversarial domain adaptation with optimal transport. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4619–4629.

[Dietterich, 2000] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

[Du et al., 2017] Du, S. S., Koushik, J., Singh, A., and Póczos, B. (2017). Hypothesis transfer learning via transformation functions. In *NIPS*, NIPS'17, page 574–584.

[Dudley, 2002] Dudley, R. M. (2002). *Real analysis and probability*. Cambridge studies in advanced mathematics. Cambridge University Press.

[Gao and Galvao, 2014] Gao, Z. and Galvao, A. (2014). Minimum integrated distance estimation in simultaneous equation models. *arXiv preprint arXiv:1412.2143*.

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

[Germain et al., 2013] Germain, P., Habrard, A., Laviolette, F., and Morvant, E. (2013). A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 738–746.

[Germain et al., 2016] Germain, P., Habrard, A., Laviolette, F., and Morvant, E. (2016). A new PAC-Bayesian perspective on domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 48, pages 859–868.

[Germain et al., 2020] Germain, P., Habrard, A., Laviolette, F., and Morvant, E. (2020). Pac-bayes and domain adaptation. *Neurocomputing*, 379:379–397.

[Germain et al., 2009] Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. (2009). PAC-Bayesian learning of linear classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 353–360.

[Germain et al., 2015] Germain, P., Lacasse, A., Laviolette, F., Marchand, M., and Roy, J.-F. (2015). Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(1):787–860.

[Habrard et al., 2013] Habrard, A., Peyrache, J.-P., and Sebban, M. (2013). Iterative self-labeling domain adaptation for linear structured image classification. *International Journal on Artificial Intelligence Tools (IJAIT)*, 22(05).

[Habrard et al., 2016] Habrard, A., Peyrache, J.-P., and Sebban, M. (2016). A new boosting algorithm for provably accurate unsupervised domain adaptation. *Knowledge and Information Systems*, 47(1):45–73.

[Hanneke and Kpotufe, 2019] Hanneke, S. and Kpotufe, S. (2019). On the value of target data in transfer learning. In *NeurIPS*.

[Hoffman et al., 2018] Hoffman, J., Mohri, M., and Zhang, N. (2018). Algorithms and theory for multiple-source adaptation. In *NeurIPS*, pages 8256–8266.

[Huang et al., 2006] Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M., and Schölkopf, B. (2006). Correcting sample selection bias by unlabeled data. In *NIPS*, pages 601–608.

[Johansson et al., 2019] Johansson, F. D., Sontag, D. A., and Ranganath, R. (2019). Support and invertibility in domain-invariant representations. In *AISTATS*, pages 527–536.

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

[Kifer et al., 2004] Kifer, D., Ben-David, S., and Gehrke, J. (2004). Detecting change in data streams. In *Proceedings of the International Conference on Very Large Data Bases*, pages 180–191.

[Kolmogorov and Tikhomirov, 1959] Kolmogorov, A. N. and Tikhomirov, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86.

[Koltchinskii and Panchenko, 1999] Koltchinskii, V. and Panchenko, D. (1999). Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, pages 443–459. Birkhauser.

[Kuroki et al., 2019] Kuroki, S., Charoenphakdee, N., Bao, H., Honda, J., Sato, I., and Sugiyama, M. (2019). Unsupervised domain adaptation based on source-guided discrepancy. In *AAAI*, pages 4122–4129.

[Kuzborskij, 2018] Kuzborskij, I. (2018). *Theory and Algorithms for Hypothesis Transfer Learning*. PhD thesis, EPFL. https://infoscience.epfl.ch/record/232494/files/EPFL_TH8011.pdf.

[Kuzborskij and Orabona, 2013] Kuzborskij, I. and Orabona, F. (2013). Stability and hypothesis transfer learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 942–950.

[Kuzborskij and Orabona, 2017] Kuzborskij, I. and Orabona, F. (2017). Fast Rates by Transferring from Auxiliary Hypotheses. 106(2):171–195.

[Lacasse et al., 2006] Lacasse, A., Laviolette, F., Marchand, M., Germain, P., and Usunier, N. (2006). PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 769–776.

[Langford and Shawe-Taylor, 2002] Langford, J. and Shawe-Taylor, J. (2002). PAC-Bayes & margins. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 439–446.

[Li and Bilmes, 2007] Li, X. and Bilmes, J. (2007). A bayesian divergence prior for classiffier adaptation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 275–282.

[Mansour et al., 2008] Mansour, Y., Mohri, M., and Rostamizadeh, A. (2008). Domain adaptation with multiple sources. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 1041–1048.

[Mansour et al., 2009a] Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009a). Domain adaptation: Learning bounds and algorithms. In *Proceedings of the Conference on Learning Theory (COLT)*.

[Mansour et al., 2009b] Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009b). Multiple source adaptation and the rényi divergence. In *UAI*, pages 367–374.

[Mansour and Schain, 2014] Mansour, Y. and Schain, M. (2014). Robust domain adaptation. *Annals of Mathematics and Artificial Intelligence*, 71(4):365–380.

[McAllester, 1999] McAllester, D. A. (1999). Some PAC-Bayesian theorems. *Machine Learning*, 37:355–363.

[McNamara and Balcan, 2017] McNamara, D. and Balcan, M. (2017). Risk bounds for transferring representations with and without fine-tuning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2373–2381.

[Monge, 1781] Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*.

[Morvant et al., 2012] Morvant, E., Habrard, A., and Ayache, S. (2012). Parsimonious unsupervised and semi-supervised domain adaptation with good similarity functions. *Knowledge and Information Systems*, 33(2):309–349.

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

[Orabona et al., 2009] Orabona, F., Castellini, C., Caputo, B., Fiorilla, A., and Sandini, G. (2009). Model adaptation with least-squares svm for adaptive hand prosthetics. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2897–2903.

[Pan et al., 2008] Pan, S. J., Kwok, J. T., and Yang, Q. (2008). Transfer learning via dimensionality reduction. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 677–682.

[Pan et al., 2009] Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2009). Domain adaptation via transfer component analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1187–1192.

[Perrot and Habrard, 2015] Perrot, M. and Habrard, A. (2015). A theoretical analysis of metric hypothesis transfer learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1708–1717.

[Re and Valentini, 2012] Re, M. and Valentini, G. (2012). Ensemble methods: a review. In *Advances in Machine Learning and Data Mining for Astronomy*, pages 563–582.

[Redko, 2015] Redko, I. (2015). *Nonnegative Matrix Factorization for Unsupervised Transfer Learning*. PhD thesis, Paris North University.

[Redko et al., 2019a] Redko, I., Courty, N., Flamary, R., and Tuia, D. (2019a). Optimal transport for multi-source domain adaptation under target shift. In *AISTATS*, pages 849–858.

[Redko et al., 2017] Redko, I., Habrard, A., and Sebban, M. (2017). Theoretical analysis of domain adaptation with optimal transport. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pages 737–753.

[Redko et al., 2019b] Redko, I., Habrard, A., and Sebban, M. (2019b). On the analysis of adaptability in multi-source domain adaptation. *Mach. Learn.*, 108(8-9):1635–1652.

[Redko et al., 2019c] Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. (2019c). *Advances in Domain Adaptation Theory*. Elsevier.

[Saitoh, 1997] Saitoh, S. (1997). *Integral Transforms, Reproducing Kernels and their Applications*. Pitman Research Notes in Mathematics Series.

[Saunders et al., 1998] Saunders, C., Gammerman, A., and Vovk, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 515–521.

[Seeger, 2002] Seeger, M. (2002). Pac-bayesian generalisation error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3(Oct):233–269.

[Sejdinovic et al., 2013] Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K., et al. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291.

[Shalev-Shwartz and Ben-David, 2014] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press.

[Shawe-Taylor and Williamson, 1997] Shawe-Taylor, J. and Williamson, R. C. (1997). A PAC Analysis of a Bayesian Estimator. In *Proceedings of the annual workshop on Computational learning theory (COLT)*, pages 2–9.

[Shen et al., 2018] Shen, J., Qu, Y., Zhang, W., and Yu, Y. (2018). Wasserstein distance guided representation learning for domain adaptation. In *AAAI*, pages 4058–4065.

[Song, 2008] Song, L. (2008). *Learning via Hilbert Space Embedding of Distributions*. PhD thesis, University of Sydney.

[Sugiyama et al., 2008] Sugiyama, M., Nakajima, S., Kashima, H., Bünau, P. V., and Kawanabe, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 1433–1440.

[Valiant, 1984] Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27:1134–1142.

[Vapnik, 2006] Vapnik, V. (2006). *Estimation of Dependences Based on Empirical Data*. Springer Science & Business Media.

[Vapnik, 1995] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc.

[Vapnik and Chervonenkis, 1971] Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280.

[Villani, 2009] Villani, C. (2009). *Optimal Transport : Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin.

[Xu and Mannor, 2010] Xu, H. and Mannor, S. (2010). Robustness and generalization. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 503–515.

[Xu and Mannor, 2012] Xu, H. and Mannor, S. (2012). Robustness and generalization. *Mach. Learn.*, 86(3):391–423.

[Zhang et al., 2012] Zhang, C., Zhang, L., and Ye, J. (2012). Generalization bounds for domain adaptation. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 3320–3328.

[Zhang et al., 2019] Zhang, Y., Liu, T., Long, M., and Jordan, M. (2019). Bridging Theory and Algorithm for Domain Adaptation. In *International Conference on Machine Learning*, pages 7404–7413.

[Zhao et al., 2019] Zhao, H., des Combes, R. T., Zhang, K., and Gordon, G. J. (2019). On learning invariant representations for domain adaptation. In *ICML*, pages 7523–7532.

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